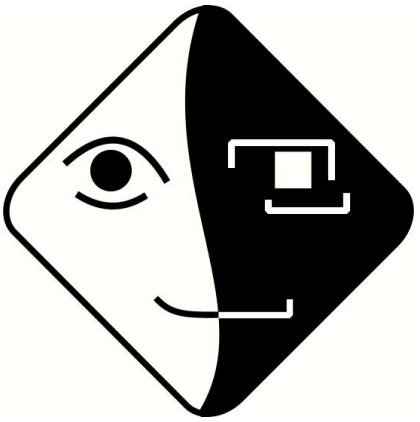
III. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2005

Szeged, 2005. december 8-9.
http://www.inf.u-szeged.hu/mszny2005

Szerkesztette: Alexin Zoltán és Csendes Dóra {alexin, dcsendes}@inf.u-szeged.hu

Felelős kiadó: Szegedi Tudományegyetem Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.

Nyomtatta: Juhász Nyomda
6771, Szeged, Makai út 4.

Szeged, 2005. november

Előszó

2005. december 8-9. között harmadik alkalommal kerül megrendezésre a Magyar Számítógépes Nyelvészeti Konferencia. Nagy örömömre szolgál, hogy a rendezvény évről évre nagyszámú érdeklődőt vonz az ország különböző tájairól. A konferencia fő célja a nyelvtechnológia (elsősorban a szöveg- és a beszédfeldolgozás) területén elvégzett vagy folyamatban lévő kutatások és fejlesztések legaktuálisabb eredményeinek bemutatása. Lehetőség nyílik kapcsolódó hallgatói projektek, ill. a számítógépes nyelvészet ipari alkalmazásainak ismertetésére is.

Az idei felhívásra beérkezett tudományos értekezések közül a programbizottság 40-et fogadott el előadás megtartására, és további 9-et poszter prezentáció, ill. 4-et laptopos bemutató megtartására.

Szeretnék köszönetet mondani a programbizottságnak: Vámos Tibor programbizottsági elnöknek, valamint Gordos Géza, László János, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság: Csendes Dóra és Alexin Zoltán munkáját.

Csirik János, a rendezőbizottság elnöke
Szeged, 2005. november

Tartalomjegyzék

I. Elmélet

II. Ontológia

III. Fordítás és szótár

IV. Morfológia és kivonatolás

V. Szintaxis és szemantika

VI. Pszichológiai szempontú szövegfeldolgozás

VII. Beszédtechnológia, kommunikáció

VIII. Poszter prezentációk

IX. Laptopos bemutatók

I. Elmélet

A világháló nyelvi vizsgálata
(Néhány tanulság a gépi feldolgozás számára)

Prószéky Gábor

MorphoLogic
1126 Budapest, Orbánhegyi út 5.
proszeky@morphologic.hu

Kivonat: A webet sokan nem pusztán a tartalom olvasására használják, hanem nyelvi készségük spontán alakítása is a weben történik. Egészen pontosan: cikkolvasás, levelezés és mindenféle tevékenység történik az internetes eszközök segítségével, így az internetező nyelvi készségére automatikusan hatással van az ott található tartalom megfogalmazási módja, stílusa, és egészen leegyszerűsítve: konkrét megjelenési formája. Kimutatásokkal támasztjuk alá a nyelvi hibák, különösen az angol nyelvi hibák internetes jelentőségét. Sokan – épp akik nincsenek abban a helyzetben, hogy ezeket a hibákat felismerjék – könnyen követhetik ezeket a mintákat is. A nyelvtechnológiai eszközöknek van ezáltal igazán feladva a lecke, hiszen az angol szövegeket sokszor nem „csak" fordítaniuk kell, hanem esetleges hibáikat felismerve és kijavítva az eredetileg szándékolt tartalom fordítását kellene elvégezniük.

1 Bevezetés

Az internet nyelvtechnológiai szempontból nézve nemcsak bájtok vagy karakterek sorozata, hanem különböző nyelveken írt információk gyűjteménye. Az emberiség több évtizedes, évszázados gyakorlata szerint a nyomtatott szövegek hitelessége magasabb, mint a kézírásosaké. Ennek oka egyszerű: a nyomtatott szövegek lektorálás nélkül ritkán jutnak el az olvasóhoz. A helyzet azonban a nyomtatás számítógépesítésével, sőt „internetesítésével" jelentősen megváltozott. Az autentikusnak tűnő formát szinte bárki előállíthatja, és már csak anyagi kérdés, hogy jó minőségű papíron, jó nyomástechnikával, profi számítógépes kiadványszerkesztők segítségével adja-e közre, vagy a nyelvi minőségi hiányokról a forma egyszerűsége is árulkodni fog.

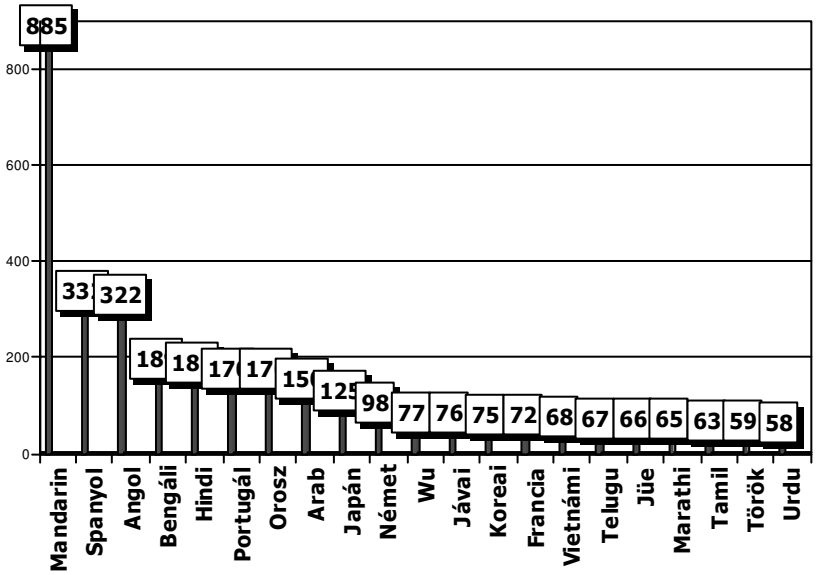
Az internetes szövegközlés még „veszélyesebb" a forma és a tartalom közötti korrelációra nézve: a hírek forrása, a megjelenés helye és még sok más paraméter homályban maradhat. Így az a tény is sokszor ismeretlen, hogy az interneten közölt tartalmat készítője nem az adott nyelv tipográfiai, sőt ortográfiai hagyományai betartásával hozta létre. Mi több, az is előfordul nem is kis számban, hogy grammatikai hibák is lesznek benne. Míg ezek a nyelvi, nyelvtani hibák egyes nyelvek esetében legfeljebb mosolygás tárgyaivá válnak, az angol estében más a helyzet. Kimutatásaink azt a hipotézist támasztják alá, hogy az angol esetében relatíve egyre kevesebb az angol

nyelven publikáló anyanyelvi beszélő, de a weben elérhető szövegek száma nő. Ennek persze lehetne az is az oka, hogy az angol anyanyelvűek sokkal szorgalmasabban használják a webet, mint mások, ám a diagramok tanúsága szerint a nem anyanyelvi beszélők által készített angol tartalom „szaporodik" igazán. Ennek következtében az az elvárás, ami az anyanyelvi beszélők, pontosabban az anyanyelven írók esetében elvárható volt, egyre kevésbé érvényesülhet.

A webet sokan nem pusztán a tartalom olvasására használják, hanem nyelvi készségük spontán alakítása is a weben történik. Egészen pontosan: cikkolvasás, levelezés és mindenféle tevékenység történik az internetes eszközök segítségével, így az internetező nyelvi készségére automatikusan hatással van az ott található tartalom megfogalmazási módja, stílusa, és egészen leegyszerűsítve: konkrét megjelenési formája. Kimutatásokkal támasztjuk alá a(z angol) nyelvi hibák internetes jelentőségét. Sokan – épp akik nincsenek abban a helyzetben, hogy ezeket a hibákat felismerjék – könnyen követhetik ezeket a mintákat is. A nyelvtechnológiai eszközöknek van ezáltal feladva a lecke, hiszen az angol szövegeket sokszor nem „csak" fordítaniuk kell, hanem esetleges hibáikat felismerve és kijavítva az eredetileg szándékolt tartalom fordítását kell elvégezniük.

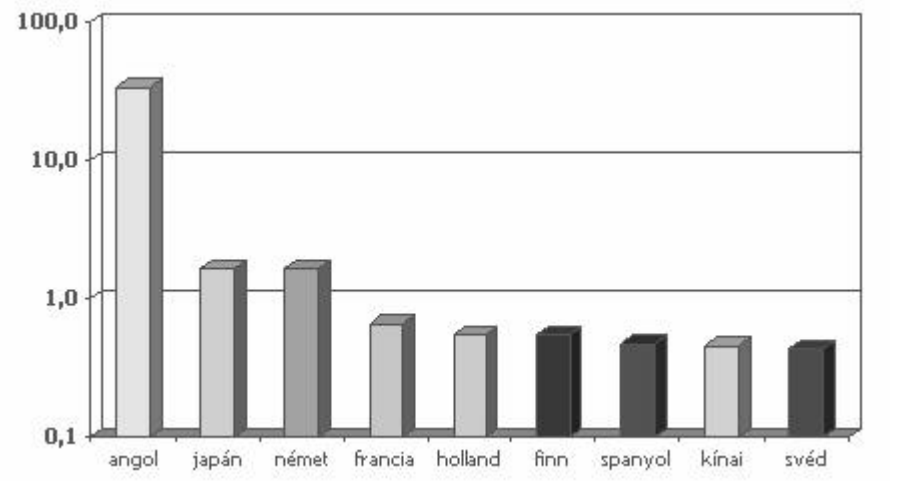
2. A világháló nyelvei – a nyelvtechnológus szemével

A világ nyelveinek beszélők száma szerinti statisztikái elsősorban a harmadik világbeli nyelvek fölényét mutatják (1. ábra). Mint érdekességet érdemes megemlíteni, hogy az első 20 nyelv közt a német az egyetlen, amelyet gyakorlatilag csak Európában beszélnek. A spanyol, az angol, a portugál, az orosz, a francia és a török más kontinenseken is nagy számú beszélővel bír, a többi nyelv pedig eleve nem európai.



1. ábra. A világ nyelvei a beszélők száma szerint

Ugyanakkor az internetes portálok nyelvi háttere meglehetősen más képet mutat, mint a beszélt nyelveké (2. ábra): itt a kis európai nyelvek – a holland, a finn, a svéd – jelennek meg a legelsők között, ám sejthető, hogy az egyes portálok nyelvei a web teljes tartalmát tekintve önmagukban nem meghatározók.



2.ábra. Az internetes portálok nyelvei

A világháló nyelvével kapcsolatos igazán meghatározó információk azok, amik elsősorban az internethasználók anyanyelvi megoszlásáról, illetve a weben található szövegek nyelveinek megoszlásáról szólnak. Az internethasználók létszámát mutatja nyelvenként millió főben az 1. táblázat [1]. Sorrendjüket az internetezők és a nyelvet beszélők aránya határozza meg. A japán nyelvű internethasználók gyakorlatilag a teljes japán társadalmat lefedik, hiszen a lakosság 16 %-a nem internetezik, ezek pedig a kicsi gyerekek és a nagyon idősek. A második helyen a norvég–dán–svéd–izlandi technikai összevonásával keletkeztetett, nem létező „skandináv" nyelvet találjuk a statisztikában. Ami feltétlen érdekes, hogy az olasz nyelv ilyen előkelő helyen áll a listában, az ilyen listákban egyébként mindig is előkelő helyeken álló holland és a német mellett. Feltűnhet, hogy az angol anyanyelvűeknek még csak a kisebbik fele használja a világhálót, a kínaiaknak, spanyoloknak és portugáloknak pedig csak az egynegyede. A világ három és fél milliárdos „maradékrészén" gyakorlatilag még nincs ott a világháló. Meglepőnek tűnő, de talán nem túlzó jóslat azt mondani, hogy a nagy sávszélesség hamarabb fog eljutni a harmadik világba, mint a sok mindent megoldani képes szociális és gazdasági segélyek. Ez viszont ezeken a területeken akár az írásbeliség erősödését és a „lappangó" szürkeállomány aktivizálását is lehetővé teheti az elkövetkező években.

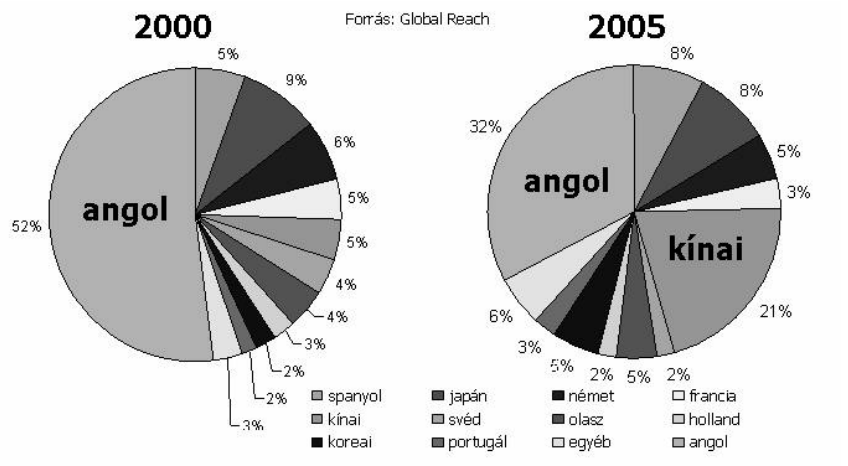
Visszatérve az internetezők száma szerinti „legnagyobb" nyelvekre, az angolra (231 millió) és a kínaira (220 millió), azt mondhatjuk, hogy mindkettőnek döntő jelentősége van. Az angol anyanyelvű webhasználók nem fogják egyhamar elveszíteni vezető szerepüket a latin betűs világban, ám a kínaiak hamarosan átvehetik a „legtöbb egynyelvű felhasználó" címet. Ha ehhez hozzávesszük, hogy Kínában valójában több, egymástól jelentős mértékben eltérő dialektus, sőt, a kínaival nem rokon nyelv

beszélői is olvasnak kínaiul, akkor azt mondhatjuk, hogy kezd előnnyé válni az „egyírásúság". Az, amiről itt, Európában sokszor úgy gondolkozunk, hogy mekkora nehézség lehet a kínai gyerekek olvasástanulása, az hamarosan úgy is megfogalmazható, hogy mekkora üzleti előny, hogy egyetlen írással lehet lokalizálni milliárdos piacot: egyetlen kézikönyv-fordítás, egyetlen használati utasítás elég, szemben a láthatóan egyszerűbb latin betűvel író, de rengeteg nyelvet beszélő, és egymást sem írásban, sem szóban nem értő Európával. Ugyanígy hatalmas előny, ha egymással beszédben nehezen szót értő emberek írásra áttérve egyértelmű kommunikációt kezdeményezhetnek. Vagyis: az internet megjelenése kimondottan kedvez a kínai írásnak. Ha belegondolunk, az ikonikus nyelv mifelénk sem ritka (csak nem több ezer éves múltra nyúlik vissza): így jelöljük a kerékpárutat, a mozgáskorlátozottak parkolóhelyét, gyakorlatilag így jelöl a KRESZ mindent, de így jelöljük a férfi és női mellékhelyiségeket, vagy épp a sportágakat az olimpián. A nehézség az, hogy ez a szimbólumegyüttes olyan szintű nyelvi kommunikáció pontos közvetítésére nem alkalmas, mint a távol-keleti ikonikus írás, a kandzsik világa.

1. táblázat: Az internethasználók nyelvei

Nyelv	1999	2002	2005	Beszélők száma	Internetezők aránya	Nem internetezik
japán	20	64	105	125	84%	20
skandináv*	8	14	15	19	75%	4
olasz	10	25	42	57	74%	15
holland	6	13	15	20	73%	5
német	14	44	71	98	72%	27
francia	10	24	49	72	68%	23
koreai	5	30	50	75	67%	25
angol	85	165	231	500	46%	269
kínai	10	75	220	885	25%	665
spanyol	13	49	80	332	24%	252
portugál	4	20	38	170	22%	132
egyéb	6	63	140	3500	4%	3360

A kérdés tehát most már az, hogy a világ számára nyelvtechnológiai szempontból a kínai vagy az angol internetes terjedése jelenti-e a nagyobb kihívást. Természetesen aki meg akar tanulni kínaiul, megtanulhat, de míg a kínai még jó darabig kizárólag a kínaiak nyelve lesz, az angol már rég nem csak az angol anyanyelvűeké. Az angol nemzetközi nyelv mivoltát az ilyen tömegesen megjelenő kínai webhasználó sem tudja veszélyeztetni. Miért? Mert az internetes szövegek nyelvi arányai más képet mutatnak, mint a világháló használóinak nyelvi arányai. Nézzük meg a Global Reach becslését [1] a világhálón fellelhető szövegek nyelveiről (3. ábra). E szerint az angol szövegek aránya – elsősorban a relatíve gyorsan növő kínai szövegmennyiség miatt – csökken, de a kördiagram azt már nem mutatja, hogy öt év alatt mekkorára nőtt maga a kör, azaz a weben elérhető szövegek összmennyisége.



3. ábra. A weben található szövegek nyelvei

Egy-egy nyelv világhálós elterjedtségének mérése alapvetően csak becslésekkel történhet. Ezek a becslések azokon a nyelvstatisztikai ismereteken alapulnak, hogy megfigyelhető, milyen gyakoriak egyes szavak a különböző nyelvek meglevő szövegkorpuszaiban. Hogy mely szavakról van szó, nem kérdés: ezeknek a kiválasztott szavaknak elég gyakoriaknak kell lenniük ahhoz, hogy bármilyen tartalmú szöveget többé-kevésbé egyformán jellemezzenek. Ilyen szavak a névelők, a kötőszók, az elöljárók – melyik-melyik különböző módon jellemző a különböző nyelvekre. Nyilván az nem elég, ha egy szó ugyan gyakori egy nyelvben, de egy másik nyelvben is az. Ilyenre lehet példa a német *die* névelő, mely az angol *die* igével azonos alakú, így nem alkalmas a német szövegek egyértelmű azonosítására. A magyar *a* névelő is egybeesik az angol *a* névelővel. A másik magyar határozott névelő, az *az* szerencsésebb, mint például az angol *an*, mely egy német elöljáróval esik egybe. Tehát nemcsak a nyelvre jellemző, hanem az egyes nyelvekre kizárólagosan jellemző szavakra van szükség. A német nyelvű szövegek méretének becsléséhez használt, csak a németre jellemző gyakori szavakat mutatja a 2. táblázat [2].

2. táblázat. Német szavak gyakorisága a német nyelvű web méretének becsléséhez

Szó	Relatív gyakoriság	Előfordulások száma	A német web becsült mérete
und	0,02892370	101 250 80S6	3 500 617 348
auf	0,00744444	24 852 802	3 338 438 082
ist	0,00886430	26 429 327	2 981 546 991
sich	0,00604594	17 547 518	2 902 363 900
eine	0,00691066	19 739 540	2 856 389 983
nicht	0,00646585	18 294 174	2 829 353 294
wird	0,00400690	11 286 438	2 816 750 605
auch	0,00581108	15 504 327	2 668 062 907
sind	0,00477555	11 944 284	2 501 132 644
oder	0,00561180	13 566 463	2 417 488 684

A fenti számok átlaga adja azt a becslést, mely alapján a világhálón található német nyelvű szavak számára 3 268 760 356 jön ki. Ugyanígy lehet becslést tenni más nyelvek weben található anyagainak méretére is. Ezt a becslést az európai nyelvek többségére Kilgariff és Greffenstette [2] elvégezte. Az általuk készített felmérés eredményét mutatja a 3. táblázat. A listában a magyar határozottan előkelő helyen áll – bár az esetleges büszkeségen túl – ebből a megállapításból semmilyen lényeges következtetést nem vonhatunk le. Abból viszont már sokkal inkább, hogy az angol nagyságrenddel előzi meg az összes többi nyelvet. Ezt fejtjük ki bővebben a következő fejezetben.

3. Gondolatok a világháló angoljáról

Az emberiség tudása elsősorban írásban rögzül. Ennek az ismeretanyagnak nagy része ma már elérhető az interneten keresztül. Ez természetesen nem jelenti azt, hogy nem volna jelentős a világháló jpg-, avi- vagy éppen mp3-tartalma, de a „közös tudás" kódolásának és a kommunikációnak elsődleges formája az írás. Azaz: a szövegek. Ezek a szövegek értelemszerűen mindig valamilyen nyelven vannak „kódolva". A web elterjedésével egyre szaporodnak a nemzeti nyelven elérhető adatok. Sőt, a globalizáció egyfajta mellékhatásaként a helyi vásárlók jobb meggyőzése érdekében egyre több helyi nyelvű, lokalizált tartalom jelenik meg. Persze a helyi cégek szintén szeretnének szerepet játszani a világban, ezért ők meg „nemzetközileg érthető nyelven", gyakorlatilag tehát angolul publikálnak. Tehát nő a helyi nyelvi tartalom, de – ettől valamivel kisebb mértékben – nő az angol is. Ez eddig csak hipotézis, de nézzük eddigi számainkat, hiszen a nyelvekkel kapcsolatos világhálós adatokból több érdekes következtetést lehet levonni.

3. táblázat. A különböző európai nyelveken írt tartalmak világhálós méretének becslése

	Nyelv	Szó a weben
1	angol	76 598 718 000
2	német	7 035 850 000
3	francia	3 836 874 000
4	spanyol	2 658 631 000
5	olasz	1 845 026 000
6	magyar	457 522 000
7	dán	346 945 000
8	finn	326 379 000
9	lengyel	322 283 000
10	szlovák	216 595 000
11	katalán	203 592 000
12	török	187 367 000
13	maláj	157 241 000
14	horvát	136 073 000
15	szlovén	119 153 000
16	észt	98 066 000

	Nyelv	Szó a weben
17	portugál	1 333 664 000
18	holland	1 063 012 000
19	svéd	1 003 075 000
20	norvég	609 934 000
21	cseh	520 181 000
22	ír	88 283 000
23	román	86 392 000
24	eszperantó	57 154 000
25	latin	55 943 000
26	baszk	55 340 000
27	izlandi	53 941 000
28	lett	39 679 000
29	litván	35 426 000
30	velszi	14 993 000
31	breton	12 705 000
32	albán	10 332 000

Egyrészt azt látjuk a kimutatásokból, hogy az internetes tartalomnak mintegy kétharmada angol. Ugyanakkor azt is észrevehetjük, hogy az internethasználók közel kétharmadának az anyanyelve nem angol. Azt is látjuk tehát, hogy a nem-angol anyanyelvűek világában a web angol tartalmának egyre nagyobb része jelenik meg. Viszont aki a webet olvassa, az növekvő számban nem-angol anyanyelvű! Aki pedig a webes tartalmat közzéteszi, az – mint a kimutatásokból látjuk – csökkenő részben angol anyanyelvű. Mit jelent ez a gyakorlatban?

Nem meglepő megállapítás, ha azt mondjuk: a nem-anyanyelvűek nagy része nem ismeri fel a nyelvi hibákat és pontatlanságokat. Ezért a weben található angol bizonyos szempontból „veszélyes". Az angol szavakból álló szövegek könnyen mintául szolgálhatnak ugyanis olyanok számára, akiknek nincs meg az az előismeretük, hogy megállapítsák, „mennyire angol" a szöveg. Így követendő mintákat vélnek sokan felfedezni a web angoljában, amiről most már nyugodtan kimondhatjuk az előzőek következményeként: nagyrészt magyarok, portugálok, litvánok és malájok, argentinok és finnek angolja, nem pedig az angol anyanyelvűeké. Bízni természetesen lehet abban, hogy aki a weben angolul publikál, nyelvi lektorral átnézeti, amit írt angolul – de mondjuk ki azt is: erre igazán kicsi az esély. Leszögezhetjük: az internetre sok olyan anyag kerül föl, amit olyan emberek írtak, akiknek nem anyanyelve az adott nyelv, de erről manapság az olvasót nem szokás tájékoztatni.

Az a következményeknek csak az egyike, hogy gyermekeink jobban fognak hinni a világhálón talált angolnak, mint talán épp saját nyelvtanáruknak – de bízhatunk abban, hogy ez nem így lesz, és az angoltanításnak a presztízsét semmilyen webes tartalom nem fogja megtépázni.

Nyelvtechnológiai szempontból viszont van egy olyan következmény, ami sokkal nehezebben védhető ki, hiszen a mai gépi fordítórendszereknek elsősorban az internet jelenti az igazi kihívást. A legtöbb felhasználó ugyanis a világhálón levő idegen nyelvű információkat fordíttatja le velük legszívesebben. A weben viszont bőven akad pontatlanul írt szó, szerkezet, sőt, mondat. Ennek oka részben az interneten publikálók figyelmetlensége, részben pedig a nyelvnek, és ezen belül is különösen az angol nyelvnek nem anyanyelvi beszélők általi terjesztése úgy, hogy az olvasó mit sem sejt a szöveg nem autentikus voltáról. Hogy milyen elütésekkel kell számolnunk, arra a kérdésre azt mondhatjuk, hogy szinte bármilyennel. Álljon ehhez itt illusztrációként három, bárki által elvégezhető internetes keresés eredménye.

Az egyik az *internet* szó elgépelt alakjainak számát mutatja (4. táblázat), egy másik az angol *have got* kifejezést és annak különféle elgépelt alakjait (5. táblázat), a harmadik pedig egy-két helytelen angol grammatikai minta előfordulását (6. táblázat) mutatja a világhálón. A hibás alakok száma ma általában nagyságrendekkel kevesebb a helyeseknél, ám összességében mégis azt látjuk, hogy jelentős számú olyan szöveg található az interneten, melyben valamely elgépelt alak szerepel.

4. táblázat. Gépelési tévesztések vizsgálata az „internet” szó segítségével

Elírt szó	Előfordulások száma a világhálón
internet	2 460 000 000
internte	67 400
interent	681 000
intenret	116 000
intrenet	193 000
inetrnet	128 000
itnernet	66 400
niternet	47 700
interne.	19 200 000
intern.t	1 940 000
inter.et	19 400 000
inte.net	2 480 000
int.rnet	436 000
in.ernet	522 000
i.ternet	441 000
.nternet	1 150 000

5. táblázat. A „have got” kifejezés különböző elgépelt alakjai különféle keresőkben

	Google	Yahoo	MSN Beta	Lycos	Altavista	WiseNut	HotBot
have got	4 450 000	139 000 000	2 455 447	29 966 017	135 000 000	12 324 316	29 961 320
have gott	1 440	9 530 000	323	2 136 594	421 000	87 649	2 136 594
have ogt	60	54 700	28	11 290	43 100	2 558	11 291
hav got	5 350	2 680	2 938	1 018	2 670	213	1 018
ahve got	2 340	537	540	157	558	47	157
hae got	737	267	1 292	106	267	88	106
haev got	219	45	68	20	47	6	20
ahev got	145	9	24	1	9	3	1
havee got	29	5	6	0	1	0	0

6. táblázat. Néhány helytelen angol nyelvi szerkezet a világhálón

Szokatlan nyelvi fordulat	Előfordulások száma	Példák
do you knows	791	*Do you knows the capital of Canada?* *How do you knows its right?*
has you got	1290	*Has you got sick much?* *Has you got comics?*
I have get	34.600	*I have get rid of a nasty bug I introduced earlier.* *I have get compliments on the shirt when ever I wear it.*
I does not	304.000	*I does not work.* *I does not affect the net.*

A gépi fordítórendszereknek tehát nem csak a helyesen formált mondatok fordításának amúgy is meglevő nehézségeivel kell megküzdeniük, hanem a helytelen mondatok automatikusan alig-alig elvégezhető korrigálásának nehézségével is. Az embert az elütések sokszor nem is zavarják, mert a szöveg számunkra sokszor így is érthető marad. A gépi fordító rendszer azonban nincs abban a helyzetben, hogy megítélje, hogy egyszerű elütéssel, vagy esetleg új szóval találkozott. A fenti példa segítségével mindössze arra próbáltunk rámutatni, hogy tömegesen hozzáférhető anyagaink – jelesül az interneten hozzáférhető szövegek – valóban komoly devianciákat mutathatnak az elfogadott akadémiai nyelvhasználathoz képest. Ha egy mondat a rendszer számára nem érthető, akár kis nyelvhelyességi módosítással azzá tehető.

Összefoglalva: a mai gépi fordítórendszereknek nem csak fordítaniuk kell, hanem „hibatűrőnek" is kell lenniük, hiszen az ilyen rendszerek használatának elsődleges célja a mások által készített, kész szövegek tartalmának megértése, nem pedig az ember által mindig sokkal jobb minőségben elkészíthető teljes fordítás.

4. Konklúzió

A webes szövegek esetében egyébként nehéz is „az angol szövegek nyelvhelyességéről" beszélni. Létezik például az American Heritage Dictionary által ajánlott amerikai lexikális sztenderd, vagy ismeretes az „oxfordi angol" fogalma, de ezek az internetes szövegek gépi elemzésében nem segítenek: igen sokszor nem úgy jelenik meg a világhálón a szöveg, ahogy azt valamelyik szabványos nyelvi leírás előírja. Egy másik következmény, hogy mivel az olvasó előbb-utóbb írni is fog, helyesírási, fogalmazási készségét viszont a sok elolvasott szöveg formája, alakja erősen befolyásolja, és ezek hatására gyakran a sokat látott formulákat fogja használni, még ha azokat nem is anyanyelvi beszélők hozták létre (amiről neki fogalma sincs). Ha viszont az emberek egy része nincs, nem lesz abban a helyzetben, hogy megítélje, mi jó, és mi nem, mit várhatunk a gépi eszközöktől? A fordítórendszerek nem üzenhetik a felhasználónak, hogy azért nem adnak fordítást, mert ez a szöveg nem angol. Megjegyezzük, hogy egy tökéletes chomskyánus grammatika pontosan ilyen intoleráns volna. A felhasználó ugyanis minden angol, vagy angolnak tűnő (!) mondatra fordítást vár, és igazán nem érdekli, hogy a létrehozó nem állt a nyelvtudás magaslatán. Ráadásul a web előtt ülve nincs is

mód a képernyőn levő szöveg megváltoztatására, tehát még ha tudná is a gépi fordítást kérő felhasználó, hogy mi a hiba, akkor sem volna módja javítani. Az esetek legnagyobb részében viszont épp azért kér gépi segítséget, mert maga nincs abban a helyzetben, hogy értelmezze az előtte levő szöveget, tehát a javítást tőle amúgy sem lehetne elvárni.

Marad tehát az eddigieknél is toleránsabb fordítószoftver kifejlesztésének lehetősége, ami viszont azért veszélyes, mert a tolerancia az egyszerű esetekben is komoly félreértelmezésekre ad lehetőséget. Az „internetes kocka el van vetve", a jelenleg ismert megoldásoknak pedig sok, eddig nem is sejtett nehézséggel kell megküzdeniük.

Bibliográfia

1. *Global Internet Statistics.* [http://global-reach.biz/globstats/index.php3]
2. Kilgariff, A., G. Greffenstette. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3) (2003) 333–348.
3. Prószéky G. *A nyelvtechnológia (és) alkalmazásai.* Aranykönyv, Budapest (2005)

Mondatrész-felfedezés önszervező tanulással

Szepesvári Csaba[1], Kálmán László[2], Lukács Ágnes[2,3],Rebrus Péter[2]

[1] MTA SZTAKI, Budapest,
szcsaba@sztaki.hu
[2] ELTE/MTA Elméleti nyelvészet szakcsoport, MTA Nyelvtudományi Intézet
{kalman,rebrus,lukacsag@nytud.hu}
[3] BME Kognitív tudományi tanszék
alukacs@cogsci.bme.hu

Kivonat: A cikkben egy újszerű gépi tanulási módszer elméleti hátterét és az első kísérlet végrehajtását ismertetjük. A módszer lényege, hogy a nyelvmodellben nem valamilyen generatív nyelvtant, hanem pusztán mondattani mintákat feltételezünk, és a rendszert nem elemzési vagy eldöntési feladattal teszteljük, hanem mondatok hasonlóságának felismerésével.

1 Bevezetés

Egy rövid évszázadra feledésbe merült Saussure-nek az a megállapítása, amely szerint az analógia nem a nyelvi változásoknak a szabályszerűtől eltérő formái mögött meghúzódó mechanizmus, hanem a szinkrón nyelvi rendszer összetartó ereje. Azért ismerjük fel egymás megnyilatkozásainak szerkezetét, mert azok más, már ismert kifejezések mintájára, analógiájára épülnek fel, és azért vagyunk képesek megérteni őket, mert a jelentésükre is ugyanez igaz. Ezeket a szerkezeti és jelentéstani mintázatokat a Saussure utáni évszázadban szabályok segítségével próbálták leírni. Pontosabban: az analógiás mintákat a szabályokkal próbálták helyettesíteni a tényleges megnyilatkozások, kifejezések szerkezetének és jelentésének magyarázatához.

1.1 Önszervező tanulás és nyelvtechnológia

A szintaktikai mintázatok felismerése és megfelelő kezelése előfeltétele a legtöbb természetesnyelv-feldolgozással kapcsolatos feladatnak. A szintaktikai mintázatok felfedézésének legkevésbé „elfogult" (biassed), vagyis a legkevesebb előre megadott nyelvészeti kategóriát feltételezõ és legkevésbé elméletterhelt megközelítése az önszervező (unsupervised) tanulás [3], melynek során pusztán a korpuszban elõforduló mondatok szolgálnak a mintázatok kinyerésére és általánosítására. A legtöbb ilyen módszer a nyelvészetbõl ismert disztribúciós elemzés [8] valamilyen változatát alkalmazza a szintaktikai összetevõk és az ezeket összekapcsoló szabályok azonosítására.

A nyelvet önszervező tanulással megközelítő elképzelés olyan nyelvmodellen alapul, mely szerint a nyelvtudás vagy a nyelvtani tudás megfelelő megközelítése nem egy absztrakt gép, amely kizárólag grammatikus mondatokat képes elõállítani, hanem egy olyan képesség, amely a nyelvi hasonlóságok és különbségek felismerésén alapul. Ennek a megközelítésnek a következménye, hogy a működő nyelvmodell szükségképpen alkalmas arra, hogy újszerű mintázatokról megállapítsa, hogy újszerűek, illetve, hogy az újszerű mintázatokat a korábbi mintázatok tudása alapján beépítse, azaz a nyelvmodell képes önszervező módon továbbfejleszteni „saját magát".

Kiinduló feltételezésünk szerint azért ismerjük fel egymás megnyilatkozásainak szerkezetét, mert azok más, már ismert kifejezések mintájára, analógiájára épülnek fel, és azért vagyunk képesek megérteni őket, mert a jelentésükre is ugyanez igaz. Az analógiák mögött rejtőzkődő szerkezeti és jelentéstani mintázatokat a nyelvészeti elméletek többsége a tényleges megnyilatkozások, kifejezések szerkezetének és jelentésének magyarázatára hivatott szabályok segítségével próbálja leírni. A mi kísérletünknek azonban nem célja sem a grammatikus megnyilatkozások nyelvtani jellemzése, „elemzése", sem egy mögöttes „nyelvtan" rekonstruálása; a modellel szembeni elvárás „mindössze" annyi, hogy a gyakori mintázatok kinyerése után a alkalmas legyen analógiás problémák (aránypárok) megoldására.

1.2 Önszervező tanulás és nyelvelsajátítás

Az önszervező gépi nyelvtanulás eredményei a nyelvelsajátítás-kutatók számára is érdekesek lehetnek. A szakértők egyetértenek abban, hogy az anyanyelv elsajátítása az önszervező tanuláshoz hasonlóan explicit tanítás hiányában történik; a vita a tanuló mechanizmusok jellegéről és a kiinduló reprezentációk gazdagságáról és absztraktságáról folyik. Az önszervező modellek megjelenése a számítógépes nyelvmodellezésben párhuzamba állítható a nyelvelsajátítás-kutatás újabb eredményeivel és modelljeivel. Ezek a generatív nyelvészeti hagyományból kinőtt, veleszületetten absztrakt szabályokat és reprezentációkat és bonyolult algoritmusokat feltételező elméletek helyett egyszerű és a nyelv kognitív tartományán kívül is általánosan működő statisztikai tanulómechanizmusokkal magyarázzák a komplex nyelvtan elsajátítását. Ahogy a nyelvészetben egyre nagyobb teret kapnak az új, lexikalista irányzatok — nemcsak a generatív táborral szemben, hanem azon belül is — a pszicholingvisztikában is az asszociatív tanulásra építő kutatási irányba tolódik el a kutatások vonala a nativistától az empirista szemlélet felé.

A nyelvelsajátítás a mai empirista felfogásban leginkább a nyelvi adatok és érzékleti bemenetek alapján történő szerkezet-elvonatkoztatás: a veleszületett tudás leginkább perceptuális primitívekből, szerkezetabsztraháló eljárásokból és a nyelv elemzésére és produkciójára szolgáló műveletekből áll. Bár egy tanulóalgoritmus sikeressége egy adott területen semmi esetre sem jelenti azt, hogy a gyerek is ugyanazzal a mechanizmussal sajátítja el az érintett nyelvi jelenséget, annyit azért elárul, hogy az adott bonyolultságú jelenség megtanulásához nem szükségszerű bonyolultabb procedúrát feltételeznünk. Az algoritmus sikertelenségéből pedig arra következtethetünk, hogy a gyerek vagy gazdagabb inputhoz fér hozzá, vagy erősebbek a kezdeti megszorításai [9].

A felügyelő nélküli tanulás egyik alapmechanizmusa a strukturalista nyelvészeti hagyományokban gyökerező disztribúciós elemzés, amely általában a nyelvi elemzés különböző szintjein címkézett elemeken (fonémák, morfémák, szavak és nyelvtani kategóriák, frázisok) történhet — ilyenkor persze nem a címkéket tanulja az algoritmus, hanem valami mást: például nyelvtani kategóriák szerint címkézett korpuszon a jelentést (szemantikát). Vannak azonban olyan algoritmusok is, amelyek címkézetlen korpuszokból, pusztán a nyelven belüli disztribúciós információból próbálnak nyelvtant tanulni. A disztribúciós elemzést általában még különböző általános statisztikai elvek egészítik ki: ilyen lehet például a súlyok valószínűségének bayesi elvek alapján történő maximalizálása, a leírás hosszának minimalizálása (válaszd azt az elemzőt és elemzést, amelyek hosszainak az összege minimális), vagy a maximum entrópia elve (az elemzési kategóriák eloszlása legyen a megszorításokkal összeegyeztethető lehető legegyenletesebb). A tanulás felügyelő nélkül történik, vagyis nincsen explicit segítség vagy visszajelzés; az elsajátítás csak közvetlenül a kijelentésekből (vagy írott szövegből) történik. A címkézetlen korpuszokon való ilyen tanulás mindenképpen hasznos lehet a nyelvelsajátítás-kutatás számára, mivel a tanulóalgoritmusnak elég explicitnek kell lennie ahhoz, hogy számítógépes program formájában implementálható és tesztelhető legyen, és így fényt deríthet arra, mi az, aminek nem előfeltétele a veleszületett tudás.

1.3 Az ADIOS rendszer

A pusztán disztribúciós elemzésre építő algoritmusok valójában számos területen meglepően sikeresnek bizonyultak. Kizárólag az együttes előfordulások statisztikájára építve eredményesen tanulnak szóosztályokat: feljegyzik azt a kontextust, amiben egy szó előfordul, majd hasonlóságot számolnak és klaszteranalízist végeznek, amelyek segítségével egybe csoportosítják azokat a szavakat, amelyek hasonló kontextusokban fordulnak elő.1 A módszer nem eleve megadott szintaktikai kategóriák szerint csoportosítja a szavakat, a létrejövő klaszterek azonban a megfelelő szinten címkézve jól megfelelnek a nyelvészeti szintaktikai kategóriáknak.

A nyelvelsajátítás felügyelő nélküli modelljei közül az ADIOS (Automatic Distillation of Structures [6] [7] [18], más hasonló megközelítésről Alignment Based Learning: [20]).

A modell nagyon jó eredményeket ért el a nyelv különböző aspektusainak elsajátításában annotálatlan korpuszokon. A modell nyelvtani szemlélete a kognitív nyelvészet és a konstrukciós nyelvtan hozzáállását tükrözi, vagyis a nyelvtan maga a nyelvi egységek listája, fokozatos általánosságot, komplexitást és absztrakciót mutató mintázatok gyűjteménye [11]. A hasznos nyelvi egységek meghatározásának eszköze ennek megfelelően a disztribúciós elemzés: azoknak a mondatoknak az azonosítása, amelyek osztoznak bizonyos szósorokban, de egy helyen paradigmatikus variabilitást mutatnak. Ez a rendszer két alapvető építőköve: a mintázat (vagy szintagma) és a változatosságot mutató helyen előforduló komplementáris disztribúcióban álló szimbólumok ekvivalenciaosztálya.

Az ADIOS tehát címkézetlen elemeken, alulról fölfelé építkezik, reprezentációs ereje három elven nyugszik: 1) azok a mintázatok fontosak, amelyek kevés szóval gyakran előfordulnak és jól általánosíthatók; 2) mivel a változatosság csak a mintázat

által meghatározott kontextusban lehet, ez a kontextusérzékeny általánosítás biztonságosabb, mint egyetemes szófajok vagy szabályok alapján; 3) a komplex mintázatok rekurzívan, hierarchikus mintázatban épülnek fel.

Az ADIOS empirikusan és automatikusan magától fedezi fel a nyelvi építőelemeket. Az egyik tanítókorpusza a CHILDES [5] [15] gyerekekhez szóló, 300000 mondatot tartalmazó beszédátirata volt. 14 nap alatt 3400 intuitíven is fontos mintázatot és 3200 szemantikailag megfelelő ekvivalenciaosztályt talált. Az új mondatokat létező mintázatok megosztott reprezentációjaként alakítja ki, és ezt használta fel azokban a tesztekben, amelyeken új inputokat kellett kezelnie. 10000 mondatos tréning után 1000 új CHILDES mondatnak illetve ugyanezek véletlen szórendű változatának elfogadhatóságát kellett megítélnie, ebben a feladatban már kevés tréning után is jól teljesített, és ugyanez igaz volt a távoli függőségi viszonyok (a modell szempontjából beágyazott ekvivalenciaosztályok) kezelésében. Szokatlan módon a teljesítményét az emberivel is összehasonlították. Egy fejlődési nyelvmegértési tesztben, ahol a feladat mondatok helyességének a megítélése volt, egy 8-8,5 éves gyerek szintjén teljesített. Kényszerválasztásos angol nyelvvizsgateszten 60%-ot teljesített (a véletlen 33% lett volna), és egy hasonló elfogadhatósági tesztben is nagyon hasonlít teljesítménye az emberére.

Összefoglalva, az ADIOS-algoritmus hátterében álló elvek:

1. a mintázat fontosságának probabilisztikus inferenciája
2. kontextusfüggő általánosítás
3. a komplex mintázatok rekurzív felépítése

2 A felhasznált módszer

2.1 Elméleti alapok

Az általunk követett eljárás egy fontos vonásban különbözik az ADIOS filozófiájától, és ennek messzire ható következményei vannak. Mi ugyanis a nyelvészet fő irányzataival szemben azt gondoljuk, hogy a nyelvtannak és a nyelvtudásnak nem az a legfontosabb célja, hogy a „jó” mondatokat a „rossz” mondatoktól elkülönítse, hanem hogy a kommunikáció céljait szolgálja, vagyis nyelvi formákhoz nyelvi jelentéseket társítson. Bár a jelentéssel egyelőre nem foglalkozunk, ennek a megközelítésnek az a közvetlen következménye, hogy a nyelvtan nem a „jó” mondatok halmazának rekurzív megadására törekszik. Ez mind a feltárni kívánt mondatszerkezetekre, mind a rendszer lehetséges tesztelésére vonatkozó megfontolásokra kihat.

A szokásostól eltérően nem feltételezzük a mondatszerkezetek szigorúan hierarchikus felépítését, tehát még közvetve sem egy újraíró szabályrendszert próbálunk rekonstruálni a korpuszból. Az újraíró szabályrendszerek a nyelv formális nyelvi modelljeiből származó, lényegében ma is egyetlen hatékony eszközei annak, hogy a nyelvet mint mondathalmazt jellemezzük. Mivel nem ilyen jellemzés a célunk, nem szükséges ezt az eszközt használnunk. Az eljárás során felismert minták átfedhetik egymást, az alá-fölérendeltségi viszonyokban nem szükséges döntenünk, sőt, azt is toleráljuk, ha a mondatokban nem minden elemet tudunk lefedni. Ez a robusztusság

szempontjából is fontos (az ismeretlen szavak a legtöbb esetben nem okoznak felismerési problémát).

Ami a tesztelés lehetőségét illeti, a mondatszerkezet felismerését a fenti okokból nem tudjuk azzal tesztelni, hogy „jól tudunk-e elemezni" korábban még nem látott mondatokat. Ehelyett olyan teszteket alkalmazunk, amelyek azt mutatják meg, hogy észreveszünk-e a tanultak alapján fontos hasonlóságokat korábban még nem látott mondatok között. [19].

Az eljárás megvalósításához elsőként azoknak a hasonlóságoknak a körét rögzítjük, amelyek elvileg relevánsak lehetnek. (Ez nyelvi szintenként eltérő lehet: például a mondattanban az *ab* és a *ba* sorozatok hasonlónak minősülhetnek, míg az alaktanban nem — a magyar morfológia ilyen elemzését ld. [10]). Ezek képezik az adott nyelvi szintről alkotott „elfogultságainkat".

A második lépésben egy gyermeknyelvi korpusz alapján (CHILDES, [5] [15]) a valóságosan relevánsnak minősülő hasonlóságokat kerestük több módszer összehasonlításával. A relevancia mutatójaként elsődlegesen a gyakoriságot és másodlagosan információelméleti kritériumokat használtunk: a gyakori hasonlóságok egyúttal gyakori különbségeket is jelentenek, hiszen ha mondatok gyakran hasonlítanak egy bizonyos tulajdonságban, akkor ennek a komplementerében gyakran különböznek. A feldolgozásnak ez a fázisa sok más kísérletre is jellemző [14] [18] [20], de a célkitűzés minden esetben eltér a mienktől, mert valamilyen mögöttes kategóriarendszer vagy nyelvtan rekonstruálására irányul. A megközelítésünkben egyedi, hogy a mintázat-jelöltek hatékony felfedezésére hatékony adatbányászati algoritmusokat (pl. A PRIORI [1] [2]) használtunk — nagyságrendekkel csökkentve a futásidőt a korábbi egyszerűbb eljárásokhoz képest [18] [20].

Végül a gyakori különbségekből és mintázatokból olyan adatbázist építünk, amelynek használatával megoldhatók a minket érdeklõ, A:B = C:x alakú aránypárok, ahol az aránypárok tagjai mondatok és mondatrészek. A nyelvtudásnak ilyen feladatokkal való tesztelése nem ismeretlen sem a mindennapi életben (pl. GRE-tesztek), sem a számítógépes nyelvészetben [12] [19]. Például a Mary is sleeping : Mary = Joe left : *x* aránypárnak *x* = Joe a megoldása, amit úgy is meg lehet fogalmazni, hogy a rendszer „felismeri az alanyt", anélkül, hogy akár mondatszerkezetet, akár mondatrész-címkéket rendelne a kifejezésekhez. Az aránypárok megoldásához feltesszük: az $A : B$ különbségnek maximálisan relevánsnak kell lennie, és meg kell egyeznie a $C : x$ különbséggel, és ugyanígy fontosnak és azonosnak kell lennie az $A : C$ és a $B : x$ különbségeknek is. Előadásunkban a módszerek bemutatása mellett összehasonlítjuk az általunk vizsgált különböző önszervező tanulási algoritmusok sikerességét ilyen analógiás feladatok megoldásában.

2.2 Az algoritmus

Az algoritmus többféle lépés iterált alkalmazásából áll, ami mindaddig folyik, amíg „érdekes jelenségeket" fedezünk fel a korpuszban. „Érdekes jelenségeknek" minősülnek a következők:

1. Gyakran ismétlődő sorozatok („kollokációk”). Ha egy sorozat kis különbségekkel ismétlődik gyakran, és a különbségek nem túl sokfélék, akkor is „kollokációról” beszélünk.
2. Érdekes kontextusok. Azok a kontextusok számítanak érdekesnek, amelyekben elég sokféle egység fordul elő ahhoz, hogy feltételezzük, egy osztályt alkotnak (ld. 3.), de nem olyan sokféle, hogy az együttes előfordulás irrelevanciájára gyanakodjunk. Itt „egységen” szavakat, szóosztályokat és ezekből alkotott gyakori sorozatokat egyaránt értünk.
3. Azonosan viselkedő osztályok. Azokról a szavakról, szóosztályokról, sorozatokról, amelyek nagyjából ugyanazokban az „érdekes” kontextusokban fordulnak elő, feltételezzük, hogy ugyanabba az osztályba tartoznak, legalábbis az illető kontextus szempontjából.

Minden iterációs lépésben átírjuk a kiinduló korpuszt a felfedezett jelenségek felhasználásával. Ez egyfajta címkézés, csak annyiban sajátos, hogy a következő lépésekben a címkék és a korábbi információ egyaránt a felfedező eljárás tárgyát képezi. Az algoritmus visszalépési lehetőséget is tartalmaz: ha azt tapasztaljuk, hogy „túláltalánosítottunk”, azaz egy osztálynak elég nagy részosztályai nem viselkednek egyöntetű módon, akkor részekre bontjuk az osztályt, és visszalépünk a keletkezésekor érvényes korpuszhoz. Osztályok egyesítését azonban nem engedjük meg (ha már különböző viselkedést tapasztaltunk, azt nem tekinthetjük meg nem történtnek).

2.3 Eredmények

A felfedező eljárás eredményét kétféleképpen teszteltük: aránypárok megoldásával és mondatkiegészítési feladatokkal. Az aránypárok megoldása bonyolultabb eljárás, de jobban megfelel az elméleti alapfeltevéseinknek, amelyekről fent szóltunk. A mondatkiegészítés egyszerűbb próba, viszont feltételezi, hogy van valamilyen módszerünk annak ellenőrzésére, hogy két mondat közül melyikre tudunk „elfogadhatóbban” felfedezett mintákat illeszteni.

A tesztelésben az az „illesztőprogram” játssza a középponti szerepet, amely a korpuszban felfedezett jelenségeket megpróbálja még nem látott mondatban felfedezni. Az illesztéseken egy kvázi-algebrai struktúrát definiáltunk (metszet, különbség), valamint egy „lefedési mértéket”, amely megadja, hogy mennyire „részletesen” fedi le az illető illesztés a mondatot.

Az $A : B = C : x$ alakú aránypárok megoldásánál (ahol x értékét megadott kifejezések közül kell kiválasztani) a következőképpen döntünk: az $A - B$ és a $B - A$ különbségekhez hasonlítjuk a $C - X$, illetve $X - C$ különbségeket (ahol X a vizsgált jelölt). Az összehasonlítás azt jelenti, hogy a metszetek nagyságát hasonlítjuk össze. Az lesz a nyertes jelölt, amelynél ezek a különbségek a legjobban hasonlítanak (a két metszet nagyságának összege a legnagyobb). A kiegészítéses feladatnál a lehetséges jelöltekkel kiegészített mondatokon futtatjuk az illesztőprogramot, és az eredmények „lefedési mértékét” hasonlítjuk össze. A „lefedési mérték” annál nagyobb, minél több szóra minél hosszabb sorozatokat tudtunk illeszteni. (A szóosztályba sorolás 1 hosszú sorozatnak minősül.)

A cikk leadásának órájában még csak megkezdtük a nagyobb korpusszal való kísérletezést. A „játék” korpuszunkból, amelyben csak névelőt, 50 főnevet, 50 melléknevet és 25 igét használtunk, és ezek minden nyelvtanilag helyes kombinációját be-

vettük (a nyelvtan természetesen „Det (A) N V” volt), tanulás után a rendszer 100% pontossággal képes volt megoldani az aránypáros és a kiegészítős feladatokat.

Bibliográfia

[1]Agrawal, R., T. Imielinski, and A. N. Swami (1993) Database Mining: A Performance Perspective. IEEE Trans. Knowl. Data Eng. 5(6): 914-925

[2]Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499

[3] Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121

[4] Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438

[5] http://childes.psy.cmu.edu/

[6] Edelman, S., Solan, Z., Horn, D. and E. Ruppin (2003) Rich Syntax from a Raw Corpus: Unsupervised Does It; a position paper to be presented at Syntax, Semantics and Statistics; a NIPS-2003 workshop, Whistler, BC.

[7] Edelman, S., Z. Solan, D. Horn and E. Ruppin (2004) Bridging computational, formal and psycholinguistic approaches to language. to appear in Proc. of the 26th Conference of the Cognitive Science Society, Chicago, IL, Aug. 2004.

[8] Harris, Z. S. (1954). Distributional structure. Word, 10:140–162.

[9] Johnson, Mark and Riezler, Stefan (2002) Statistical models of syntax learning and use. Cognitive Science, 26, 239-253.

[10] Kálmán, L., P. Rebrus and M. Törkenczy (2005): Hungarian linking vowels: An analogy-based approach. Paper presented at the 7th International Conference on the Structure of Hungarian Veszprém, Hungary, May 29-31, 2005

[11] Langacker, R. W. (1987) Foundations of cognitive grammar: Theoretical Prerequisites. Stanford, CA: Stanford University Press.

[12] Lepage, Y. (1998) Solving Analogies on Words: an Algorithm. Proceedings of COLING-ACL'98, Montréal, August vol. I, pp. 728-735.

[13] van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)

[14] Lieven, E., Tomasello, M., Behrens, H. Speares, J. (2003) Early syntactic creativity: a usage-based approach. Journal of Child Language, 30, 333-370

[15] MacWhinney, B. and C. Snow (1985) The child language exchange system. Journal of Computational Lingustics, 12:271–296.

[16] Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)

[17] Pedersen, B., Edelman, S., Solan, Z., Horn, D., and E. Ruppin, (2004) Some Tests of an Unsupervised Model of Language Acquisition, in Proc. COLING-2004 Workshop on Psycho-computational Models of Human Language Acquisition, Geneva, Switzerland.

[18] Solan, Z., Horn, D., Ruppin, E. and Edelman, S. (2003) Unsupervised Context Sensitive Language Acquisition from a Large Corpus; in Proc. NIPS-2003..

[19] Turney, P.D., and Littman, M.L. (2005), Corpus-based learning of analogies and semantic relations, Machine Learning, 60 (1-3), 251-278.

[20] van Zaanen, M. M. (2003) Theoretical and Practical Experiences with Alignment-Based Learning. Proceedings of Australasian Language Technology Workshop, Melbourne, Australia, December.

Hunpars: mondattani elemző alkalmazás

Babarczy Anna[1], Gábor Bálint[1], Hamp Gábor[2], Kárpáti András[3], Rung András[4], Szakadát István[2]

[1] Kognitív Tudományi Tanszék, BME, 1111 Budapest, Stoczek u. 2.
{babarczy, bgabor}@cogsci.bme.hu
[2] Szociológia és Kommunikáció Tanszék, BME, 1111 Budapest, Stoczek u. 2.
hampg@eik.bme.hu, syi@axelero.hu
[3] Klasszika-filológia Tanszék, PTE, 7624 Pécs, Ifjúság útja 6.
karpati.andras@t-online.co.hu
[4]Nyelvtudományi Intézet, MTA-ELTE, 1068 Budapest, Benczúr Gy. u. 33.
runga@artitude.hu

Kivonat: A Hunpars-projekt célja egy nyílt forráskódú elemző alkalmazás létrehozása, amely automatikusan végzi el bármilyen értelmezhető magyar mondat szintaktikai elemzését, konkrétan a mondatot alkotó szócsoportok és azok egymáshoz való viszonyának azonosítását. Az elemzőt egy többkomponensű rendszer részeként képzeljük el: a fejlesztés alatt álló modul bemenete egy előzőleg tokenizált mondat, amelyben a szavak morfológiai jegyeikkel felcímkézve szerepelnek. A szintaktikai elemzés szabályalapú: elsősorban egy szintaktikai kategóriákra épülő frázis-struktúra nyelvtan és kiegészítésként különböző lexikális tárak felhasználásával valósul meg. Az alkalmazást irodalmi, jogi, tudományos-ismeretterjesztő és sajtószövegből származó, kvázi-véletlenszerűen kiemelt mondatokon teszteltük. A tesztmondatok 72%-ára helyes elemzést kaptunk, további 11% elemzésének hibája szótári hiányosságra vezethető vissza.

1 Bevezetés

A következőkben egy magyar nyelvre alkalmazható, mondattani elemző működését mutatjuk be. Az alkalmazás fejlesztése 2003-ban kezdődött, és eredetileg egy kérdésmegválaszoló rendszer[1] egyik moduljának készült. Később az elemző fejlesztése függetlenedett az eredeti projekttől, így ma már nem csak egyszerű kérdő mondatok, hanem bármilyen magyar nyelvű mondat elemzésére is használható. Azt a cél tűztük ki tehát, hogy létrehozzunk egy olyan nyílt forráskódú alkalmazást, amely magyar nyelvű természetes mondatok frázisainak és a frázisok közötti viszonyoknak az azonosítását végzi el automatikusan, kézi beavatkozás nélkül.

A mondattani elemzés elsődleges feladata, hogy egy mondatban ne csak az összetartozó szócsoportokat azonosítsa, hanem meghatározza a viszonyokat az egyes szavak és a szavakból alkotott szerkezetek közt is. Ez a cél megvalósítható szigorúan

[1] *Szavak hálójában*, Budapest Műszaki és Gazdaságtudományi Egyetem és Axelero Rt. (jelenleg T-Online Rt.) projekt NKFP és OM támogatással.

lexikalista alapon is [5], amikor a szintaktikai szerkezetet a szavak közötti kapcsolatok határozzák meg. Erre a megközelítésre példa a magyar GeLexi-projekt, ahol a szintaktikai elemzés alapja a gazdag szóleírásokat tartalmazó lexikon, amely megadja az egyes szavak kapcsolódási lehetőségeit [1].

Egy másik lehetséges – és általunk is választott – megközelítésben az elemző a frázisstruktúra nyelvtanokhoz hasonlóan a szavakat hierarchikus szerkezetekbe, frázisokba szervezi, és ezt követően a viszonyokat már ezek között a hierarchikus szerkezetek között határozza meg (a módszer áttekintésére lásd [2]). Például az (1) mondat szerkezetét a (2)-es zárójelezett változat jeleníti meg.

1. Az előadás után meglehetősen leverten álltam a lepusztult mozi előtt.
2. [[Az előadás] után] [meglehetősen [leverten]] [álltam] [a [lepusztult [mozi]]] előtt].

Minden frázisnak (zárójelezett egységnek) van feje, amely egy olyan szó, amely meghatározza a frázis viselkedését a mondatbeli hierarchia következő szintjén. A frázisstruktúra nyelvtanok kiegészíthetők lexikális függőségi információval [8]. A mondat szerkezetének helyes elemzéséhez szükségünk van az adott nyelv módosítóinak nyelvtanára is (például hogy az *előtt* névutó jelölhet egy szabadon előforduló hely- vagy időhatározót), és a régensek szubkategorizációs követelményeire, azaz hogy milyen argumentumai lehetnek egy adott régensnek, amelyek jelenlétében az adott szerkezet jól formált lesz. Egy ilyen típusú komplex nyelvtant meghatározhatunk lexikális és általánosított konstrukciós minták halmazával (Kálmán et al. 2003) vagy frázisstruktúrákat létrehozó szabályok egymás után való rendezésével és lexikális függőségi adattárak alkalmazásával.

A Hunpars alkalmazás az utóbbi eljárást használja kisebb módosításokkal, mint azt a következőkben részletezzük. A megközelítésünkhöz hasonló, de csak szócsoportok azonosítására fókuszáló kutatások folynak a Nyelvtudományi Intézetben [9], [10], illetve ide sorolható a szintén szabályokat létrehozó, de automatikus módszereket alkalmazó HumorEsk [6] és Hócza András kutatásai [3].

2 A Hunpars felépítése

Projektünk nem törekszik arra, hogy egy adott elméletet minél hívebben adaptálva hozzon létre egy mondattani elemzőt. A elemző tervezésénél a nagy lefedettség elérése volt az elsődleges cél, azaz hogy egy viszonylag egyszerű szabályrendszer és néhány jól megválasztott algoritmus rugalmas kombinációjával minél többféle természetes nyelvi mondatot tudjunk elemezni, beleértve az esetleg egyedien formált, pontatlan vagy nem teljes mondatokat is.

A Hunpars alkalmazás tehát három pilléren nyugszik:

- Frázis-struktúra nyelvtan
- Lexikális adattárak
- Elemzési algoritmusok

2.1 A nyelvtan

A nyelvtant kifejezetten a magyar nyelvre fejlesztettük, szem előtt tartva a nyelv gazdag morfológiai rendszerét és az ezzel összefüggésbe hozható variálható szó- és konsztituenssorrendet. A Hunparsnak szüksége van az elemezni kívánt mondat szavainak morfológiai elemzésére, ehhez a Hunmorph morfológiai elemzőt [7] használjuk. A Hunmorph által adott elemzés az egyes szavak szófaji besorolása mellett megadja a szóalakok teljes morfológiai jegyhalmazát is. A *dolgoknak* szóalakhoz így például a következő elemzés tartozik: `dolog/NOUN<PLUR><CAS<DAT>>`. Az elemző nyelvtana tehát elsősorban a morfológiai elemzés kimenetére és másodsorban a szavak lineáris elhelyezkedésére épül.

A frázisok fejének megválasztásakor a lexikalista hagyományokat követjük. A mondattani elemzés során a zárójelezett frázisok alapértelmezésben öröklik a fej jegyeit, illetve egyes esetekben a frázis más alkotószavainak jegyei is öröklődhetnek. Így például *a lepusztult mozi* konstrukció nemcsak a tartalmi *mozi* fej jegyeit, hanem a névelő jegyeit is hordozza.

Ha a mondatban szereplő szavak bármelyike morfológiailag többértelmű, akkor az adott mondatnak ennek megfelelően újabb változatait hozzuk létre, és ezek mindegyikére lefut az elemzés. Azaz, ha egy mondatban három kétértelmű és egy háromértelmű szó található, akkor az adott mondatnak akár $2^3 \times 3$, azaz 24 különböző elemzése is lehet. A többértelműségek nagy része egy statiszikai egyértelműsítő modullal kiszűrhető. Bár ilyen modul jelenleg nem áll rendelkezésünkre, néhány egyszerű előszűrő szabállyal is jelentősen sikerült csökkentenünk a többértelműségek számát. Az előszűrést követően megmaradt mondatváltozatokat a nyelvtan tovább szűri: a szabályrendszert ki nem elégítő változatokat elvetjük.

2.2 Lexikális adatbázisok

A morfológiai jegyeken túl nyelvtanunk lexikális adatbázisokban található információkat is használ. Például melléknevek (*egy fiára büszke anya* , de **fiának büszke *fiával büszke*), névutók (*a házzal szembe*, **a házra szembe*, **a házba szembe*) és a későbbiekben majd igék bővítményeire vonatkozó megszorításokat. A Hunpars használja továbbá az igék és igekötők lehetséges kombinációira vonatkozó információs tárat.

2.3 Elemzési algoritmusok

Első lépésben az összetett mondatokat az elemző tagmondatokra bontja, melyeket külön elemez a továbbiakban, és a folyamat végén ezeket a részelemzéseket egyesíti. Az elemző algoritmus sorrendezett elemzési fázisokból áll. Mindegyik fázis szabályillesztések egy sorozatát és/vagy egyéb algoritmikus lépéseket foglal magában, melyek egy adott frázistipus zárójelezését végzik. A szabályillesztések során a mondatokban jobbról-balra vagy balról-jobbra keresünk olyan szót, mely morfológiai jegyei alapján lehet a keresett frázis feje a mindenkori szabálynak megfelelően. A fej azonosítása után a nyelvtani szabály által meghatározott (kötelező vagy opcionális) egyéb elemeket a fejhez csatoljuk. Ha egy frázis elemeit megtaláltuk, az elemző tovább halad a tagmondatban.

Miután az elemző egy fázison belül zárójelezte a lehetséges frázisokat, továbblép a következő fázisra. Az elemzés későbbi lépéseinél az előzőleg lezárt frázisokat egy egységnek kezeljük.

A keresés irányát a nyelvtan határozza meg, ez a fázisok szabálycsoportjaiban különbözhet. A keresés irányának meghatározó szerepe van. Az irányváltoztatásnak bizonyos rekurzív tulajdonságokat mutató szerkezetek elemzésénél (pl. birtokos szerkezet) van kiemelt szerepe, amelyekben így elkerülhető volt, hogy vermet igénylő rekurzív szabályillesztéseket használjunk.

Az elemző másik fontos eszköze beágyazott szerkezetek kezelésére a lezáratlan frázis funkció. A lezáratlan frázisok elemei az elemzés későbbi lépései során (speciális szabályok segítségével) még bővíthetők.

A nyelvtan szükség esetén – mint láttuk – a lexikai adatbázisokhoz fordul.

A Hunpars a következő elemzési fázisokat tartalmazza.

- Előfeldolgozás: morfológiai egyértelműsítés
- Előfeldolgozás: tagmondatokra bontás
- Az igei frázis és a tagmondat régensének felismerése
- Határozószói frázisok elemzése
- Számnévi frázisok elemzése
- Melléknévi frázisok elemzése
- Főnévi frázisok elemzése
- Névutói frázisok elemzése

Bizonyos fázisokat követően egy mellérendelői szerkezeteket azonosító fázis is lefut az adott szinten lévő mellérendelések azonosítására. Ez a fázis akkor ismer fel egy mellérendelést, amikor a kötőszó előtt és után álló mellérendelő viszonyban frázisok már elemzésre kerültek. Például mellérendelés a *piros labda és kék szalag* kifejezésben csak a főnévi frázisok azonosítása után jöhet létre, míg a *piros és kék labda* esetében már a melléknévi szakasz után azonosítható.

A fázisok lefutása során az azonosított fejekkel egy frázisba kerülnek módosítóik és bővítményeik egy hierarchikusan szervezett frázisstruktúrát alkotva.

A fázisok jellegének szemléltetéséhez az alábbiakban bemutatjuk a szabályok leírására alkalmas formalizmus egy rövid kivonatát és példaként a melléknévi fázis leírását:

`A ::= B1 B2 … Bn`

B_1 B_2 … B_n egymás után alkosson A frázist.

`A ::= { B1 B2 … Bn }`

B_1 B_2 … B_n bárhogy elhelyezkedve alkosson A frázist.

`Phase(<name>):`

A következő szabályok a <name> nevű elemzési fázishoz tartoznak.

`group:`

Egy szabálycsoport szabályai következnek, a következő `group:`-ig, vagy a következő fázis kezdetéig. Egy szabálycsoporton belüli szabályokat egy keresés során illeszthetünk, melynek iránya alapértelmezésben jobbról balra. Ha egy szabálycsoporton belül több szabálynak is ugyanaz a feje, és a mondat egy összetevője fejként mind a két szabályillesztést lehetővé tenné, akkor azt az illesztést kell végrehajtani, amelyik hosszabb frázist eredményez. A group kulcsszó után zárójelben a szabályok fejének kereséséről adhatunk meg paramétereket:

`group(left2right)`: a keresés balról jobbra haladjon

`group(unfinal)`: a fejet lezáratlan frázisokban keresse

Szögletes zárójelben (`[]`) adhatunk meg a szabály használatára vonatkozó feltételeket, ezek lehetnek tokenszintűek vagy szabályszintűek. Tokenszintű esetén a token után rögtön szögletes zárójelben szerepel a rá vonatkozó feltétel, szabályszintű esetén a szabály után található egy több tokenre vonatkozó feltétel. A szabályszintű feltételben úgy hivatkozhatunk a tokenekre, hogy azokat egy / jellel és egy számmal megindexeljük.

Pl.:

tokenszintű: `AdjP ::= Adv[canModify(Adj)] Adj`

szabályszintű: `AdjP ::= Noun/1 Adj/2 [arg(1,2)]`

Alapértelmezésben a létrejövő frázis feje a jobb oldalon a legszélen lévő elem, egyéb esetben a fejet `/HEAD`-del jelölhetjük meg.

A tokenek számának meghatározására a reguláris kifejezésekben megszokott jelöléseket használjuk.

A szabály végén kettőspont után utalhatunk a szabályhasználat miként jére vagy következményeire . Ha nem szerepel semmi, akkor a szabályt nem illeszthetjük újra a létrejövő frázisra, ha `:repeat` szerepel, akkor rekurzívan többször is illeszthetjük, ha `:break`, akkor ezt és más a fázishoz vagy szabálycsoporthoz tartozó szabályt sem illeszthetünk többet, ha `:error` akkor az elemzésünk hibajelzéssel leáll. Ha `:rel(<reláció>)` szerepel a szabály után, akkor a megadott relációt a szabály alkalmazhatósága esetén igazra kell állítanunk. `:call(<eljárás>)` esetén a megadott a szabály alkalmazása után végre kell hajtanunk a megadott eljárást, `:split` esetén a szabály illesztésekor két elemzési variánst kell létrehozni: egyikben illesztjük a szabályt, másikban nem.

Ha a `::=` jel bal oldalán nincs semmi, akkor illesztés esetén nem jön létre új frázis, de a szabályhasználat megadott következményeit végre kell hajtanunk.

A melléknévi fázis rövid leírása:

```
Phase('adj'):

group:

AdjP ::= Adj
```

Az -ú/-ű végű melléknév előtt nem -ú/-ű végű melléknév állhat:

```
AdjP ::= Adj[adjType!=_U] Adj[adjType==_U] :repeat
```

Nem -ú/-ű végű előtt bármilyen melléknév állhat:

```
AdjP ::= Adj Adj[adjType!=_U] :repeat
```

Melléknevet módosító határozószó lehet melléknév előtt, de ez nem ismételhető szabály:

```
AdjP ::= Adv[canModify(Adj)] Adj
```

Ha a melléknév igenév, akkor bármilyen határozószó állhat előtte:

```
AdjP ::= Adv Adj[adjType==PART]
```

Főnév vagy melléknév vonzatkerettár ellenőrzéssel szerepelhet, lezáratlan frázist hoz létre, és ilyenkor más szabályt már nem alkalmazhatunk:

```
AdjP/1 ::= Noun/2 Adj/3 [argCheck(2,3)] :unfinal(1)
:break

AdjP/1 ::= Num/2 Adj/3 [argCheck(2,3)] :unfinal(1)
:break
```

3. Teszteredmények és továbblépési lehetőségek

Mivel jelenleg nem áll rendelkezésre különféle szövegműfajokat jól reprezentáló, kézzel elemzett magyar nyelvi korpusz, az elemző tesztelése nem automatizálható, s ezért ezt kézzel kellett elvégeznünk. Erre a célra egymástól független, magyar mondatokból álló korpuszt állítottunk össze kvázi véletlenszerű módon különböző forrásokból. A források között kortárs irodalmi művek, sajtó-, tudományos-ismeretterjesztő és jogi szövegek szerepeltek. Mivel elemzőnk nyelvtana jelenleg nem terjed ki vonatkozói mellékmondatok elemzésére, így az azokat tartalmazó mondatokat eltávolítottuk a korpuszból. Az elemző a teljes korpusz elemzése során a bemeneti mondatokra – illetve ha egy mondathoz több változat is szerepelt, azok mindegyikére – az elemzés eredményét tartalmazó annotált text fájlt hoz létre. Az ellenőrzés megkönnyítésére az text fájlok mellett ezekből létrehozott, grafikusan megjelenített elemzési fákat állítottunk elő; az 1-es ábrán egy ilyen elemzési fa látható. Az ellenőrzést egy nyelvészeti szaktudással rendelkező szakértő végezte, akinek nem volt alapos ismeretei a nyelvtan részleteiről és a Hunpars algoritmusairól.

Az elemző működésének kvantitatív értékelésékor *sikeresnek* minősítettünk egy elemzést, amikor az elemző az elemzési folyamat során legalább egy mondatváltozatot nem utasított el. Azonban ezen elemzéseknek csak az a része minősült *helyesnek*, amelyet a kézi ellenőrzés során is hibátlannak találtunk.

A teszt kvantitatív kiértékelését az 1-es táblázat mutatja be:

1. Táblázat: Tesztelési eredmények

Bemeneti mondatok száma	309
Sikeres elemzések száma	600
Azon mondatok aránya, amelyeknek legalább egy sikeres elemzése volt	97%
Azon mondatok aránya, amelyeknek legalább egy helyes elemzése volt	72%

Az eredmények hibáinak kiértékelése során a helytelen elemzéseket a következő csoportokba soroltuk:

- Helytelen az elemzések lexikális adatbázisok hiányosságai vagy hibái miatt. Ide tartoznak a morfológiai elemző által fel nem ismert vagy rosszul besorolt szavak vagy olyan bővítmények, amelyek a lexikai adatbázisokban nem szerepeltek, illetve idiómák és tulajdonnevek fel nem ismeréséből fakadó hibák (a bemeneti mondatok 11%-a).
- Helytelen elemzések a nyelvtan hibájából kifolyólag (a bemeneti mondatok 17%-a).
- Helytelen elemzés implementációs hiba miatt (a bemeneti mondatok kevesebb mint 1%-a).

Az elemző hibáinak elemzése azt mutatja, hogy teljesítménye jelentősen javítható lenne a lexikális adatbázisok, különösen a névutók lehetséges bővítménytárának pontosításával. További fejlődést a nyelvtan bővítésétől várunk. A hibák vizsgálata megmutatta, hogy a beágyazott igeneves szerkezetek okozzák a problémák nagy részét. Az alábbi példákban zárójelezéssel emeljük ki a nem helyesen elemzett szerkezeteket:

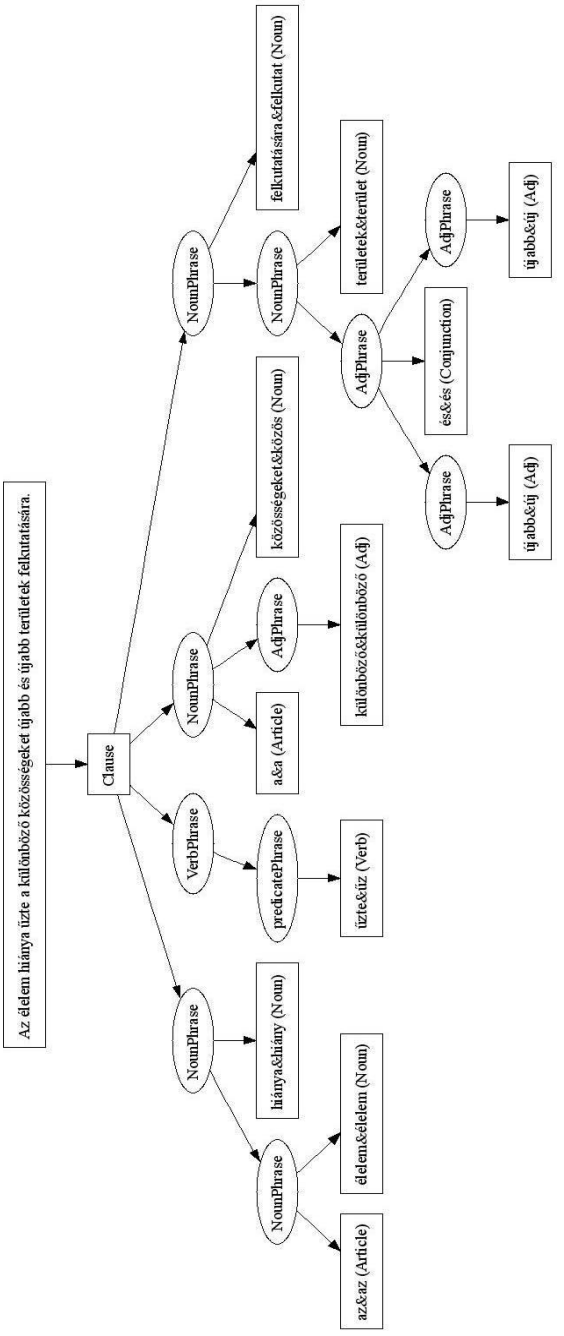
- A férfi a szóbeszéd szerint [egy [a felesége telefonjában talált] SMS] miatt kezdett gyanakodni házastársára.

3. A legnagyobb hazai gyorséttermi lánc múlt vasárnaptól [az egyik fizetős wifiszolgáltatóval együttműködve] drótnélküli internettel csalogatja a fizetőképes keresletet.
4. A vipassana meditáció gyakorlása során a meditáló [[[[a testében megjelenő] pszicho-fizikai jelenségek] természetének] helyes megértésére] törekszik.

További problémát okoznak azok a többértelmű mondatok, amelyek esetében az anyanyelvi beszélő számára egyértelmű, hogy a lehetséges helyes elemzések közül melyiket kell kiválasztani. A Hunpars jelenleg nem tud szemantikai információt felhasználni. Erre mutat példát a 6-os mondat, amelyben a Hunpars helytelenül azonosította a megjelölt frázist:

5. [Az országos televízió főműsoridőben] legalább húsz perc, országos rádió legalább tizenöt perc önálló hírműsort köteles egybefüggően szolgáltatni.

Hosszú távú tervünk, hogy az ilyen ilyen jellegű hibákat egy nagyméretű annotált korpuszon tanított, statisztikai módszereket használó komponens segítségével kerüljük el.



1. ábra *Az élelem hiánya űzte a különböző közösségeket újabb és újabb területek felkutatására.* mondat elemzésének grafikus megjelenítése

Bibliográfia

1. Alberti G., Kleiber J., Viszket A.: Főnévi GeLExi projekt: GEneratív LEXIkonon alapuló mondatelemzés. In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) In: MSZNY (2003) 79–846.
2. Appelt, D.E., Israel D.: ANLP-97 Tutorial: Building information extraction systems. (1997). Available as http://www.ai.sri.com/appelt/ie-tutorial.
3. Hócza A.: Teljes mondatszintaxis tanulása és felismerése. In: In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004)MSZNY (2004) 127–135.
4. Kálmán L., Balázs L., Erdélyi Szabó M.: Tudásalapú természetesnyelv-feldolgozás. In: In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) MSZNY (2003) 109–114.
5. Karlsson, F., Voutilainen, A., Heikkila, J., Anttila, A. (1995): *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
6. Kis B., Naszódi M., Prószéky G.: Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) In: MSZNY (2003) 145–152.
7. Németh L., Halácsy P., Kornai A., Trón V.: Nyílt forráskódú morfológiai elemző. In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004) 163–171.
8. Sag, I., Wasow, T.: *Syntactic Theory: A Formal Introduction*. Stanford : CSLI Publications, (1999).
9. Váradi T., Gábor K.: A magyar Intex fejlesztésről. In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004) 3–10.
10. Váradi T.: Főnévi csoportok annotálása a CLaRK rendszerben. In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) 65–70.

A sz.ot.ag
Optimalitáselmélet szimulált hőkezeléssel

Bíró Tamás

Eötvös Loránd Tudományegyetem, Budapest
University of Groningen, Hollandia
birot@nytud.hu

Kivonat: Az SA-OT algoritmus bevezeti a *szomszédsági struktúra* (*neighbourhood structure*; geometria vagy topológia) fogalmát a jelöltek halmazán. A szimulált hőkezelés egy véletlen bolyongást valósít meg azon a halmazon, amelynek a legjobb elemét keresi, és ehhez a szomszédsági struktúra adja meg azt, hogy egy adott állapotból (SA-OT esetén: jelöltből) mely szomszédos állapotokba lehet lépni. Sőt, a topológiának azt is meg kell határoznia, hogy melyik szomszédot milyen *a priori* valószínűséggel választjuk. A jelen cikk alkalmazza az SA-OT algoritmust a szótagolásra, különböző lehetséges topológiákat hasonlít össze, és azt mutatja be, hogy a topológia definíciója – a többi paraméterhez hasonlóan – jelentősen befolyásolja az algoritmus kimenetelét.

1 Keresés a jelöltek halmazán

Az Optimalitáselmélet (*Optimality Theory,* OT, Prince és Smolensky, 1993 / 2004; magyarul lásd például: Rebrus, 2001) alapfeltevése szerint a mögöttes reprezentációból generált jelöltek halmaza (*candidate set*) optimális eleme (vagy elemei) felel(nek) meg a leírandó nyelv felszíni alakjainak. Az *E* Evaluációs (vagy Harmónia-) függvény egy-egy vektort rendel az egyes *w* jelöltekhez, amelynek *i*-ik komponense az adott jelölt számára az *i*-ik *constraint* (C_i) által kiosztott sértések száma, $C_i(w)$:

$$E(w) = \left(C_N(w), C_{N-1}(w), \ldots, C_0(w)\right) \quad (1)$$

Az OT-paradigmán belül dolgozó hagyományos nyelvész (leggyakrabban fonológus) feladata meghatározni, hogy egy adott jelölt, mint potenciális nyelvi alak, mennyi sértést gyűjt be egy-egy constraint-től (megszorítástól, korláttól) – azaz, milyen constraint-eket definiáljunk, univerzális nyelvi megfigyelések alapján. Mi most ezeket a C_i megszorításokat adottnak vesszük, hiszen a klasszikus OT filozófiája szerint a constraint-ek halmaza univerzális (csak éppen nem jelenleg ismert).

Ha a vektorkomponensek sorrendje megfelel a constraint-ek csökkenő hierarchiájának, akkor az optimalitást a vektorok *lexikografikus rendezése* határozza meg. A szavak ábécésorrendjéhez hasonló módon, két vektor közül azt a vektort tekintjük harmonikusabbnak, amelynek az első nem-azonos komponense kisebb:

Definíció: Legyen w_1 és w_2 a jelöltek halmazának két eleme. A w_1 jelölt akkor és csak akkor *harmonikusabb* a w_2 jelöltnél, ha létezik olyan $k \in \{N, N\text{-}1,\ldots, 0\}$, hogy $C_k(w_1) < C_k(w_2)$, továbbá minden $i \in \{N, N\text{-}1,\ldots, 0\}$-re: ha $i > k$, akkor $C_i(w_1) = C_i(w_2)$.

A w_1 és w_2 jelöltek *ekvivalensek*, ha minden $i \in \{N, N\text{-}1,\ldots, 0\}$-re $C_i(w_1) = C_i(w_2)$.

Azt a C_k megszorítást, amely a w_1 és w_2 jelöltek összehasonlítása során eldönti, hogy melyik jelölt a harmonikusabb, *fatális constraint*-nek fogjuk nevezni. Bebizonyítható formálisan (Bíró, 2006), hogy bizonyos, a nyelvészeti alkalmazásokban általában teljesülő feltételek mellett, a jelöltek halmaza egy *jólrendezett* halmaz, sőt bármely részhalmazának van *optimális részhalmaza*:

Definíció: Legyen *S a jelöltek valamely (rész)halmaza.* Ekkor létezik $S_{ö} \subseteq S$ úgy, hogy (a) ha w_1 és $w_2 \in S_{ö}$, akkor w_1 és w_2 ekvivalensek; (b) ha $w_1 \in S_{ö}$ és $w_2 \notin S_{ö}$, akkor w_1 harmonikusabb w_2-nél. Ezt az optimális részhalmazt $S_{ö}$ = opt (*S*)-sel jelöljük.

Ezek alapján formálisan is megfogalmazhatjuk az Optimalitáselmélet alapfeltevését. Mint oly sok nyelvészeti modell az elmúlt ötven évben, egy *UR* mögöttes reprezentáció és egy *SR* felszíni alak közötti leképezést szeretnénk mi is megadni. Ha *GEN*(*UR*) jelöli az *UR*-ből az univerzális *Generátor függvény* által legyártott jelöltek halmazát, *E*(*w*) pedig az univerzális constraint-ek alapján (1) szerint definiált *Evaluációs függvényt*, akkor az UR-nek megfelelő grammatikus alak(ok) azok, amelyek *GEN*(*UR*)-en optimalizálják az *E*(*w*) függvényt:

$$SR(UR) = \underset{w \in GEN(UR)}{\arg\operatorname{opt}} \big(E(w)\big) \qquad (2)$$

Véges, kevés elemből álló jelölthalmaz esetén az optimális elem megtalálása nem jelent nehézséget. Nagyobb, és főleg végtelen jelölthalmaz esetén viszont az Optimalitáselmélet egy ritkán végiggondolt komputációs kihívást jelent – legyen szó a modell számítógépes implementációjáról vagy a kognitív plauzibilitásáról. Eisner (2000) bebizonyította, hogy az optimális elem megtalálása NP-teljes probléma a nyelvtan méretében. Bizonyos erősen korlátozott feltételek között használhatóak véges állapotú modellek (pl. Ellison, 1994; Frank és Satta, 1998; Gerdemann és van Noord, 2000; Bíró, 2003, 2005a). Dinamikus programozással (Tesar és Smolensky, 2000; Kuhn, 2000) egy jóval tágabb problémahalmazra implementálható az Optimalitáselmélet, de ez utóbbi technika – a véges állapotú automaták megépítéséhez hasonlóan – nem csekély számítási kapacitást igényel.

Kérdés, persze, hogy van-e egyáltalán szükség olyan algoritmusra, amely habár jelentős számítógépes kapacitások (memória, idő) igénybe vétele árán, de garantáltan megtalálja a jelöltek halmazának optimális elemét. A nyelvtechnológia talán még használhatná ezeket az eljárásokat – feltéve, hogy a nyelvtechnológia bármikor hasznosítani fogja az OT-t –, de nehéz elképzelni, hogy biológiai (pszichológiai, kognitív) szempontból adekvátak lenének. Az emberi beszédprodukció ugyanis nem hibátlan, viszont olyan algoritmusra van szükségünk, amely valós időben működik. Beszéd közben nem forgathatjuk a homokórát a képernyőn („bocs, számolok"), de hajlandóak vagyunk a pontosságból áldozni, különösen akkor, ha ezáltal felgyorsítható a számolás. Az SA-OT algoritmus ezt lehetővé teszi.

2 A szimulált hőkezelés alkalmazása az Optimalitáselméletre

Egy házi dolgozatban (Bíró, 1997), a disszertációmban (Bíró, 2006), valamint néhány cikkben (Bíró, 2005b,c) javasoltam a *szimulált hőkezelés* (Kirkpatrick et al., 1983)[2] alkalmazását a legjobb jelölt megkeresésére és a beszédtempó modellezésére. A statisztikus fizikából származó (Metropolis *et al.*, 1953[3]), az elmúlt húsz évben széles körben elterjedt optimalizációs algoritmus előnye ugyanis az egyszerűsége és a kis számítási igénye.

Az SA-OT (*Simulated Annealing Optimality Theory*; 1. ábra) algoritmus ún. heurisztikus technika (például ld. Reeves, 1995), vagyis nem garantálja, hogy megtalálja a keresett legjobb megoldást. Lassú futtatás (sok iterációs lépés) esetén nagy valószínűséggel rátalál a „helyes" megoldásra, míg gyors futtatás (kevés iteráció) mellett könnyebben hibázik. Tehát, akárcsak az emberi beszéd, az algoritmus felgyorsítható, amelynek az árát a pontossággal kell megfizetni. Viszont nem bármilyen hibát követ el az algoritmus: csak bizonyos „helytelen" alakokat talál meg, amelyek egy sikeres modellben épp a gyorsbeszéd jellegzetes alakjainak felelnek meg.

```
algorithm SA-OT
    w := w_init;
    for K = K_max to K_min step K_step
        for t = t_max to t_min step t_step
          choose random w' in Neighbourhood(w);
          calculate <k,d> = | E(w')-E(w) | ;
          if d <= 0 then   w:=w';
          else             w:=w' with probability
              P(C,d;K,t)   = 1         , if k < K
                           = exp(-d/t) , if k = K
                           = 0         , if k > K ;
        end-for;
    end-for;
```

1. ábra: Az SA-OT algoritmus (*Simulated Annealing Optimality Theory*) (Bíró, 2006)

Az SA-OT algoritmus megértése érdekében először tekintsük az egyik legegyszerűbb optimalizációs eljárást. Tegyük fel, hogy egy $E(w)$ függvényt szeretnénk *minimalizálni* egy S alaphalmazon. Definiáljunk egy *topológiát* (egy gráfszerű geometriát, szomszédsági struktúrát) az S alaphalmazon: egy *Neighbour*(w) függvényt, amely S mindegyik w eleméhez hozzárendeli S egy részhalmazát, w szomszédjait. Ezen szomszédsági struktúra lehetővé teszi azt, hogy egy képzeletbeli bolhát (egy *véletlen bolyongást*) indítsunk el S-en, valamely w_{init} elemből indulva.

Egy iteráció két részből áll. Amikor a bolha a w pontban tartózkodik, kiválasztjuk véletlenszerűen w valamelyik w' szomszédját (azaz *Neighbour*(w) egyik elemét), majd eldöntjük, hogy a bolha vajon w'-ba ugrik-e, vagy w-ben marad. Első megközelítésben legyen az a szabály, hogy a bolha akkor és csak akkor ugrik w'-be, ha $E(w') \leq$

[2] Az ELTE fizikus szakán *szimulált hőkezelésnek* neveztük azt a technikát, amelyet a BME-n *szimulált lehűtésként* ismernek. A cikkben az előző elnevezést fogjuk használni.

[3] Érdekesség, hogy az említett cikk utolsó szerzője Teller Ede.

$E(w)$. Az algoritmus addig tart, amíg adott számú iterációt el nem végeztünk, vagy pedig amíg a bolha be nem ragad egy *lokális minimumba*. Ugyanis a lokális minimumnak, definíció szerint, nincs olyan szomszédja, ahova a bolha átugorhatna. Az algoritmus az S alaphalmaznak azon elemét adja vissza, amelyben a bolha végül leledzik. Ha az eljárás elég hosszan fut, az algoritmus kimenetele egy *lokális minimum* lesz – de semmi sem garantálja, hogy megtaláljuk a keresett *globális minimumot*. Sőt, ha rossz „völgyből" indulunk, el sem juthatunk a globális minimumhoz.

A hagyományos *szimulált hőkezelés* (vagy *szimulált lehűtés*, *simulated annealing*), ezzel szemben, lehetővé teszi „hegyek" megmászását is azáltal, hogy a bolha bizonyos valószínűséggel akkor is átugorhat w-ből w'-be, ha $E(w') > E(w)$. Ezt az *állapotátmeneti valószínűséget* egy T paraméter segítségével számoljuk ki:

$$P(w \rightarrow w' \mid T) = \begin{cases} e^{-(E(w')-E(w))/T} & \text{ha } E(w') > E(w) \\ 1 & \text{ha } E(w') \leq E(w) \end{cases} \quad (3)$$

A T paramétert *hőmérsékletnek* hívjuk, utalva a statisztikus fizikai (termodinamikai) gyökerekre. Jelentése az az $E(w') - E(w)$ érték, amennyit a bolha $1/e = 0{,}37$ valószínűséggel ugrik felfelé. Kisebb ugrás valószínűsége magasabb, nagyobb ugrásé alacsonyabb. A szimuláció kezdetén T értéke magas, és a bolha bármely szomszédos állapotba könnyen átugorhat. Majd a T hőmérsékletet lépésről lépésre csökkentjük, a rendszert „hűtjük": a bolha fokozatosan fárad, és csak egyre kisebb „dombokat" hajlandó megmászni. Végül a rendszer „megfagy", és a bolha leereszkedik a legközelebbi völgy aljára. A szimulált hőkezelés sikerét az adja, hogy minél lassabban csökkentjük a hőmérsékletet (vagyis minél több iterációs lépést engedünk közepes hőmérsékleten), annál nagyobb a valószínűsége annak, hogy a bolha a globális minimumot találja meg az algoritmus végén, mert kikerüli a többi lokális minimum csapdáját.

Végül ezt a technikát alkalmazzuk az Optimalitáselméletre. A nehézséget az adja, hogy – a hagyományos szimulált hőkezeléssel ellentétben – az optimalizálandó $E(w)$ függvény nem valósértékű, általános esetben nem helyettesíthető egy valósértékű függvénnyel. Hogyan értelmezzük akkor az exponenciális kifejezést a (3) egyenletben? A különböző megfontolásokból javasolt megoldás (Bíró 2005c, 2006) szerint legyen két, (1) szerinti $E(w)$ vektor „különbsége" a következő rendezett pár:

$$\left|E(w') - E(w)\right| := < k, C_k(w') - C_k(w) > \quad (4)$$

ahol C_k a fatális constraint w és w' jelöltek összehasonlításakor. Vagyis két OT-jelölt *Evaluációs függvényének* különbségét az a legmagasabb C_k korlát adja meg, amelyben eltérnek. A különbség egy rendezett pár, amelynek az első eleme C_k indexe, a második eleme pedig az ezen megszorítás által kiosztott sértések számának a különbsége.

Most már érthetővé válik az 1. ábrán bemutatott SA-OT algoritmus magja. A kettős ciklus belsejében előbb a jelenlegi w állapot egy szomszédos w' szomszédját választjuk ki véletlenszerűen, majd (4) segítségével kiszámoljuk a megfelelő *Eval*-függvények értékének a különbségét. Ha a különbség második tagja nem-pozitív, a bolha átugrik a w-nél harmonikusabb w' állapotba. Ellenkező esetben az átmenet valószínűségét a különbség két tagján kívül a $T = <K,t>$ hőmérséklet is befolyásolja, és a valószínűséget a (3) egyenlet analógiájára számoljuk ki. Ezt az analógiát alaposabban kidolgozza Bíró (2005c, 2006). A $P(w \rightarrow w' \mid T)$ állapotátmeneti valószínűség

jelentése az, hogy generálunk egy *r* véletlen számot 0 és 1 között egyenletes eloszlással, és ha $r < P(w \rightarrow w' \mid T)$, akkor a bolha átugrik *w*-ből *w'*-be.

Már megelőlegeztük azt, hogy az SA-OT algoritmusban a *T* hőmérséklet sem egy valós szám, hanem éppúgy egy rendezett pár, mint $|E(w') - E(w)| = <k,d>$. E két mennyiség azonos szerkezetét megköveteli a (3) egyenlet, illetve annak az interpretációja: *T* jelentése azon ugrás magassága, amelynek a valószínűsége $1/e = 0{,}37$. Az SA-OT algoritmusban szintén igaz, hogy ha a $|E(w') - E(w)|$ és *T* egyenlő (ugyanaz a rendezett pár), akkor a bolha $1/e$ valószínűséggel ugrik felfelé.

Végezetül, a hőmérséklet komplex voltából következik az, hogy a hagyományos szimulált hőkezelés egyszeres ciklusával ellentétben, az SA-OT algoritmus egy kettős ciklus segítségével csökkenti a $T = <K,t>$ hőmérsékletet. A két ciklus kétszer három paramétere (K_{max}, K_{min}, K_{step}, t_{max}, t_{min}, t_{step}) határozza meg az iterációk számát, vagyis a globális minimum megtalálásának a valószínűségét (Bíró 2005b, 2006).

3 A topológia jelentősége

A jelenlegi cikk célja a topológia jelentőségének a bemutatása. A topológiát eddig mint egy *Neighbour*(*w*) függvényt definiáltuk, amely – az SA-OT esetén – a jelöltek halmazát képezi le ezen halmaz hatványhalmazára: minden jelölthöz hozzárendeli a jelöltek halmazának egy részhalmazát. De valójában ennél többre van szükségünk, méghozzá egy *valószínűségi eloszlásra* minden egyes *Neighbour*(*w*) halmazon.

Ugyanis a szomszédsági struktúrát arra használjuk, hogy a bolha jelenlegi *w* helyének egy *w'* szomszédját véletlenszerűen kiválasszuk. A $P(w'|w)$ valószínűség fogja megadni azt, hogy milyen valószínűséggel választjuk *w'*-t következő potenciális ugrási célpontként, feltéve, hogy $w' \in Neighbour(w)$. A $P(w'|w)$ valószínűséget mostantól *a priori valószínűségnek* nevezzük, ellentétben a korábban, például (3)-ban, definiált $P(w \rightarrow w' \mid T)$ *állapotátmeneti valószínűséggel*. Annak az esélye, hogy a bolha a következő iteráció során a *w'* pontban lesz, feltéve, hogy most *w*-ben található, e két valószínűség szorzata: előbb a $P(w'|w)$ *a priori* valószínűséggel választjuk ki *w'*-t, majd a $P(w \rightarrow w' \mid T)$ állapotátmeneti valószínűség határozza meg, hogy a bolha ténylegátugrik-e oda.

E két valószínűség egymástól független. Az állapotátmeneti valószínűségek a véletlen bolyongás tájképének „függőleges struktúrájától", az *Eval*-függvény által meghatározott hegyek magasságától, völgyek mélységétől, lejtők meredekségétől függ, valamint a fokozatosan csökkenő hőmérséklettől. Ha a constraint-eket más sorrendbe rakjuk, azaz megváltoztatjuk az E-függvényt, nagyon különböző valószínűségeket kapunk. Ezzel szemben, az *a priori* valószínűségek függetlenek a constraintek-től, azok rendezésétől, és a szimuláció során sem változnak. Az említett tájkép „vízszintes struktúráját" határozzák meg, az egyes jelöltek „szomszédsági fokát", „közelségét".

Amint a hagyományos OT a jelöltek halmazát egyetemesnek feltételezi (a *Richness of the Base* alapelve és a GEN leképezés univerzalitása folytán), úgy a jelöltek halmazának a struktúráját sem tekintem – alapvetően – nyelvspecifikusnak. A legegyszerűbb javaslat szerint $w' \in Neighbour(w)$ akkor és csak akkor, ha *w*-ből egy elemi transzformáció segítségével konstruálható *w'*. Mivel ezek az elemi transzformációk – például egy szegmentum beszúrása vagy törlése, egy szerkezeti határ eltolása – a nyelvi jel

szerkezetéből adódnak, a jelöltek halmazának a topológiája természetes módon kell, hogy kapcsolódjon a jelöltek univerzális ábrázolási formalizmusához.

Jegyezzük itt meg, hogy a topológia definiálását néhány további megszorítás is befolyásolja. Habár nem szükségszerű, de valószínűleg a szomszédsági relációt ésszerű szimmetrikusnak megadni. Fontos viszont a véletlen bolyongás szempontjából, hogy a jelöltek halmaza egy összefüggő struktúrát alkosson, vagyis bármely elemből el lehessen jutni bármely másik elembe véges számú lépéssel. Harmadszor, valószínűleg nem érdemes egy jelölthöz túl sok, egyaránt valószínű szomszédot rendelni, hiszen ekkor a szimulált hőkezelés – amelynek a lényege a szomszédsági struktúra értelmes kihasználása – elveszíti a hatékonyságát, és egy hihetetlenül ügyetlen, véletlenszerű próbálkozássá redukálódik.

Mindazonáltal, a topológiához kapcsolódó valószínűségi eloszlás definiálása már nem ennyire triviális feladat. Amennyiben mindegyik jelöltnek néhány szomszédja van, a legegyszerűbb megközelítés mindegyik szomszédnak egyenlő valószínűséget rendel: $P(w'|w) = P(w'' \mid w)$, ha mind w' és $w'' \in Neighbour(w)$. Ez a javaslat például sikerrel járt a holland gyorsbeszédben megfigyelhető hangsúlyeltolódások magyarázata során (Bíró, 2005b). De más lehetőségek is elképzelhetőek, és az alábbiakban a topológia, amelybe ezentúl beleértjük a $P(w'|w)$ *a priori* valószínűségeket is,[4] szerepét vizsgáljuk az SA-OT-ben.

Ezen a ponton kicsit árnyaljuk a topológia univerzalitásáról tett korábbi állításunkat. A topológiát meghatározó alapelvek, az *a priori* valószínűségek kiszámításának a logikája valóban univerzális, de a pontos részleteket esetleg paraméterek határozhatják meg. Amint rövidesen látni fogjuk, elképzelhető például, hogy egy jelölt szomszédjait úgy kapjuk, hogy bizonyos számú elemi transzformációk közül pontosan egyet hajtunk végre, de ezen elemi transzformációkhoz nem szükséges egyenlő valószínűségeket rendelni. Lehetséges tehát, hogy az egyes elemi transzformációk pontos valószínűsége, mint a topológia paraméterei, nyelvenként, regiszterenként, esetleg beszélőnként kis mértékben változhatnak.

[4] Eddig a $P(w'|w)$ *a priori* valószínűséget adott w esetén a $Neighbour(w)$ halmazon vezettük be. De tulajdonképpen a $Neighbour(w)$ függvényre nem is lesz szükségünk mostantól, elegendő lesz a $P(w'|w)$: $S \times S \rightarrow [0,1]$ leképezést megadnunk. A $Neighbour(w)$ pedig azon w'-k halmaza lesz, amelyekre $P(w'|w)$ pozitív (vagy nagyobb egy adott ε-nál). Más megközelítésben, $Neighbour(w)$-t felvehetjük úgy is, mint a jelöltek teljes S halmazát, csak éppen e halmaz jelentős részén nulla (elenyésző) az *a priori* valószínűség. Ne felejtsük el azt sem, hogy bármely w-re a

$$\sum_{w'} P(w' \mid w) = 1$$

feltételnek teljesülnie kell, mivel a valószínűségi eloszlást 1-re kell normálni.

4 Szó.ta.go.lás

4.1 Az SA-OT komponensei a klasszikus szótagolási modellhez

A jelen cikkben az SA-OT algoritmust az egyik klasszikus optimalitáselméleti példán, a szótagoláson mutatom be: hogyan bontsunk szótagokra (szótagkezdetre, szótagmagra és kódára) egy magánhangzó-mássalhangzó sorozatot? (Magyarul lásd például: Rebrus, 2001.) A paradigma nem csupán a *faktoriális tipológia* klasszikus példája 1993. óta, de az Optimalitáselmélet különböző implementációit is ezen a feladaton illusztrálták (véges állapotú automatákkal: Gerdemann és van Noord, 2000; dinamikus programozással: Tesar és Smolensky, 2000). Az egyszerű constraint-ek ellenére a legjobb jelölt megtalálása nem triviális feladat, mivel a jelöltek halmaza végtelen a rekurzív epentézis (*over-parsing*) lehetőségének a következtében.

A feladat ismert: egy bemeneti sztringhez, mint mögöttes reprezentációhoz, rendelünk egy kimeneti sztringet (fűzért), mint felszíni alakot, amely alapvetően a bemeneti sztring másolata, de

(a) a bemeneti sztring néhány karaktere törölhető (*underparsing*, alulelemzés),
(b) epentetikus szegmensek szúrhatók be (*overparsing*, túlelemzés),
(c) a szótaghatárok jelölve vannak (például egy ponttal).

A szótagok pontosan egy *nukleuszt* (*szótagmagot*) tartalmazhatnak, és nyelvenként változik, hogy mi lehet szótagmag. Az egyszerűség kedvéért feltételezzük most, hogy a nyelv hangkészletének egy része (magánhangzók, diftongusok, szillabikus szegmentumok) szerepelhetnek csak nukleuszként a szótagban, és ezek a szegmentumok szerepet nem is tölthetnek be.

A szótagnak a nukleuszt megelőző részét *szótagkezdetnek* (*onset*-nek), a nukleuszt követő darabját pedig *kódának* nevezzük. Ismét az egyszerűség kedvéért azt is feltételezzük, hogy a szótagkezdet és a kóda nem lehetnek komplexek (elágazóak), csupán egyetlen fonémát tartalmazhatnak. A nyelvek egy részében ez valóban létező megszorítás, és most elsősorban az algoritmus tulajdonságai érdekelnek bennünket.

Adott mögöttes reprezentációhoz tartozó jelölthalmaz a fenti feltételeknek megfelelő valamennyi sztring. Gerdemann és van Noord (2000) bemutatja, hogy habár végtelen, de a jelölthalmaz egy reguláris nyelvet alkot. Ezen halmaz optimális elemét ezek után az alábbi constraint-ek valamilyen rendezése mellett kell megtalálnunk:[5]

ONSET: a szótagkezdettel *nem* rendelkező szótagok száma
NOCODA: a kódával rendelkező szótagok száma
PARSE: az *alulelemzett* (a mögöttes formából törölt) szegmentumok száma.
FILLNUCLEUS: az *túlelemzett* (epentetizált), szótagmagi pozícióban található szegmentumok száma.
FILLONSET: az *túlelemzett* (epentetizált), szótagkezdeti pozícióban található szegmentumok száma.

[5] A constraint-eket, a szokásos fonológiai irodalommal ellentétben, nem azon feltételek megadásával határozom meg, amelyeket az alakoknak lehetőleg teljesíteniük kell. Hanem, mivel a constraint-ek egészértékű függvények, a jelölteknek kiosztott sértések számát definiálom.

Hogyan lássunk most neki e modell SA-OT-vel történő implementációjához? Egy SA-OT szimuláció előkészítése a következő négy lépésből áll:

1. A jelöltek halmazának (azaz a *Generátor- függvénynek*) a megadása.
2. A topológia és az *a priori* valószínűségek definiálása a jelöltek halmazán.
3. A constraint-ek és azok rendezésének (hierarchiájának) a meghatározása.
4. Az SA-OT paramétereinek, elsősorban a „hűtés menetének" (*cooling schedule*, a hőmérséklet csökkentési ritmusának) a megadása.

Ezen teendők közül az első és a harmadik pont a hagyományos Optimalitáselméletből ismert, és a nem-számítógépes nyelvészet területe. A szótagolás kapcsán fentebb már definiáltuk a jelöltek halmazát és a constraint-eket, és rövidesen kipróbáljuk a modellt több lehetséges hierarchiára (v.ö. *factorial typology*). Az utolsó pontra kitérünk rövidesen, így most összpontosítsunk a topológiára, pontosabban szólva az *a priori* valószínűségekre.

Két jelöltsztringet akkor tekintünk szomszédnak, ha pontosan egy *elemi lépés* (*elemi transzformáció*) segítségével eljuthatunk egyikből a másikba. Az elemi lépések – első megközelítésben – a következők:

1. A szóalak hosszának növelése epentetikus fonéma beszúrásával (magánhangzó nukleuszi pozícióba, mássalhangzó szótagkezdeti vagy kódai pozícióba).
2. A szóalak hosszának a csökkentése egy epentetikus szegmentum törlésével, vagy egy eredeti (nem-epentetikus) szegmentum alulelemzésével. Egy alulelemzett eredeti szegmentum visszaelemzése.
3. Egy szótaghatár eltolása, vagyis egy szótagkezdet átértelmezése kódává, vagy egy kóda átértelmezése szótagkezdetté.

Valójában, az SA-OT implementációnk nem *választ* egyet az előre kiszámolt szomszédok közül, hanem *legyártunk* egyet közülük. Az *a priori* valószínűségeket az határozza meg, ahogyan pontosan eljárunk. Először is, eldöntjük, hogy újraszótagolunk-e (a 3. lehetőség), vagy megváltoztatjuk-e a szó hosszát. Az újraszótagolás esélye $p_{reparse}$, míg az első két lehetőségé $1 - p_{reparse}$. A tapasztalat szerint a 60% körüli $p_{reparse}$ érték volt a legcélravezetőbb. Amennyiben az újraszótagolás mellett döntünk, véletlenszerűen választjuk ki, hogy melyik intervokális mássalhangzónak változtassuk meg a státuszát (ha *onset* volt, kódává tesszük, és fordítva). Ellenkező esetben, 50% valószínűséggel beszúrunk egy epentetikus szegmentumot (40%, 40% és 20% rendre az esélye annak, hogy szótagkezdetet, nukleuszt vagy kódát szúrunk be). És szintén 50% eséllyel törlünk egy szegmentumot[6]: ha epentetikus volt, töröljük, ha parszolt eredeti szegmentum volt, alulelemezzük, de ha eddig alulelemezve volt, akkor újraparszoljuk. Végül, amint eldöntöttük, hogy melyik műveletet választjuk, a jelöltsztringen belüli pozíciók közül egyenlő eséllyel kiválasztjuk, hogy hol hajtsuk azt végre – amennyiben ezáltal megengedett alakot kapunk.

[6] Az epentézis és a törlés egyenlő valószínűségének a célja az, hogy egyensúlyban legyenek a stringet növelő és az azt csökkentő operációk. A szimuláció korai fázisában, amikor még a magas hőmérséklet megenged bármely állapotátmenetet, fontos, hogy a véletlen bolyongó bejárhassa a jelöltek halmazának egy jelentős részét – az epentetikus szegmentumokat törlő operációk túlsúlya ezt megakadályozná. Ugyanakkor, a túl sok epentézis szükségtelenül megnöveli, „felfújja" a sztringet.

4.2 Az SA-OT futtatása

Ha ezt a szimulációt lefuttatjuk, csak a legritkább esetben kapjuk vissza a várt alakot, mivel egy sor *lokális minimum* állít csapdát a modell számára. Korábbi modellekkel ellentétben (például Bíró 2005b), e lokális optimumok nem felelnek meg performanciahibáknak. Így ki kell egészíteni a topológiát olyan további, *ad hoc* transzformációknak megfelelő szomszédsági relációkkal, amelyek megszüntetik ezeket a csapdákat. E *posztprocesszálás* során $p_{postproc}$ valószínűséggel két transzformációt hajtunk végre az előbbiekben leírt algoritmus eredményeként kapott sztringen. Egyrészt törölhetünk egy szótagot, ha pontosan egy epentetikus szótagkezdetből és egy epentetikus szótagmagból áll: az ilyen szótag törlése ugyanis két lépésből áll, de az első lépés az eredetinél rosszabb jelöltet hozna létre. Hasonló okokból nekünk kell közbelépnünk, ha egy epentetikus nukleusz mellett egy alulelemzett magánhangzó, vagy egy epentetikus onset mellett egy alulelemzett eredeti mássalhangzó áll.

1. Táblázat: A $p_{reparse}$ és a $p_{postproc}$ paraméterek hatása az SA-OT algoritmus pontosságára

$p_{reparse}$	%	$p_{reparse}$	%
0.00	15	0.60	20
0.10	15	0.70	15
0.20	15	0.80	14
0.30	16	0.90	9
0.40	14	1.00	3
0.50	17		

$p_{postproc}$	%	$p_{postproc}$	%	$p_{postproc}$	%
0.00	19	0.35	19	0.70	14
0.05	11	0.40	15	0.75	15
0.10	8	0.45	12	0.80	16
0.15	10	0.50	13	0.85	14
0.20	14	0.55	11	0.90	16
0.25	18	0.60	11	0.95	21
0.30	14	0.65	14	1.00	25

Az 1. Táblázat azt mutatja be, hogy a két tárgyalt paraméter – $p_{reparse}$ és $p_{postproc}$ – miként befolyásolja az SA-OT algoritmus pontosságát, a topológia *a priori* valószínűségeinek modulálása révén. Az *an.ta* bemenettel indítottuk el az algoritmust, mindegyik ($p_{reparse}$, $p_{postproc}$) paraméterkombináció esetén tízszer-tízszer; a használt hierarchia ONSET ≫ FILLNUCLEUS ≫ PARSE ≫ FILLONSET ≫ NOCODA volt ebben az előzetes kísérletben. Mindkét táblázat csak az egyik paraméter szerepét mutatja be, miközben a másik paraméterre átlagoltunk. Látható, hogy bizonyos paraméterkombinációk jelentősen eltérő valószínűséggel találják meg a helyes alakot – esetünkben *Tan.ta*-t, egy szókezdő epentetikus mássalhangzóval. Az SA-OT algoritmus paraméterei a következők voltak: $K_{max} = 5$, $K_{min} = -2$, $K_{step} = 1$, $t_{max} = 4$, $t_{min} = 0$, $t_{step} = 0{,}1$. Az öt constraintet a 0, 1,…, 4 indexekkel azonosítottuk.

Hasonló feltételek mellett futottak azok a szimulációk, amelyekről a 2. Táblázat számol be. Ezúttal a topológia paraméterei rögzítettek: $p_{reparse} = 0{,}60$, $p_{postproc} = 0{,}95$.

Figyeljük meg, hogy különböző hierarchiák nagyon eltérő pontossággal produkálhatók. Ezt a tényt interpretálhatjuk úgy, hogy az SA-OT lehetőséget biztosít arra, hogy habár egyes nyelvtípusokat elvileg megengedne a faktoriális tipológia (azaz az emberi agy), de megmagyarázzuk, miért nem fordulnak elő (egyáltalán nem, vagy csupán ritkán) a világ nyelvei között.

Ugyanis ezeket az elvileg lehetséges nyelveket nehéz produkálni, valamint a sok hiba miatt a következő generáció is nehezen sajátítja azt el. Vagyis az adott nyelvtípus evolúciósan nem stabil.[7] Ha az SA-OT algoritmus, legalább elvileg, képes ilyen típusú magyarázatra, az jó hír a számunkra: ugyanis reményt ad arra, hogy az Univerzális Grammatika modelljeit nem kell túlkomplikálni, ha a faktoriális tipológia történetesen nem képes megmagyarázni egy kis űrt a megfigyelhető nyelvtipológiában. Lehetséges ugyanis, hogy ezt az űrt nem az UG szintjén kell megmagyarázni: az UG megengedné, de azért nem fordul elő a típus, mert nehéz lenne helyesen produkálni.

2. Táblázat: Az SA-OT algoritmus pontossága különböző hierarchiák esetén

Hierarchia	*%*
prs fio fin noc ons	31
prs fio fin ons noc	26
prs fio noc fin ons	78
prs fio noc ons fin	84
prs fio ons fin noc	14
prs fio ons noc fin	72
prs fin fio noc ons	38
prs fin fio ons noc	25
prs fin noc fio ons	30

5 Összefoglalás

A jelen cikk célja egyrészt az volt, hogy a magyar nyelvű szakmai közönségnek bemutassam a szimulált hőkezelés társítását az Optimalitáselmélettel, azaz az SA-OT algoritmust. Az eljárás részletei mellett több korábbi cikkben, valamint a készülő disszertációmban részletesen érvelek, különböző típusú – matematikai, „filozófiai", empirikus – érveket hozva. Az érdeklődő kipróbálhatja az algoritmus demonstrálására elkészített weblapot is a *http://www.let.rug.nl/~birot/sa-ot/* oldalon.

A cikk fő célja viszont annak a bemutatása volt, hogy milyen szerepet játszik az algoritmusban a jelöltek halmazán bevezetett topológia (szomszédsági struktúra), különösen pedig az *a priori* valószínűségek. Ezen fogalmak újak a hagyományos

[7] Ez a gondolatmenet beleilleszkedik abba a kutatási programba, amely a Chomskyánus megközelítést – ha egy nyelvtípus nem létezik, akkor azt az Univerzális Grammatika modelljének kell kizárnia – más megközelítésekkel igyekszik kiegészíteni. Jäger (2003) megmutatja, hogy a faktoriális tipológia által megengedett egyes típusok azért nem léteznek a világ nyelvei között, mert ezek a típusok evolucionárisan nem stabilak. Boersma (2004) pedig olyan magyarázatot javasol, amely szerint egyes típusokat nehezebb elsajátítani egy tanulóalgoritmussal.

Optimaitáselméletben, habár *Paul Smolensky* konnekcionista munkái is használnak egy szomszédsági struktúrát. Ebből a célból megvizsgáltuk, hogy miként alkalmazható az SA-OT az Optimalitáselmélet talán leggyakrabban használt példájára, a bemenet szótagstruktúrájának a kiszámítására. Láttuk, hogy hogyan vezethetők be nem-triviális *a priori* valószínűségek néhány paraméter segítségével – szemben például a Bíró (2005b) által használt modellel, amely a topológia szempontjából rendkívül egyszerű. Valamint azt is láttuk, hogy e paraméterek variálásával az algoritmus kimenete, pontossága erősen változhat. Hipotézisem szerint az *a priori* valószínűségek kiszámításának a módja univerzális, de a paraméterek pontos értékei változhatnak nyelvenként, regiszterenként, vagy akár beszélőnként.

A cikkben bemutatott eredmények egyelőre nem meggyőzőek abból a szempontból, hogy konkrétan a szótagolás jól implementálható dinamikus programozással (Tesar és Smolensky, 2000), és a véges állapotú automatákkal is meglepően jól működik, szemben az elméleti várakozásokkal (Gerdemann és van Noord, 2000), miközben az SA-OT algoritmus egyelőre meglehetősen alacsony pontosságot produkált. De a bemutatott eredmények egy folyamatban levő munka részeredményei, amelyek továbbgondolásra szorulnak. Továbbá, a pontosság radikálisan növelhető, ha néhány szimulációt párhuzamosan futtatunk. Még ha egy szimuláció csupán 20% valószínűséggel találja is meg a helyes alakot, ha tíz szimulációt futtatunk párhuzamosan, 90%-ot meghaladja annak az esélye, hogy a tíz közül legalább az egyik a keresett alak lesz. Márpedig tíz output alak közül megkeresni a legjobbat egy könnyű feladat, összehasonlíthatatlanul könnyebb, mint egy végtelen halmaz legjobb elemének a megtalálása. Emlékezzünk, épp ebből az utóbbi célból alkalmaztuk a szimulált hőkezelést.

Köszönetnyilvánítás

A köszönetemet fejezem ki a hollandiai Groningeni Egyetem (Rijksuniveriteit Groningen) *High Performance Computing* programjának, valamint témavezetőimnek és konzulenseimnek, *John Nerbonne*-nak, *Gosse Bouma*-nak és *Gertjan van Noord*-nak a segítségükért a cikk alapjául szolgáló kutatásban.

Bibliográfia

1. Bíró, T.: Az Optimalitáselmélet megvalósítása szimulált hőkezeléssel, szemináriumi dolgozat Rebrus Péter órájára, ELTE Elméleti Nyelvészet, 1997.
2. Bíró, T.: Quadratic Alignment Constraints and Finite State Optimality Theory, In: Proc. FSMNLP, pp. 119-126, Budapest, (valamint: *ROA-600,* http://roa.rutgers.edu), 2003.
3. Bíró, T.: Squeezing the Infinite into the Finite: Handling the OT Candidate Set with Finite State Technology, Proc. FSMNLP, Helsinki, 2005a.
4. Bíró, T.: When the Hothead Speaks: Simulated Annealing Optimality Theory for Dutch Fast Speech, In: Cremers, C., Reckman, H., Poss, M., van der Wouden, T. (eds.): Selected Papers of the 15th Meeting of Computational Linguistics in the Netherlands, Leiden, 2005b.
5. Bíró, T.: How to Define Simulated Annealing for Optimality Theory? In: Proc. 10th Conf. on Formal Grammar and 9th Meeting on Mathematics of Language, Edinburgh, 2005c

6. Bíró, T.: Finding the Right Words: Implementing Optimality Theory with Simulated Annealing, PhD disszertáció, University of Groningen, Hollandia, 2006.
7. Boersma, P.: Review of *B. Tesar & P. Smolensky* (2000): Learnability in Optimality Theory, *ROA-63*8 (Rutgers, Optimality Archives, http://roa.rutgers.edu), 2004.
8. Eisner, J.: Easy and Hard Constraint Ranking in Optimality Theory: algorithms and complexity, In: Eisner, J., Karttunen, L., Thériault, A. (eds.): Finites-State Phonology, Proceedings of SIGPHON 5, pp. 57-67, Luxemburg, 2000.
9. Ellison, T.M.: Phonological Derivation in Optimality Theory, In: Proc. COLING-94, pp. 1007-1013, Kyoto, 1994.
10. Frank, R., Satta, G.: Optimality Theory and the Generative Complexity of Constraint Violability, In: Computational Linguistics, 24(2):307-315, 1998
11. Gerdemann, D., van Noord, G.: Approximation and Exactness in Finite State Optimality Theory, In: Eisner, J., Karttunen, L., Thériault, A. (eds.): Proc. SIGPHON, 2000.
12. Jäger, G.: Simulating language change with functional OT,. In: Kirby, S. (ed.): *Language Evolution and Computation, Proc. of the Workshop at ESSLLI, Vienn*a, pp. 52–61, 2003.
13. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science no. 4598, vol. 20, pp. 671-680, 1983
14. Kuhn, J.: Processing Optimality-theoretic Syntax by Interleaved Chart Parsing and Generation, In: Proc. ACL-2000, pp. 360-367, Hongkong, 2000.
15. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of State Calculation by Fast Computing Machines. J. Chemical Physics, vol. 21, no. 6, pp. 1087-1092, 1953.
16. Prince, A., Smolensky, P.: Optimality Theory: Constraint Interaction in Generative Grammar, RuCCS-TR-2, 1993 / Blackwell, Malden, MA, 2004.
17. Rebrus, P.: Optimalitáselmélet, In: Siptár, P. (szerk.): Szabálytalan fonológia, Tinta, 2001.
18. Reeves, C.R.: Modern Heuristic Techniques for Combinatorial Problems, McGraw-Hill, London, 1995.
19. Tesar, B., Smolensky, P.: Learnability in Optimality Theory, MIT Press, Cambridge, 2000.

II. Ontológia

Réteges struktúra, alaprelációk

Szakadát István

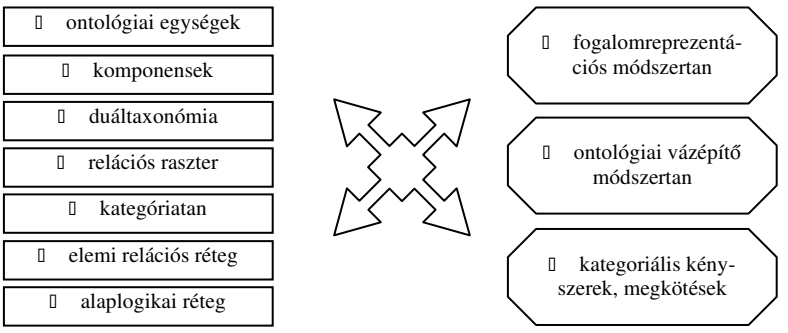
BME Szociológia és Kommunikáció Tanszék, MOKK, 1111 Budapest, Sztoczek u. 2.
i@syi.hu

Kivonat: A MEO-projektben felépítendő csúcsontológia szerkezete rétegekre bontható. A tanulmány először felvázolja a rendszer struktúráit, az egymásra épülő modulok, rétegek kapcsolódási, ellenőrzési lehetőségeit, majd röviden bemutatja, hogyan használják a fontosabb szakmai közösségek a paradigmatikus relációk két legfontosabb típusát, a generikus és a partitív relációt, illetve hogyan lehet ezeket formális nyelven definiálni.

1. Ontológiai rétegmodell

Minden axiomatikusan építkező, formális rendszer komponensekre, rétegekre bontható. A különböző rétegekben adott ideig „szabadon" folyik az építkezési munka, de úgy, hogy az egyes rétegek közti kapcsolatokat, 'interfészeket' folyamatos kölcsönös kontroll alatt kell/lehet tartani, és ha valamelyik rétegben folyó munkák (és döntések) eredményeként szükségessé válik egy másik réteg tartalmának, szerkezetének, struktúrájának megváltoztatása, akkor azt megtehetjük az adott rétegen belül. Ebben a rendszerben a változtatási igény mindkét irányból kezdeményezhető (fentről lefelé vagy fordítva).

A MEO projektben az alkalmazott rétegek, illetve a réteges építkezés munkáját támogató módszertanok az alábbi rendben épülnek egymásra:



A legalsó két rétegben a legelemibb logikai alapfogalmak, illetve az algebrai tulajdonságokkal jellemezhető elemi relációs fogalmak „helyezkednek el", mint például a matematikában ismert relációs tulajdonságok (reflexivitás, szimmetricitás, tranzitivi-

tás stb.) vagy olyan nevezetes alaprelációk, mint a rendezés, ekvivalencia, egyenlőség, tolerancia stb. Ahhoz, hogy a réteges struktúránk következő szintjeire léphessünk, szükségünk van arra, hogy olyan a legfontosabb paradigmatikus relációkat definiálni tudjuk.

2. A generikus és partitív relációk használati gyakorlata

A szemantikai relációk fontosságát sok tudományág elismerte, amiből fakadóan egymástól eltérő definíciók és tipizálások jöttek létre. Legalább három nagyobb szakmai közösség, a nyelvészeti, a tudásreprezentációs-informatikai és a könyvtártudományi-osztályozáselméleti szakma saját fogalomkészletet alakított ki ezen a területen. A relációs fogalmak közül kiemelve a két legjelentősebbet, a *generikus* és a *partitív* relációt, a következőkben röviden összefoglaljuk, hogyan jellemezhetők, hogyan illeszthetők egymáshoz a különböző megközelítések, fogalomhasználati gyakorlatok.

2.1. Módszertani kitérő

Mielőtt témánk kifejtésébe vágnánk, két rövid módszertani kitérőt kell tennünk.
1.) A természetes nyelvekben nagyon gyakori, hogy a relációk megnevezésére használt predikátum (reláció) „rövidül", vagyis terminussá (függvénnyé) változik, s ezáltal „önállósodik", vagyis a reláció valamelyik relátumának nevévé válik. A bináris relációk közül példát véve a **szülője** reláció két EMBER között teremt egy **rákövetkezés** relációt, amit a következőképpen jelölhetünk:

szülője(ember1,ember2)

A reláció inverze a **gyereke** reláció, amelynek ugyanazok az argumentumai:

gyereke(ember2,ember1)

A nyelvhasználati gyakorlatban azonban a **szülője** reláció első relátumát adó ember1 argumentum helyett megjelenik a SZÜLŐ mint önálló és nem-relációs terminus, míg a **gyereke** relációban az ember2 argumentum helyett a GYEREK mint önálló és nem-relációs terminus, és ez a két függvény betehető a két alapreláció argumentumai helyére, vagyis a két reláció a következőképpen írható le:

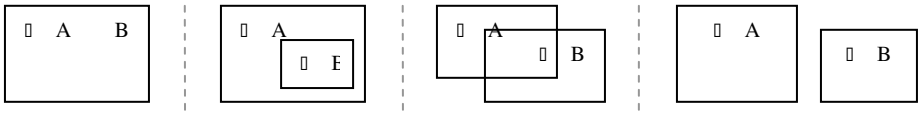
szülője(SZÜLŐ, GYEREK)
gyereke(GYEREK, SZÜLŐ)

A továbbiakban a szóban forgó relációk különböző használati gyakorlatát vizsgálva mind a „relációs", mind a „függvényes" alakot alkalmazni fogjuk – igazodva az elemzett szakterület tényleges gyakorlatához.
2.) A nevezetes relációk jelentős részénél értelmes és gyakran használt a reláció inverzének a fogalma. A cikkünkben vizsgált két hierarchikus relációnál is ez a helyzet. Olykor azonban előfordulhat, hogy egy relációs terminus esetében pongyola fogalomhasználati gyakorlat alakul ki, és az alapreláció megnevezését alkalmazzák akkor is, amikor az inverzét kellene (a későbbiekben látni fogjuk: a taxonima vagy a szubszumpció relációmegnevezéseknek nincs meg az inverz párjuk). A „pongyolaság" ilyenkor nyilván annak köszönhető, hogy az adott fogalom használatát a kontextus egyértelműsíti, s ezáltal elkerülhető a félreértelmezés lehetősége.

2.2. NYK-gyakorlat

A nyelvészközösség (NYK) világában merült fel először a szemantikai relációk vizsgálatának igénye. Saussure a *szintagmatikus* és *asszociatív* relációkat különböztette meg egymástól [14], melyből számunkra utóbbiak az érdekesek, mivel a fogalmak egymás közti viszonyával, nem pedig a szavak mondatokba illeszkedésével akarunk foglalkozni. Később Saussure, majd az egész nyelvészközösség az asszociatívról a *paradigmatikus* relációra váltott, ami számunkra azért szerencsés, mert legalább ezen a ponton elkerülhetjük azt az értelmezési problémát, ami abból fakadhatna, hogy időközben más szakmákon belül (pl. az objektumorientált modellezés vagy a tezauruszok világában) az *asszociáció* relációt újra elkezdték használni (természetesen egészen más jelentés mentén).

A nyelvészetben a paradigmatikus relációk fogalma alatt a szavak, lexikai egységek közti *lexikai relációkat* kezdték el tipizálni – előbb Lyons [10], majd Cruse [2]. Utóbbi vetette fel, hogy az alapvető lexikai relációkat halmazelméleti relációkra támaszkodva érdemes definiálni – ahogy ezt az alábbi ábra mutatja:



A fenti relációk a következő algebrai tulajdonságokkal jellemezhetőek:

azonosság:	reflexív, szimmetrikus, tranzitív, antiszimmetrikus –	azonosság
inklúzió:	reflexív, antiszimmetrikus, tranzitív –	nem szigorú rendezés
	aszimmetrikus, (irreflexív), tranzitív –	szigorú rendezés
átfedés:	szimmetrikus, reflexív –	tolerancia
különbözőség:	irreflexív, szimmetrikus, nem-tranzitív –	különbözőség

Néhány megjegyzés: A későbbiekben használni fogjuk azt a tényt, hogy a magába foglalás kétféleképpen is minősíthető attól függően, hogy a kapcsolatra megengedjük-e a reflexivitást vagy sem. A toleranciarelációt hasonlósági relációnak is szokták nevezni. A különbözőség reláció nem-tranzitív, azaz se nem tranzitív, se nem intranzitív:

A bal oldali ábra az intranzitivitást szemlélteti, hiszen A elkülönült B-től, B C-től, de A nem elkülönült C-től, míg a jobb oldalon igaz a tranzitivitás, hiszen A elkülönült B-től, B C-től és A C-től.

Az azonosság segítségével a **szinonima** lexikai relációt határozhatjuk meg, de ezzel a jelen tanulmányban nem foglalkozunk (bár egy alaposabb tárgyalás esetén inkább az ekvivalenciarelációt kellene használnunk). Az inklúzió (magába foglalás) segítségével a **hiponima** (hyponymy) relációt, az átfedéssel a **kompatibilitás**

(compatibility) relációt, míg az elkülönüléssel az **inkompatibilitás** (incompatibility)relációt határozhatjuk meg. A legtöbb relációnak van inverzrelációja, ezért itt megemlítjük, bár a továbbiakban nem foglalkozunk a kérdéssel, hogy a **hiponima** inverzrelációja a **hiperonima** (hyperonymy).

A nyelvészet területén bevett gyakorlat, hogy adott nyelvi szabályszerűség megállapításához, ellenőrzéséhez illeszkedési mintákat, teszteket adnak meg. A hiponima lexikai relációhoz hozzárendelt minta, nyelvi teszt egy jól-ismert szerkezet:

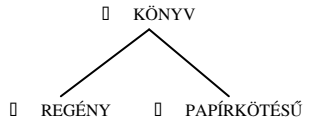
is-a

Ha két lexikai egység kielégíti ezt, akkor – a nyelvészet világában – hiponimáról beszélhetünk. Az alábbi példák mind beleesnek a **hiponima** körébe:

REGÉNY **is-a** KÖNYV
VERSESKÖTET **is-a** KÖNYV
PAPÍRKÖTÉSŰ (KÖNYV) **is-a** KÖNYV
KEMÉNYFEDELŰ (KÖNYV) **is-a** KÖNYV

A hiponima reláció előnyös tulajdonsága, hogy a tartományába tartozó elemek hierarchikus struktúrába szervezhetőek. A fenti példánk alapján az alábbi két hierarchia rajzolható fel:

Kérdés, mi az oka annak, hogy bár a hiponima relációt a fenti négy esetben érvényesnek tartottuk, mégis két hierarchiát rajzoltunk fel belőlük, és nem érezzük összevonhatónak őket? Vagy a kérdésben rejlő problémát másként megfogalmazva válaszra vár az is, miért nem érezzük jól-formáltnak a következő hierarchiát:



A fenti hierarchiával az a probléma, hogy a KÖNYV alá rendelt két terminus, a regény és a papírkötésű, átfedő viszonyban van egymással, hiszen lehet papírkötésű regény. Másképpen fogalmazva: a két alárendelt elem valamilyen mértékben kompatibilis egymással. A fenti példa a hierarchiába rendezés problémáját mutatja. Mivel a **hiponima** reláció nincs semmilyen kapcsolatban az **inkompatibilitási** relációval, ezért önmagában semmi sem is biztosítja azt, hogy egy terminusnak alárendelt két másik egység inkompatibilis legyen egymással. Márpedig ezt várjuk el a hierarchikus struktúrától. Ez viszont azt jelenti, hogy a hierarchiát nem lehet egyetlen relációval megadni, az alá-fölérendeltséget leíró hiponima reláció mellett ki kell fejezni az ugyanazon fölérendelt alá tartozó elemek közti inkompatibilitást is. Nem teljesen korrekt megoldásként Cruse (és még sokan mások) ezen a ponton ajánlják bevezetni a taxonima lexikai relációt, melynek nyelvi tesztje, „jele”:

[van] fajta/is-a-kind-of (olykor: is-a-type-of, is-a-sort-of)

Azért nem teljesen korrekt ez a megoldás, mert az ’is-a-kind-of’ teszt (és megnevezés!) inkább csak jelzi a problémát, de önmagában nem oldja meg. Mindenesetre az új lexikai reláció bevezetésével egy fontos különbségre már rámutathatunk. A

hiponima és a **taxonima** relációk kétfajta elemi relációtípus alá sorolhatók be, mert a reflexivitást illetően eltérően viselkednek. Mondhatjuk ugyanis azt, hogy:

KUTYA **[van]** KUTYA /DOG **is-a** DOG (**hiponima**),

de nem mondhatjuk, hogy:

KUTYA **[van] fajtája** KUTYA/dog **is-a-kind-of** dog (**taxonima**).

A taxonima szűkebb terjedelmű a hiponimához képest. A két reláció minőségére vonatkozóan az alábbiakat állapíthatjuk meg:

hiponima:	antiszimmetrikus, reflexív, tranzitív	nem szigorú rendezés
taxonima	aszimmetrikus (irreflexív), tranzitív	szigorú rendezés

A taxonima bevezetésével a hiponima egyik fajtájaként egy hierarchia-képző relációt kaptunk. A WordNet projekt keretében Fellbaum bevezetette a hiponima egy másik típusát, a **troponima** (troponymy) relációt. Az új reláció a lexikai egységek szófaji szerinti szűkítésén alapul, hiszen a troponima igékre alkalmazható. Az ellenőrző tesztkeret angolul a következő:

is-a-…-in-some-manner

Egy-két példát véve:

FUTÁS/RUNNING **is-a-…-in-some-manner** MOZGÁS/MOVING

SÉTÁLÁS/WALKING **is-a-…-in-some-manner** MOZGÁS/MOVING

A tesztkeretet nehéz magyarra fordítani, „[van] valami módon” lenne a szöveghű magyar változat, ami nem használható. Működőképesebbnek tűnik a „[van] valami fajta” minta. Érdekes kérdés, vajon meg lehet-e engedni azt, hogy a különböző nyelvek közti tesztekben ilyen eltérések lehessenek, de ez a probléma már nem tartozik elemzésünk tárgykörébe.

 ***

A nyelvészet fontos fejleménye volt, amikor felismerték, hogy a főnevek között nem csak a generikus reláció mentén lehet kapcsolatot teremteni. Pustejovsky különböztette meg a **generikus** relációtól a **szerep** relációkat, melyeket másként lehet jellemezni [13]. Példája a HÁZIÁLLAT/PET volt (további példái voltak: vásárló, élelmiszeráru, mosoda). Pustejovsky értelmezése szerint a kutya és állat közti **hiponima** reláció lényegesen különbözik a háziállat-állat kapcsolattól. Háziállatnak lenni ugyanis csak bizonyos körülmények között, csak adott kontextusban lehetséges. Ez a kategória időben nem stabil, nem állandó, és nem tesz hozzá semmit a felettes kategóriához. Ez a különbségtétel felhívta a figyelmet arra, hogy az ontológiaépítés során pontosan el kell tudni választani **szerep** alapú főneveket a **generikus** hierarchiába rendezhető főnevektől.

 ***

A nyelvészet területén, bár elismerték a fontosságát, közel sem olyan alapossággal foglalkoztak a *rész-egész* (part-whole) viszonnyal, mint a **hiponima** relációval. Lyons, Cruse, sőt a WordNet projekt is egyaránt elismerte e kapcsolattípus fontosságát, sajátos szerepét a nyelv világában, de a **partonima** reláció mindig „másodrendű maradt” a nyelvészek számára.

Sokakat foglalkoztató kérdés volt, hogy a partonima reláció tranzitívitásának problémája. Már Lyons megmutatta, hogy nem minden esetben áll fent a tranzitivitás („az ajtónak része a kilincs, a háznak része az ajtó, de a háznak nem része a kilincs…”). Cruse megpróbált tesztkeretet adni a partonima reláció ellenőrzésére, amely azonban nem minden esetben működött jól (ezt Cruse maga is elismerte). Cruse szerint a

partonima fennállásához elégséges feltétel, ha két terminus (X és Y) eleget tesz az alábbi két tesztnek (fordított szerepben):

Y **van/has-a** X

X **része/is-part-of** Y

Cruse szerint azért szükséges két feltétel, mert az első tesztkereten való megfelelés még nem elégséges a **partonima** fennállásához, hiszen a FELESÉG és a FÉRJ osztályfogalmak átmennek a teszten:

FELESÉG **van/has-a** FÉRJ

Viszont a második feltételnek már nem felelnek meg (nem is áll fenn köztük a része reláció):

* FÉRJ **része/is-part-of** FELESÉG

A **partonima** létezésének eldöntéséhez szükséges tesztek bemutatása után Cruse megpróbálta a reláció tipusait is megkeresni, ám – a későbbi tipizálási kísérletek fényében ezt nyugodtan mondhatjuk – nem igazán sikeres ebben az igyekezetében.

Összegzésképpen megállapíthatjuk, hogy a nyelvészet területén a **partonimával** kapcsolatban – még a **taxonima** reláció gyenge formalizálási kísérletéhez képest sem – történt komoly előrelépés, vagyis nem kezdődött meg a reláció formalizálása.

▯ ***

A nyelvészek által használt paradigmatikus relációk áttekintését lezárva meg kell még állapítanunk, hogy a nyelvészet nem foglalkozott érdemben az individuumokra vonatkozó kijelentésekben megjelenő, szintén az **is-a** tesztkerettel leírható, de az **eleme** reláció alá sorolható lexikai relációval, melynek szemléltetésére vegyük az alábbi példát:

`Rozi` **[van]** KUTYA/`Rozi` **is-a** DOG

Bármennyire is igaz azonban az, hogy ugyanazzal a tesztkerettel vizsgálható ez a reláció, mint a **hiponima**, mégis lényeges különbség van a kettő között, mivel az **eleme** reláció *heterogén*, azaz a reláció két relátumába másfajta halmaz elemeiből válogathatunk, ezzel szemben a **hiponima** (sőt, az összes eddig tárgyalt reláció) *homogén* volt.

Az **eleme** reláció a következő tulajdonságokkal rendelkezik tehát:

aszimmetrikus, (irreflexív), nem-tranzitív, heterogén

Ha nem is nyelvi tesztkeretként, de pontosabb megnevezésként használhatjuk a következő kifejezést: **[van] eleme/is-element-of**.

2.3. TIK-gyakorlat

Amennyire nem került be az individuumok problémája a nyelvészek érdeklődési körébe, annyira foglalkoztatta ugyanez a kérdés a tudásreprezentációs és informatikai közösség (TIK) tagjait, az informatikai rendszerek fejlesztésével foglalkozó szakembereket. Ez persze érthető, hiszen az ipari rendszerek működőképességéhez mindig szükség volt a világ valamely szegmensén belül konkrét individuumokra vonatkozó állításokra. Az ilyen – a működőképesség kényszerével terhelt – rendszerek fejlesztése megkövetelte a rendszerek elméleti leírását, és ez szakmacsoport a leggyakoribb megnevezéseként vélhetőleg az **is-a** kifejezést vette használatba. A gyakorlatorientáltságból fakadó érdekek nem igazán engedték meg, hogy, hogy a fogalom és terminus egységes értelmezés mentén maradjon széleskörű használatban, amit már 1983-

ben szóvá is tett Brachman egy híressé vált, „beszédes” című cikkében: *What IS-A link and isn't…* [1]. Az **is-a** reláció többértelmű használata persze csak a problémák egyik típusát jelentette. Ezen a szakterületen is nagyjából ugyanazokat a relációtípusokat termelték ki az évek során, mint a nyelvészek, de a relációk megnevezésében jelentős eltérések adódtak. A **hiperonima** reláció helyett a tudásreprezentáció területén gyakran alkalmazták az **is-a** terminust, a logika területén főleg a **szubszumpció** (subsumption) kategóriáját vették használatba, míg az objektumorientált programozás terjedésével a **szubtípus/szupertípus**, **pontosítás/általánosítás, részosztály/osztály**, **specializáció/generalizáció** terminuskettősök mindegyike forgalomba került (az első két pár függvényes, a második két pár relációs megnevezésű). Mindegyik elnevezéssel ugyanarra a fogalomra, az osztályozás alapját jelent **taxonimára** szándékoztak hivatkozni a terminusok alkalmazói. Az osztályozási hierarchiák pedig egyértelműen a felsőbb szintű típusok (osztályok) tulajdonságainak örökölhetősége miatt voltak fontosak.

Az informatikai fejlesztések világában szükség volt az individuumok és a típusok egyértelmű elkülönítésére, ezért kétfajta egyedfogalmat vettek használatba: a típus (osztály) és az individuum fogalmát (utóbbira további megnevezések is elterjedtek: *példány*, *instancia*, *előfordulás*, *elem*, *tag* – ezek mind függvényes megnevezések!). A **taxonima** (és a később tárgyalt *része*) reláció mellett (melyeket teljesen külön kezeltek) mindenfajta új relációt definiálni lehet, és ezeknek a relációkat is jellemezni lehetett aszerint, hogy típusok (osztályok) közti vagy példányok közti relációról van-e szó. Ez annyit jelent, hogy a relációk esetében is beszélhetünk *relációosztályokról*, illetve *relációelőfordulásokról* (az előbbi esetben *asszociáció*, az utóbbiban *link* a reláció típusának megnevezése).

***

Az informatikai rendszerek fejlesztésekor (főleg az objektumorientált rendszerek megjelenésével – talán a programozás komponens-elvűségének kényszere miatt?) a **taxonima** reláció mellett kiemelten kezelték a **partonima** relációt. Természetesen ezen a szakterületen is versengő teminushasználati gyakorlat alakult ki, és a *has-a*, *has-part*, *is-part-of*, *part-of* kifejezések mindegyike elterjedtté vált. Az OO-UML-szakma is új relációneveket vezetet be rész-egész kezelésére, és az *aggregáció/dezaggregáció*, illetve a *kompozíció/dekompozíció* relációk a **partonima** reláció megfelelői lettek ebben a világban (az aggregáció és kompozíció különbsége lényegtelen gondolatmenetünk szempontjából).

Bármennyire is fontos volt az aggregáció-kompozíció kategóriája az objektumorientált modellezés számára, nem igazán foglalkoztak a **partonima** reláció formalizálásával. Ezt a feladatot sokkal inkább a logika felől kezdték (és tudták) megoldani. Az ezzel foglalkozó új tudományág, a *mereológia* megalapozását, illetve a **partonima** reláció új megnevezését (**meronima**) Stanisław Leśniewski nevéhez lehet kötni [15], aki mellett – elméletalapítóként – még gyakran hivatkoznak a Leonard és Goodman szerzőpárra is [9]. Az igazán érdemi előrelépést e téren P. Simons 1987-ben megjelent könyve jelentette, amely megfelelő alapokat és döntő lökést adott **meronima** reláció beható tanulmányozásához [15]. Ekkortól kezdve a szakterület tanulmányai vagy a meronima reláció tipizálásával vagy annak formalizálásával foglalkoztak.

Winston, Chaffin és Hermann 1987-es cikke volt az első kísérlet a **partitív** relációk tipologizálására, melynek az alábbi hat altípus lett az eredménye [18]:

összetevő-objektum (component-object): pl. ág/fa
elem-gyűjtemény (member-collection): fa/erdő

porció-tömeg (portion-mass): szelet/torta
anyag-objektum (stuff-object): alumínium/repülőgép
jellemző-tevékenység (feature-activity): fizetés/vásárlás
helyszín-terület (place-area): Philadelphia/Pennsylvania

A szerzők egy évvel későbbi a fenti hat típushoz hozzáadtak még egyet, és a legtöbbet erre a felosztásra hivatkoznak a szakirodalomban [19]:

fázis-folyamat (phase-process): kamaszodás/felnövés

Iris, Litowitz, and Evens [8] a fenti hétből párat összevonva már csak négy típust "hagyott meg":

funkcionális rész (functional part)/fázis-folyamat + jellemző-tevékenység: kormány/bicikli,
szegmentum (segmented part)/összetevő-objektum + helyszín-terület: szelet/kenyér,
gyűjtemény-elem (collection-member)/elem-gyűjtemény + anyag-objektum: birka/nyáj,
részhalmaz (subset)/porció-tömeg: hús/étel.

Iris és társai szerint csak a *szegmentumra* és a *részhalmaz* relációra érvényes a tranzitivitás. A *funkcionális részre* és a *gyűjtemény-elem* reláció lehet tranzitív, de nem minden esetben, tehát a tranzititás nem húzható rá e két relációtípus egészére.

Pribbenow – saját megközelítését konstruktivistának nevezve – két fő részképzési lehetőséget különített el egymástól elválasztva az a priori adott, bizonyos szempont szerint állandónak minősíthető strukturálási, illetve az időlegesen megteremthető szegmentálási lehetőséget [12]. Az első típus része a fogalmi, szakterületi tudásunknak, a második típus nem. A fogalmilag állandó rész-egész relációtípusok Pribbenow szerint:

komponens-komplex egész (component-complex)
elem-gyűjtemény (element-collection)
anyagrész(mennyiség)-anyag/részhalmaz-halmaz (quantity-mass/subset-set)

Az időlegesen, tetszőleges szempont szerint konstruált részképzési lehetőségeket, melyek nem tartoznak a világtudásunk körébe, Pribbenow partícióknak, szegmenseknek nevezi.

Simon 1987-ben megjelent korszakos könyve után, a kilencvenes évek közepétől kezdve A. Varzi, B. Smith és C. Pontow voltak azok, akik Simon mereológiai elméletét el kezdték továbbformálni [17,11,12]. A mereológiai elméletek egymásra épülését, axiomatizálását tanulmányunk másik részében mutatjuk be, itt a fejezet zárásaként csak annyit jegyeznénk meg, hogy a **rész-egész** reláció egyik alesetének tartott **eleme** reláció ugyanaz, mint a példánya reláció, amit korábban a – gyakorlatban nagyon pongyolán kezelt – **is-a** reláció egyik típusaként mutattunk be.

2.4. KOK-gyakorlat

A könyvtártudományi és osztályozáselmélet közösséget (KOK) jellemezte talán a legnagyobb önreflexió a relációk értelmezése és alkalmazása során. A könyvtártudományi szakma az osztályozási rendszerek, tezauruszok építése során szembesült a relációk definiálásának és tipizálásának problémáival. Az osztályozáselmélet szakemberei értelemszerűen az osztályozás alaprelációját, a **taxonima** relációt vizsgálták. A tezauruszépítés feladatai megkívánták, hogy minél pontosabb és egyértelműbb gyakorlatot alakíthassanak ki a katalogizálási munka során, s evégett saját relációs terminusokat vezetettek be, mi több, ezeket szabványokban definiálták is (nemzetközi szabvány: ISO 2788, 1986, amerikai szabvány: ANSI/NISO Z39.19, 1994, [3]).

A tezauruszt alkotó lexikai egységek között pontosan definiált relációkat lehet felvenni, és ezen relációk mentén haladva lehet megtalálni azokat a lexikai egységeket, melyekkel leginkább lehet jellemezni a dokumentumok tartalmát. A tezauruszokon belül is a hierarchikus relációk a legfontosabbak. A nemzetközi szabványokban a hierarchikus relációknak három fajtáját különítették el egymástól:

a **generikus** relációt (generikus fölérendeltje/alárendeltje),
a **partitív** relációt (partitív alárendeltje/fölérendeltje), illetve
a **példánya** relációt.

Ez a hármas felosztás megegyezik az eddig tárgyalt relációtipológiával. A **generikus** reláció a **taxonimának**, a **partitív** reláció a **partonimának**, míg a **példánya** reláció a példánya relációnak felel meg. Mivel a tezauruszok hierarchikus relációinak korábbi általánosítása a szűkebb/tágabb terminus (Narrower Term/Broader Term – NT/BT) relációpárja volt, ezért a fenti hármas relációpárt rendre az alábbi rövidítésekkel jelölik: BTG, BTP, BTI, illetve NTG, NTP, NTI. A magyar gyakorlatban a **generikus** relációk jelölésére használják még az **A** és **F** jeleket (az Alárendelt és Fölérendelt szavak kezdőbetűit), míg a **partitív** relációkra a **P** és a **T** betűket (a Pars/rész, illetve a Totum/egész szavak alapján).

A tezauruszokban a *szinonimitást* az *ekvivalenciarelációk* sajátos típusaként kezelik, és a *lásd* (UF – use for) relációt, illetve a *helyette* (USE) inverzrelációt használják e célra. Itt kell megjegyeznünk, hogy a tezauruszokban kétféle lexikai egységet használnak: vannak *deszkriptorok* és *nem-deszkriptorok*. Mindkét terminustípust használják az indexelésben és a keresésben, de csak a deszkriptorok között építik fel a relációs kapcsolatokat, tehát csak ezek között lehet navigálni. A nem-deszkriptorok kizárólag a szinonimakapcsolat mentén vannak a tezaurusz elemei közé bekötve.

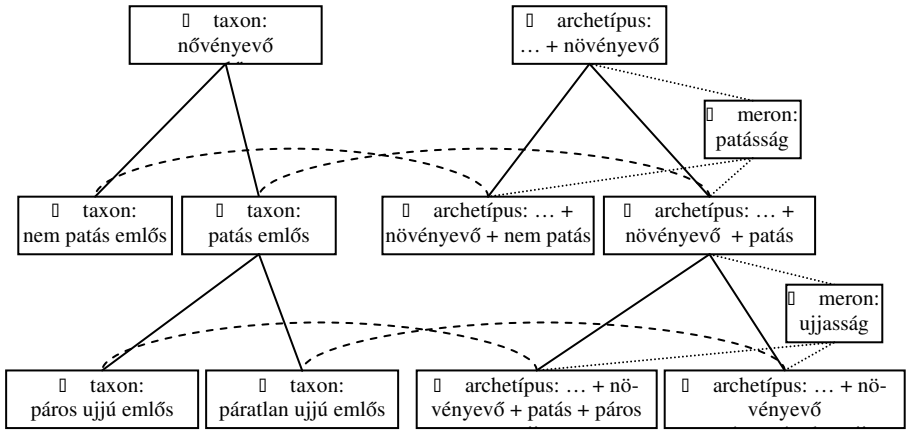
A hierarchikus kapcsolatokon és szinonimarelációkon túl minden más relációt (vagyis a „maradékot") az *asszociatív* reláció gyűjtőfogalma alá sorolták be (például az oksági, következési, megelőzési relációkat). A magyar tezauruszépítő gyakorlatban ezt rokonsági relációnak hívják. (Halkan jegyezzük meg: az 'asszociatív', a 'rokonsági' és az angol 'related term' név nem túl fantáziadús megnevezése az „egyéb relációnak").

***

Az eddig bemutatott, hasonló fogalmat jelölő, egymással versengő terminusok közül a következőkben a **taxonima** relációra a **generikus**, míg a **partonima** esetében a **partitív** reláció megnevezését fogjuk használni – egyrészt a szakterülethez kötődő

hagyományok önmagában vett értéke, másrészt az osztályozáselmélet alapok létezése miatt. J. Srejder ugyanis igen fontos megjegyzéseket tesz az osztályozási tevékenység

duális rendszerével kapcsolatban [16]. Šrejder szerint az osztályozás során úgy hozunk létre új, egymás alá rendelt részosztályokat (Šrejder terminológiájában: *taxonokat*), hogy emellett párhuzamosan és folyamatosan végzünk egy archetípusépítési munkát, amelynek során új és új részeket, tulajdonságokat, állapotokat veszünk fel, melyeket a leírandó új osztályokhoz (taxonokhoz) rendelünk. Šrejder új terminust vezet be erre: ez a *meron*. Az *archetípus* az általa jellemzett objektumok belső felépítését, tulajdonságait és külső kapcsolatait írja le, tehát voltaképpen különböző meronok értékeiből, állapotaiból áll össze. Amikor osztályozási tevékenységet végzünk, a taxonok formálásán, egymás alá- és mellérendelésén túl az archetípusokat és a meronokat is folyamatosan építjük, ezért hát a *taxonómia* és a *meronómia* (nem mereológia!) egyetlen duális rendszert alkot. Ezáltal újra fogalmazható az arisztotelészi elv is: minél speciálisabb jelentésű fogalomról, osztályról van szó, annál kisebb az osztály/taxon terjedelme, de annál bővebb az archetípus tartalma.



A meronok segítségével pontosabban leírhatjuk és értelmezhetjük a hiponima és inkompatibilitási reláció függetlensége, illetve a taxonima és hiponima különbsége kapcsán már jelzett problémát, ti. hogy az azonos taxon alá tartozó résztaxonok között egymást kizáró feltétel esetén beszélhetünk „igazi" generikus relációról. A meron fogalmának segítségével ezt a feltételt úgy fogalmazhatjuk meg, hogy amikor adott taxon alatt a generikus reláció révén hozunk létre új résztaxonokat, akkor a taxonhoz tartozó archetípus bővítéseként felvett új meron résztaxonokhoz rendelt állapotértékeinek különbözőnek kell lenniük.

Tanulmányunkban nem térünk ki rá, de itt jelezzük, hogy az OntoClean módszertan metatulajdonságokra támaszkodó ellenőrzések, mechanizmusok révén próbálja elősegíteni a taxonómiák konzisztenciájának növelését [7].

2.5. A relációnevek összehasonlító táblázata

A három nagyobb szakmai közösség relációhasználati gyakorlatának áttekintése után, mintegy zárásként, tanulságos lehet közös táblázatban feltüntetni a különböző terminusok és fogalmak egymáshoz való viszonyát (relációs és függvényes nevekkel):

		NYK-nevek	TIK-nevek	KOK-nevek
inklúzió, magába foglalás	alárendelt	hiponima	is-a, alárendelés	-
		taxonima	is-a-kind-of, is-a-type-of, is-a-sort-of, ako, részosztály, szubtípus, szubszumpció, specializáció	generikus alárendeltje, faja, szűkebb terminus, NTG, A
		troponima		
	fölérendelt	hiperonima	tartalmazás, fölérendelés, generalizáció, általánosítás	-
		inverz taxonima	szupertípus, szuperosztály	generikus fölérendeltje, bővebb terminus, BTG, neme, F
		inverz troponima		
rész-egész	alárendelt	partonima, meronima	része, is-part-of, dezaggregáció, komponensre bontás, dekompozíció	partitív alárendeltje, Pars, NTP, P
	fölérendelt	holonima	egésze, has-a, has-part, aggregáció, kompozíció	partitív fölérendeltje, Totum, BTP, T
példány-típus	alárendelt	-	is-a, is-element-of, eleme, tagja, példánya, instanciája, előfordulása	példánya, előfordulása, NTI
	fölérendelt	-	osztálya, típusa, is-class-of	fogalma, BTI

3. A relációk formalizálása

A történeti kitekintés után vegyük sorra a generikus és partitív reláció formalizálását.

3.1. Generikus reláció

A generikus reláció (G) osztályfogalmak között teremt kapcsolatot, amelyek a formális nyelvekben relációnak (függvénynek) számítanak. Ebből fakadóan a generikus reláció másodrendű nyelvvel írható le, tehát a generikus reláció másodrendű. A generikus reláció meghatározása és algebrai tulajdonságai a következők [6]:

(g1) $\Box\forall R\forall Q\forall x((R(x)\rightarrow Q(x))\wedge\neg(Q(x)\rightarrow R(x)))$ G generikus reláció

(g2) $\forall x\forall y(G(x,y)\rightarrow\neg G(y,x))$ G aszimmetrikus

(g3) $\forall x\forall y\forall z((G(x,y)\wedge G(y,z))\rightarrow G(x,z))$ G tranzitív

A konzisztens alárendeltségi viszonyok, a hierarchikus struktúra kialakításához szükség van egy különbözőségi relációra, melyet az alábbi módon határozhatunk meg:

(g4) $\forall x\forall y(D(x,y)\leftrightarrow x\neq y)$ D különbözőség reláció

3.2. Partitív reláció

A partitív reláció (P) formális definiálásakor több lépésben hajthatjuk végre, melynek során többféle mereológiai elméletet állíthatunk fel, fogadhatunk el. Először a **része** relációt alapját jelentő **parciális rendezés** relációt vesszük fel:

(p1)	$\forall x(P(x,x))$	P reflexív
(p2)	$\forall x\forall y(P(x,y)\wedge P(y,x)\rightarrow x=y)$	P antiszimmetrikus
(p3)	$\forall x\forall y\forall z(P(x,y)\wedge P(y,z)\rightarrow P(x,z))$	P tranzitív

A következő lépésben a P relációval néhány új mereológiai relációkat definiálunk, majd az újakkal még újabbakat hozunk létre:

(1)	$\forall x\forall y(PP(x,y)) := \forall x\forall y(P(x,y)\wedge\neg P(y,x))$	valódi része
(2)	$\forall x\forall y(O(x,y)) := \forall x\forall y\exists z(P(z,x)\wedge P(z,y))$	lefedése
(3)	$\forall x\forall y(U(x,y) := \forall x\forall y\exists z(P(x,z)\wedge P(y,z))$	lefedettsége
(4)	$\forall x\forall y(PO(x,y)) := \forall x\forall y(OX(x,y)\wedge OX(y,x))$	valódi lefedése
(5)	$\forall x\forall y(PU(x,y)) := \forall x\forall y(UX(x,y)\wedge UX(y,x))$	valódi lefedettsége

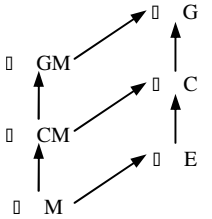
A része, lefedés és lefedettsége relációk segítségével először meghatározhatjuk a *kiterjeszthetőség (erős) elvét* (p4), majd a (p5)-(p7) mereológiai axiómák elfogadásával definiálhatjuk a *zárt (extenzionális) mereológia* elméletét:

(p4)	$\forall x\forall y(\neg P(x,y)\rightarrow\exists z(P(z,x)\wedge\neg O(z,y)))$
(p5)	$\forall x\forall y(U(x,y)\rightarrow\exists z\forall w(O(w,z)\leftrightarrow(O(w,x)\vee O(w,y))))$
(p6)	$\forall x\forall y(O(x,y)\rightarrow\exists z\forall w(P(w,z)\leftrightarrow P(w,x)\wedge P(w,y)))$
(p7)	$\forall x\forall y\exists z((P(z,x)\wedge\neg O(z,y))\rightarrow\exists z\forall w(P(w,z)\leftrightarrow(P(w,x)\wedge\neg O(w,y))))$

További axiómák felvételével egyrészt definiálhatjuk az általános (extenzionális) mereológia elméletét (p8), másrészt meghatározhatjuk az atomos (p9), illetve atom nélküli mereológiákat (p10):

(p8)	$\exists x\phi\rightarrow\exists z\forall y(O(y,z)\leftrightarrow\exists x(\phi\wedge O(y,x))$	általános (extenzionális) mereológia
(p9)	$\forall x\exists y(P(y,x)\wedge\neg\exists zPP(z,y))$	atomos mereológia
(p10)	$\forall x\exists y(PP(y,x))$	atom nélküli mereológia

A (p1)-(p8) axiómák segítségével tehát egymásra építhető mereológiai elméleteket hozhatunk létre. A köztük létező viszonyt az alábbi ábrával szemléltethetjük:



- mereológiai elméletek:
- GEM – általános extenzionális mereológia: (p1)-(p4), (p8)
- GM – általános mereológia: (p1)-(p3), (p8)
- CEM – zárt extenzionális mereológia: (p1)-(p7)
- CM – zárt mereológia: (p1)-(p3), (p5)-(p7)
- EM – extenzionális mereológia: (p1)-(p4)
- M – mereológia: (p1)-(p3)

A mereológia fenti axiomatizálása az utóbbi években egyre szélesebb körben elfogadottá vált, bár újabban az elmélet bizonyos pontjait érték kritikák (a kiterjeszthetőség erős vagy gyenge elvének felcserélhetőségével kapcsolatban) [11]. Az elmélet hasznosítási lehetőségeinek felkutatása a MEO-projekt fontos feladatai közé tartozik.

4. Bibliográfia

1. Brachman, Ronald J., What IS-A link and isn't: An Analysis of Taxonomc Links in Semantic Networks, *IEEE Computer*, 16 (10); October 1983, pp. 30-36.
2. Cruse, D. A., *Lexical Semantics*. New York: Cambridge University Press, 1986.
3. Dextre Clarke, Stella G., Thesaural Relationships, in: Bean, C.A., Green, R., *Relationship in the Organization of Knowledge*, Dordrecht-Boston-London: Kluwer, 2001, pp.37-52.
4. Fellbaum, Troponymy, in: Green, R. et al. (eds.), *The Semantics of Relationships: an Interdisciplinary Perspective*, Kluwer, 2002, pp. 23-34.
5. Green, R., Bean, C.A., Myaeng, S.H., (eds.), *The Semantics of Relationships: an Interdisciplinary Perspective*, Kluwer, 2002.
6. Guarino, N., Welty, C., Identity and subsumption, in: Green, R. et al. (eds.), *The Semantics of Relationships*, pp. 111-126.
7. Guarino, N., Welty, C., Supporting ontological analysis of taxonomic relationships, in: *Data & Knowledge Engineering,* 2001, (39), pp. 51-74.
8. Iris, Litowitz, Evens, Problems of the part-whole relation, in M. Evans (ed.) *Relational models of the lexicon*, Cambridge: Cambridge University Press, 1988.
9. Leonard, H.S., Goodman, N., Calculus of Individuals, in: *Journal of Symbolic Logic*, 1940 (5), pp.45-55.
10. Lyons, J., *Semantics, 1-2,* New York: Cambridge University Press, 1977.
11. Pontow, C., A Note on the Axiomatics of Theories in Parthood, in: *Data and Knowledge Engineering*, 2004 (50), pp. 195-213.
12. Pribbenow, S., Meronymic Relationships: From Classical Mereology to Complex Part-Whole Relations, in: Green, R. et al. (eds.), *The Semantics of Relationships,* pp. 35-50.
13. Pustejovsky, J., *The Generative Lexicon*, Cambridge, MA: MIT Press, 1995.
14. Saussure, F., *Bevezetés az általános nyelvészetbe*, Budapest: Corvina Kiadó, 1997.
15. Simons, P., *Parts: a Study in Ontology*, Oxford: Clarendon Press, 1987.
16. Šrejder J. A., Rendszerek és modellek, in: Ungváry Rudolf, Orbán Éva, *Osztályozás és információkeresés*, Budapest: OSZK, 2001, II. kötet, 297-325.
17. Varzi, A.C., Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology, in: *Data and Knowledge Engineering*, 1996 (20), pp.177-198.
18. Winston, M.E., Chaffin, R., Hermann, D.J., A taxonomy of part-whole relations, in: *Cognitive Science*, 1987 (11), pp. 417-444.
19. Winston, M.E., Chaffin, R., Hermann, D.J., An empirical taxonomy of part-whole relations: Effects of the part-whole relation type on relation identification, in: *Language and Cognitive* Processes, 1988 (31), pp. 17-48.

A cikk a NKFP-2/042/04.sz. Magyar Egységes Ontológia (MEO) projekt keretén belül jött létre.

Szerepfogalmak az ontológiákban - az OntoClean metodológia továbbfejlesztése

Szőts Miklós, Lévay Ákos

Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7
allbox@all.hu

Kivonat: az előadás az OntoClean módszertanból kiindulva a szereptípusú fogalmakat vizsgálja. Esettanulmányok alapján feltárja a szerep fogalomtípust definiáló metatulajdonságok megállapításának problémáit és különböző értelmezéseit. A pontosabb reprezentálás céljából elemzi a szerepek és a természetes fogalmak közti relációkat – azt a relációt, amely a szerepek függését fejezi ki, és azt, amely azt fejezi ki, milyen előfordulások játszhatják a szerepet. A szerepfogalmak elemzése rávilágít néhány olyan alapvető kérdésre, amelyet az ontológiaszerkesztés kezdetén el kell dönteni. Másrészt a szerepekkel összefüggő relációk standardizálása módot ad arra, hogy egy eljövendő ontológiaszerkesztő rendszer segítse a felhasználót az ontológia módszeres építésében. Kutatásunk célja úgy kiegészíteni a metodológiát, hogy az ontológiaszerkesztést és -kezelést segítő eszköztár alapja lehessen.

Bevezetés

Az OntoClean módszertant N. Guarino vezette csoport fejlesztette ki. Rendeltetése az, hogy az ontológia céljától és a konkrét ontológiai elkötelezettségektől független kritériumokat adjon ontológiák vizsgálatára, segítse ontológiák felépítését. Előnye nemcsak az, hogy egyetlen, praktikus is: jelentős eredményeket is értek el használatával, lásd pl. [8]-t. Ennek ellenére az alapos vizsgálat több problémát fed fel ([12]). Előadásunk feltár néhány problémát, és a metodológia továbbfejlesztését mutatja be egy fogalomtípus (a szerep) elemzése alapján.

Mivel az ontológia terminológiája nem kiforrott, először tisztáznunk kell szóhasználatunkat, amely a MEO projekt során alakult ki, l. [13]. Az ontológia egységei a **fogalmak**. A fogalomnak van tartalma és terjedelme: a **tartalom** a fogalom leírása (definíciója, jellemzése), mi feltételezünk valamilyen formális logikai nyelvet a tartalom leírására. A **terjedelem** az előfordulások halmaza. Logikai megközelítésünknek megfelelően a terjedelmet a lehetséges világokban értelmezzük. A logika terminológiájában az extenzió felel meg a terjedelemnek, a tartalom pedig az intenzió axiomatizálása. Az **egyed**ek (particular az OntoClean terminológiában) térben és időben elhatárolt entitások. Vannak fogalmak, amelyek előfordulásai egyértelműen egyedek (pl. `ember`, `revolver-esztergapad` stb.). Azonban másoknál, el-

sősorban pszichikai jelenségek, tulajdonságok, állapotok esetén, nehéz így értelmezni, bár [13] megmutatja, hogy lehet. Mi azonban inkább úgy mondjuk, hogy az *egyedek hordozzák az előfordulásokat*. Tehát pl. a `piros szín` előfordulásait hordozzák a piros színű dolgok, előfordulásaik nem maguk a dolgok, hanem színük. Ha két dolog színe megegyezik, ugyanazt a `szín` előfordulást hordozzák. Így értelmezhető lesz a `színe` reláció, és az olyan kifejezések, mint „*ugyanolyan színű*”.

A logikai leírásokban a fogalmakat relációkkal reprezentálják, a különböző ontológia leíró nyelvek különböző típusú relációkat engednek meg. A MEO projektben megengedjük az egyargumentumú relációkat (osztályfogalmak) és a kétargumetumúakat (relációfogalmak).

Az OntoClean (angol nyelvű) terminológiáját ismerteti pl. [5] és [6]. Az OntoClean a fogalmakat kizárólag egyargumentumú relációkkal jelöli (property-nek nevezi) – ezért a továbbiakban a fogalom szó az előadásban is osztályfogalmat jelent.

1. Az OntoClean módszertan – rövid ismertetés

Az OntoClean módszertan a következő összetevőkből áll:

- a fogalmakhoz rendelt metatulajdonságok,
- a generikus (subsumption, is_a, subclass, hypernym-hyponym) reláció alkalmazásának ellenőrzésére szolgáló szabályok, amelyek a metatulajdonságok definíciójából következnek,
- a fogalmak osztályozása a metatulajdonságok alapján,
- az osztályozáson alapuló felső szintű ontológia szerkezet.

Arra természetesen itt nincs mód, hogy a teljes módszertant ismertessük, [5] és [6] együtt alapos összefoglalás, [7] egy példán ismerteti használatát, [12] részletesen tárgyalja. A következőkben röviden áttekintjük azokat a fogalmakat, amelyek az előadás megértéséhez szükségesek.

Az OntoClean metodológia által bevezetett **metatulajdonság**ok a fogalmak tulajdonságai, tehát logikailag másodrendű, egyargumentumú predikátumok. A következő metatulajdonságokra lesz szükség a fogalmak osztályozásához:

- **Rigiditás**:
  Egy tulajdonság egy egyed **lényeges** tulajdonsága akkor és csak akkor, ha kötelezően áll rá (minden lehetséges világban és időpontban, ahol és amikor az egyed létezik). Egy fogalom **rigid** (**+R**) akkor és csak akkor, ha minden előfordulásának a fogalomhoz való tartozása lényeges tulajdonsága. Egy fogalom **antirigid** (**~R**) akkor és csak akkor, ha a fogalomhoz való tartozás egyik előfordulásának sem lényeges tulajdonsága. Ha egyik feltétel sem áll rá, a fogalom szemirigid (**¬R**). Az, hogy egy fogalom antirigid, öröklődik a generikus reláció szerint (a rigiditás nem).
  Példa: a `személy` fogalom rigid mert amíg egy személy él, addig személy[8]. Azonban a `tanár` antirigid, hiszen a tanárságot el kell nyerni, és el lehet veszíteni. Általában a foglalkozások, a családi állapotok (pl. `férj`) antirigidek, de

[8] példáink tükröznek valamilyen ontológiai elkötelezettséget – az előadás mondanivalójának szempontjából lényegtelen, ha valaki ezt nem osztja, és ezért az említett fogalmakhoz más metatulajdonságot rendelne.

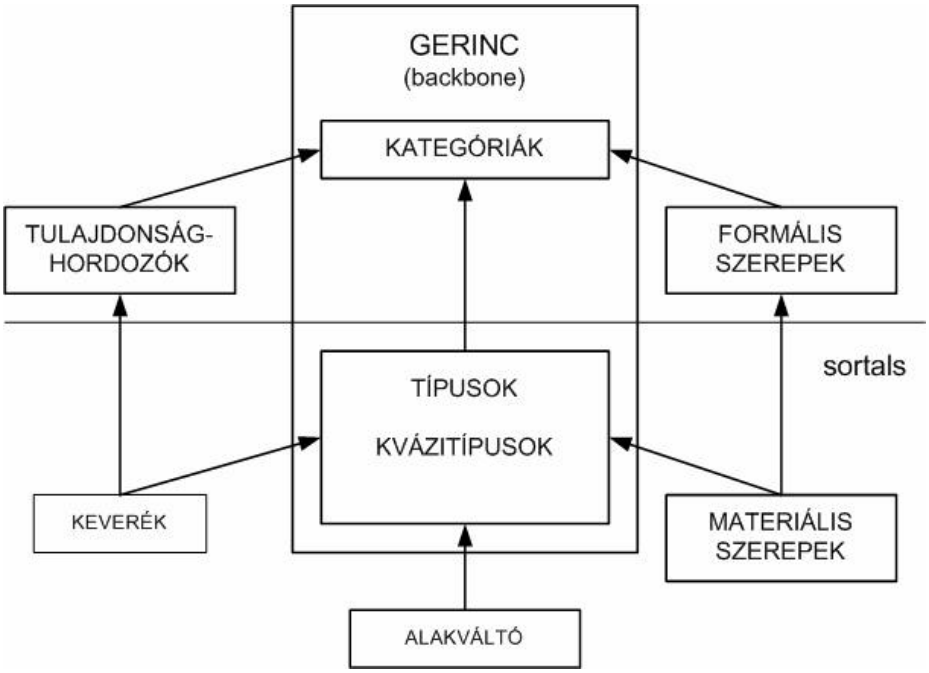
antirigid a munkanap fogalom is. A piros dolog fogalom szemirid, hiszen lehet olyan dolog, ami törvényszerűen piros (példa: az emberi vér), de a legtöbb nem ilyen.

- **Függés**:
  Egy A fogalom fogalmilag **függ** a B fogalomtól, ha A egy előfordulása csak akkor létezhet, ha létezik B egy előfordulása is. A **fogalmi** függés azt jelenti, hogy a fogalom tartalma (meghatározása) feltételez egy másikat. Tehát pl. semmi sem függ a részeitől, anyagától stb.; és a világ törvényszerűségei szerinti függést sem tekintjük fogalminak (tehát pl. fogalmilag a panda nem függ a bambusztól). Azt mondjuk, hogy egy fogalom **függő**, ha fogalmilag függ egy másik fogalomtól. Jelölés: **+D**, ill. **–D** ha nem.
  Példa: az ember nem függő fogalom, de a tanár függő (**+D**) (függ többek között a tanítástól, a tanított tárgytól és a tanítványtól). Hasonlóképpen függő fogalom a férj, de nem függő a munkanap fogalom.
  A függés is öröklődik, de ha nem függő egy fogalom, fajtája lehet függő.
- **Azonossági feltétel**:
  Az **azonossági feltétel** (identity condition, IC) olyan ismérv, amely alapján kimondható két egyed azonossága, ill. különbözőségük. Azaz az, hogy minek az alapján ismerek fel egy előfordulást. A félreértések elkerülése végett: nem arról van szó, hogy felismerjük-e, valami a fogalomhoz tartozik-e, hanem arról, hogy a fogalomhoz tartozó előfordulásokat a fogalomhoz kötött ismérv alapján azonosítjuk-e, különböztetjük-e meg. A fogalmak jellemzése szempontjából az érdekel bennünket, hogy egy fogalom előfordulásai rendelkeznek-e közös IC-vel. Egy fogalom **hordoz azonossági feltételt**, ha van egy azonossági feltétel, amely minden előfordulására alkalmazható. Jelölés: **+I**, **-I**. A tulajdonságot a fogalom azonossági feltételének nevezzük.
  Példa: a fizikai létező k azonossági feltétele lehet a téridőben elfoglalt hely-idő, az állatoké ezen felül pl. a DNS struktúrájuk; az anyagöszleté (amount of matter) pedig az, hogy tetszőleges részük megegyezik (mereológiai IC). Tehát a fizikai létező, állat, anyagöszlet fogalmak hordoznak azonossági feltételt. A munkanap fogalom szintén hordoz IC-t – a nap-tól örökli. Viszont a tulajdonság, ágens fogalmak nem.
  Látható, hogy valaminek általában több IC-je van. Az, hogy egy kritérium egy fogalom azonossági feltétele, öröklődik; így az is öröklődik, hogy egy fogalomnak van azonossági feltétele.

A metatulajdonságok által meghatározott fogalomtípusokat az 1. táblázat, az ez alapján megkonstruált ontológia szerkezetet az 1. ábra mutatja (l. [5] és [6]).

1. táblázat: fogalomtípusok

+I	**+R**	**±D**	típus (type), kvázi-típus (quasi-type)	„sortal”
	~R	**+D**	materiális szerep (material role)	
	~R	**-D**	alakváltó (phased sortal)	
	¬R	**±D**	keverék (mixin)	
-I	**+R**	**±D**	kategória	
	~R	**+D**	formális szerep (formal role)	
	~R	**-D**	tulajdonsághordozó (attribution)	
	¬R	**±D**		



1. ábra: ontológiaszerkezet
(a nyilak a kötelező öröklődést jelzik)

Az ontológia **gerinc**ébe kerülnek a rigid fogalmak, ezek megvalósítják az egyedek egy teljes osztályozását. Minden egyednek tartoznia kell egy típushoz. Sokan (pl. [3], [7]) a rigid fogalmakat **természetes**nek nevezik. Kategória a `tulajdonság`, ennek alárendelt típus pl. a `szín` (IC-je lehet a színekre jellemző hullámhossztartomány). Típus az `ember`, `gépkocsi`, stb. is.

A prototipikus példa a tulajdonsághordozóra a `piros színű dolog`, az alakváltóra a `hernyó` (ha a hernyót és az abból átalakuló lepkét egy egyednek vesszük), de a `munkanap` szintén ebbe a fogalomtípusba tartozik. A keverékre N. Guarino

publikációi nem hoznak értelmes példát, de a `nőtlen` jó példa rá, ha ezt a tulajdonságot gyermekekre is alkalmazhatónak vesszük.

A szerepekkel fogunk részletesen foglalkozni, a 2.1 szakaszban hozunk példákat erre a fogalomtípusra.

A kevés számú fenti példából látható az, hogy a metatuljadonságok alkalmazása, és a fogalomtípusba sorolás ontológiai elkötelezettségektől függ. Ugyanakkor látszik az is, hogy a fogalomtípusokra bevezetett elnevezések által tükrözött intuíció nem mindig felel meg a metatulajdonságok szerinti definíciónak.

Ahogy a bevezetésben említettük, az OntoClean metodológia alapos vizsgálata több problémát fed fel ([12]). Az OntoClean metodológia jelenlegi – ill. az általunk ismert jelenlegi – állapotának legfontosabb hiányosságai:

- a fogalomtípusok definíciója finomítást igényel;
- a metodológia nem foglalkozik a relációkkal.

Jelen előadás a **szerep** fogalomtípus részletes elemzésével foglalkozik. Bár épp a szerep az OntoClean egyik legjobban jellemzett fogalomtípusa, ennek pontosabb elemzését elsőrendűen fontosnak tartjuk gyakorlati fontossága miatt. Másrészt a relációknak a metodológiába való beemelését ezen a témán jól illusztrálhatjuk.

2. A szerep fogalomtípusról

A szerep fogalmát nem az OntoClean hozta a köztudatba, már régen a tudásreprezentáció (pl. [3]) és a modellezés (pl. [11]) bevett, de nem egyértelmesült fogalma.

A továbbiakban feltételezzük, hogy az egyedek típusok előfordulásai (azaz rigidek), és a következő szóhasználattal élünk, függetlenül attól, hogy mennyire tudatos egy egyed és mennyire „önként” lesz egy szerep előfordulása:

- egy egyed egy szerepet **játszik**, ha előfordulása egy szerepfogalomnak,
- egy egyed **felvesz**, ill. **letesz** egy szerepet, amikor elkezdi, ill. abbahagyja egy szerep játszását.

Bevezetőjében [11] egy alapos áttekintést ad arról, hogy a szerepnek milyen tulajdonságokat tulajdonítanak. Néhány azok közül, amelyek minket érintenek:

1.) A szerepeknek vannak saját tulajdonságaik.
2.) A szerepek csak valamilyen viszony kontextusában értelmesek.
3.) Egy egyed dinamikusan felvehet ill. letehet szerepet.
4.) Különböző típusok egyedei játszhatják ugyanazt a szerepet, sőt szerep játszhat szerepet.
5.) Feltételekhez köthető, hogy egy egyed játszhat-e egy szerepet.

Egyes tulajdonságokat lefed az a tény, hogy felvesszük az ontológiába a szerepfogalmakat (1.), és az OntoClean féle osztályozást alkalmazzuk: a szerepek függőek (2.) és antirigidek (3.). Azonban az OntoClean metodológia jelenlegi eszköztára nem teljesen alkalmas a 4. és 5. tulajdonság kifejezésére.

Az OntoClean metodológia megkülönbözteti a formális és a materiális szerepet. A különbség a metatulajdonságok szerint abban van, hordoz-e a fogalom azonossági feltételt. Ha a szerep modellezés szempontjából való jellemzésre fordítjuk le ezt a

különbséget, azt jelenti, hogy meg tudjuk-e adni azt a típust, amely egyedei játszhatják a szerepet, vagy több típusból is jöhetnek (v.ö. 4. tulajdonsággal).

A továbbiakban a következő kérdésekkel foglalkozunk:

- mennyire találó és alkalmazható az OntoClean szerepdefiníciója,
- hogyan reprezentálhatóak a szerepek tulajdonságai.

2.1. Néhány esettanulmány

Az OntoClean publikációk (pl. [6], [7]) egyik kedvenc példája a materiális szerepre az élelem[9] (food) fogalom. A következő metatuljadonságokat rendeli hozzá:

- **~R**, azaz antirigid, mivel „semmi sem szükségszerűen élelem”,
- **+D**, azaz függő, az evés eseményétől függ,
- **+I**, mert az anyagöszlet fajtájának veszi, és attól örökli a mereológiai azonossági feltételt.

Látható, hogy N. Guarino szerint élelem az „*amit valaki valamikor ténylegesen megeszik*”. Ugyanakkor ez az értelmezés nem illik az élelem fogalomhoz: mondhatunk olyat pl., hogy az „ételt senki sem ette meg”, és itt ugyanazt a fogalmat fedi az étel szó mint az élelem, amelyet jellemezni akarunk. Továbbá ételt főzünk akkor is, ha még nem ette meg senki.

Másik példa az építőanyag fogalma, amelyet első, ösztönös reakcióval materiális szerepnek vélünk. Kétféleképp érthető:

építőanyag$_1$: mindaz, amit építménybe beépítenek(ettek);

építőanyag$_2$: mindaz, amit azért gyártanak, forgalmaznak, hogy építménybe beépítsék.

Az első esetben az építőanyag fogalom nyilvánvalóan antrigid (**~R**) és függő (**+D**). Ugyanakkor nem hordoz azonossági feltételt (**-I**), mivel anyagöszlet és formával rendelkező tárgy fajtái egyaránt előfordulnak benne: pl. homok és vasbeton gerenda. Tehát **formális szerep** – ez részben megfelel intuíciónknak, sőt ha összehasonlítjuk a prototipikus formális szereppel, azaz az ágens-sel, megerősíti az OntoClean osztályozás eredményét: az ágens is azért lesz „formális”, mert a legkülönbözőbb egyedek lehetnek előfordulásai.

A második értelmezésnél a függés „szándékbeli”. Hajlunk arra, hogy függésnek fogadjuk el így is: építkezés nélkül nem lenne építőanyaggyártás és -kereskedelem. A problémát a rigiditás okozza, mivel az építőanyag fajtái különbözőképp viselkednek. A homok mint építőanyag nyilvánvalóan antirigid, de az építőanyag céljából gyártott tégla, vasbeton gerenda már rigid fogalom. E szerint az építőanyag már szemirigid lenne, így nem lehetne szerep. Márpedig intuíciónk szerint annak kell lennie.

Az igazi kérdés az, hogy melyik értelmezést akarjuk az ontológiában az építőanyag fogalomnak adni. Az első (építőanyag$_1$) minden esetre a jobban kezelhető: nemcsak az OntoClean metodológiában, de leíró logikában ([1]) is: definiálható az építmény fogalom és az áll_belőle reláció segítségével: építőanyag≡∃áll_belőle.építmény. Mindenképp értelmezett az áll_belőle reláció az építőanyag és az építmény fogalmak közt, de a

[9] az élelem, étel, ennivaló szavakat ebben a kontextusban szinonimáknak tekintjük.

különböző módon: az első esetben az építőanyag$_1$ terjedelme pontosan a reláció értelmezési tartománya, az építőanyag$_2$ terjedelme viszont az értelmezési tartományt csak tartalmazza.

Vegyük észre, hogy az élelem fogalmánál is a fenti kettősség jelentkezett. Ha az élelmet úgy definiáljuk, hogy élelem az, *amit annak készítenek el*, intuíciónknak jobban megfelelő fogalmat írunk le, amely a fentebb felvetett szóhasználatba nem ütközik, de a metatuljadonságokkal való felékesítése ugyanazokat a problémákat okozza, mint amelyeket az építőanyag második értelmezése esetén láttunk.

A fenti esetekben (és bárki még akárhányat generálhat) van egy fogalmunk, amely intuíciónk szerint szerep, de a metatulajdonságok szerinti osztályozás szerint más fogalomtípusba esik.

Két problémába ütközünk:

- a függés értelmezése,
- a szemirigiditás.

A **függés** kérdése emlékeztet a *nem monoton logikák* megszületésének szituációjára. Akkor a problémát az okozta, hogy a „madarak repülnek" generikus állítást kellett következtetésekben való használat céljából a logika nyelvére fordítani, ahol is a „minden madár repül" mondat lett belőle. Most is megfogalmazható úgy a probléma, hogy „az élelem az, amit megeszünk" definíciót generikusan értjük, ezért a függést értelmezhetjük generikusan is. Azaz tipikusan megesszük.

A fenti példákban a probléma úgy jelentkezik, hogy pl. az építőanyag az építés-től függ nyilvánvalóan, amikor előfordulásait beépítik egy építménybe, de ugyanakkor a függést értelmezzük akkor is, amikor már be van építve, és amikor még nincs. Innen már csak egy lépés a függést értelmezni akkor is, ha nem vagyunk biztosak abban, hogy be fog épülni. A megkülönböztetést az épp fennálló függés és a generikusan fennálló közt megtalálhatjuk Sowa csúcsontológiájában [10] is, amikor a role kategória alatt szerepel a participant, amely utóbbi terjedelme azok az előfordulások, amelyek valamilyen eseményszerűségben[10] résztvesznek (így függenek tőle). Példája a sofőr, amely terjedelme az épp gépkocsit (busz, teherautó stb.) vezető személy, aki a participant előfordulása, és a hivatásos autóvezető, aki a roles előfordulása, de a participant-é csak akkor, ha vezet. Ugyanígy értelmezhető pl. a tanár esetén a kettősség: „az, aki éppen tanít" vs. „a tanár foglalkozású" személy. Természetesen nem tartjuk szükségesnek ezt a kettősséget általában bevezetni az ontológiákba, de tudnunk kell róla. Viszont speciális célokra szolgáló fogalmi sémák esetén természetes ez a megkülönböztetés, pl. a rendőrség tudásbázisában, ahol feladat az egyes balesetek, szabálysértések elemzése. Vegyük észre, hogy a „participant" értelmezésű szerepfogalmakat a leíró logika segítségével automatikusan képezni lehet – már csak ezért sem érdemes a fogalmakat megkettőzni az ontológiában.

A **szemirigiditás** komolyabb probléma. Itt nem használ az a megközelítés, hogy „általában", vagy „tipikusan" antirigid az élelem – mert épp a tipikus esetekben nem az.

A fenti probléma oka az, hogy az emberi civilizáció bizonyos szerepekre speciális termékeket állít elő: vasbeton gerendát az építmények födémének kiképzésére, pa-

[10] az eseményszerűség az esemény, folyamat stb. közös neme, az angol eventuality szakszó bevett fordítása; használják még rá az occurent, perdurant terminusokat.

prikás csirkét ebédre és kalapácsot a szögek beverésére. Ezek a tárgyak és az általuk játszott szerepek összemosódnak tudatunkban.

Ezzel a problémával a tudásreprezentáció és a modellezés már régóta küszködik, és pontos megoldása a szerepek és a szerepet játszó tárgyak elkülönítése. [3] egy olyan rendszerről számol be, ahol ezt szigorúan végrehajtották: külön szerepel a `kalapács` és a `kalapács szerep`. Ebben az esetben is, mint más problematikus esetekben, figyelembe kell venni, hogy milyen célra készül az ontológia. Például ha a cél információkeresés, a „kalapács szerep" fogalom érdektelen, viszont természetesnyelv-feldolgozás esetén esetleg kezdeni kell tudnunk valamit olyan kifejezésekkel is, mint „kalapácsnak használta a cipőjét".

Van egy másik probléma is a szerepekkel, mégpedig az, hogy minden materiális szerepnek kell egy általánosabb fogalomhoz tartoznia, a megfelelő formális szerephez. Sokszor ez nem probléma: pl. a tanár az `ágens` fajtája. Az `építőanyag` és az `élelem` esete már fogósabb kérdés: talán az `eszköz` fajtái („valamivel elverte az éhét"). Viszont mit kezdjünk `férj` fogalommal? Tipikus materiális szerep, de mi lenne a formális megfelelője?

2.2. A szereptípusú fogalmak relációs kapcsolatai

Az ontológiákban az osztályokat kifejező fogalmakat relációk strukturálják. Elsőrendű fontosságú a generikus reláció (részosztály, is_a stb.), de elengedhetetlen más relációk használata. Ha a relációkat gondosan választjuk meg, tisztán a relációk alkalmazásával jelentős mértékben jellemezni tudjuk a fogalmakat.

A szerepek esetén két fontos relációtípusról kell szólni:

- a szerep függését kifejező relációkról,
- a „role filler" vagy „játszhatja" („played by") relációról.

2.2.1. A függésről

A szerepeket elsősorban függő voltuk különbözteti meg.

Vegyünk megint először egy példát: a `tanár` szerep, mert függ pl. a `tanuló`tól, a `tananyag`tól vagy a `tanítás`tól.

A fenti függések nem egyformák: a `tanuló` éppúgy függ a `tanár`tól (és a `tanítás`tól), viszont a `tanítás` nem függ az előzőktől[11], valamint a `tanítás` eseménye rigid. Így a `tanítás` típus lesz, míg a `tanuló` szerep. Hasonlóan minden foglalkozás (tipikus szerepfogalmak) függ valamilyen – egy, vagy több – eseményszerűségtől.

Tipikus szerepfogalmak a különböző eszközök, pl. `kés`[12] függ `vágás`tól, `fegyver` a `katoná`tól és a `harc`tól. Ezek is mindig függnek valamilyen eseményszerűségtől[13].

---

[11] természetesen az utóbbi állítást a külső függés megfelelő definíciója teszi igazzá.

[12] pontosabban: a `kés szerep` – emlékezzünk a 2.1-ben mondottakra a szerepre készített tárgyakról.

[13] a függés nem kizárólagos, nem függvény reprezentálja, hanem reláció; tehát a fegyver függhet nemcsak a harctól, hanem sok más eseményszerűségtől: a vadászattól, gyilkosságtól stb.

Másik tipikus szerepfogalom-összesség a családi szerepek: férj, anya stb. Ezek a fogalmak tipikusan egymástól függnek: a férj a feleségtől, anya a gyermektől stb. – legalábbis ez a legnyilvánvalóbb függés. Azonban ez az egymástól való függés nem túlságosan informatív – és valóban van típus is, amelytől függenek:

- a férj és feleség a házasságtól – a házasság, amíg fennáll házasság, tehát rigid, azaz típus;
- az anya és a gyerek a szülés eseményétől.

Ilyen példák alapján a következő ismérveket adunk a szerepek jellemzéséhez:

- *függenek valamilyen gerincbeli fogalomtól;*
- *a függést kifejezhetjük egy relációval (szerepreláció).*

Példánkban tanítás a típus, amelytől függ mind a tanár, mind a tanuló; sőt a tananyag is. A függést meg is nevezhetjük: a tanár a tanítás **ágense**, a tananyag a **témája**, a tanuló a **kedvezményezett**[14]. Ha a tipikus szerep fogalmakat végigvesszük, jelentős részéről találjuk úgy, hogy függ valamilyen eseményszerűségtől: ilyen minden foglalkozás, de ilyen volt az előző pontban elemezett építőanyag és élelem is. Tehát a szerepek fontos osztályára kapunk a szemantikus értelmezést segítő rendszert:

- *az eseményt, folyamatot jelentő fogalmaktól függő szerepek megfelelhetőek a nyelvészetből ismert **tematikus szerepek**nek.*

Természetesen nem vehető át automatikusan egy nyelvészetben kidolgozott tematikus szerep készlet – mert nem mondatok értelmezése, hanem az ontológia fogalmai tartalmának megfogalmazása a célunk. Igényeinkhez, legközelebb az R. Jackendoff (l. [14]) és a J.F. Sowa (l. [10]) által kidolgozott rendszer áll – a MEO projekt keretében most dolgozunk egy olyan rendszeren, amely az ontológia követelményeit legjobban kielégíti. Mind a nyelvészeti, mind a Sowa féle tudásreprezentációs szemléletben a szerepek mint fogalmak jelennek meg: Mi azonban, visszatérve a forráshoz, azaz Davidson eseményszerűségek leírására szolgáló sémájához ([2]), *azokat a relációkat kívánjuk rendszerbe foglalni, amelyek az egyes eseményszerűséghez kötik az egyes szerepeket játszó egyedeket osztályba foglaló fogalmakat*.

A tanár-tanuló függés természetesen létezik, de **származtatható** az eseményszerep típusú függésekből: ágense•kedvezményezettje^{-1}. Sajnos találó elnevezést erre a viszonyra nem találtunk. Viszont más esetekben természetesen adódik elnevezés a származtatott relációkra. Vegyük a harc, katona és fegyver esetét: a fegyver a harc eszköze, a katona a harc ágense és az eszköze•ágense^{-1} reláció szintén egy eszközt kifejező reláció, csak nem a cselekvésé, hanem a cselekvőé, a pontosság kedvéért eszköze' jelöléssel megkülönböztetjük a tematikus szerep eszköz relációtól. Az itt bevezetett eszköze' reláció nemcsak ebben a példában használható, hanem általában: pl. míg az ecset és festés fogalmak terjedelme közt az eszköze reláció, az ecset és a festő fogalmak terjedelme közt az eszköze' reláció értelmezett, stb. Nem állítjuk, hogy a kétféle eszközt kifejező relációt az ontológiákban okvetlen különböző névvel kell szerepeltetni, hiszen az értelmezési tartomány megkülönbözteti ezeket. Azonban tudni kell a különböző-

[14] beneficiens – nem a Parson-féle, angol nyelvhez kötött szintaktikus ismérvet (l. [9]) vettem figyelembe, hanem a fogalmi tartalmat.

ségről, és lehetséges, hogy az ontológia céljától függően a logikai reprezentáció megköveteli a megkülönböztetést. A két példát a 2. ábra illusztrálja.

2. ábra: a függést kifejező relációk
(a szaggatott nyilak a származtatott relációkat jelölik)

Sokszor számszerű megszorítást tudunk adni egy eseményszerűségben egy szerepet játszó előfordulásokra: pl. a `sakkozás`nak pontosan két ágense van. Bár ez az információ inkább a `sakkozás` jellemzéséhez tartozik, mint a `sakkozó`éhoz.

A függést kifejező viszony a generikus reláció szerinti **öröklődik**. Példa: a `matematika tanár ágense` lesz egy eseményszerűségnek, és csak a `tanítás`, vagy annak fajtája lehet (`matematikatanítás`). Ezt az öröklődés fordítva is fennáll: a `matematika tanítás ágense` csak a `tanár` vagy annak fajtája lehet. Az öröklődés ellenőrzése segíthet az ontológiák szerkesztésében.

2.2.2. A „játszhatja" reláció

A `játszhatja` reláció azt határozza meg, hogy mely előfordulások játszhatnak egy szerepet. Bár a szerepek általános ismertetésénél említettük, hogy szerep is állhat szereppel `játszhatja` kapcsolatban (pl. csak `katona` lehet `tiszt`), minden szerepből egy `játszhatja` relációkból álló lánc vezet gerincbeli fogalmakhoz. (A `játszhatja` reláció nyilvánvalóan tranzitív). Több ontológiában szemantikai hibához vezetett, hogy a `játszhatja` reláció helyett a generikus relációt alkalmazták, pl. az `aktor` alá rendelték a `gép` és `ember` fogalmakat.

Van viszont olyan eset, amikor megengedhető a generikus reláció használata erre a célra: amikor azokat az egyedeket, amelyek a materiális szerepet játszhatják, egy fogalom terjedelmében találhatjuk. Például `tanár` csak `ember` lehet, tehát a generikus reláció használata az ember és tanár fogalmak közt megengedhető. Ungváry Rudolf ezt a használatot „kvázigenerikus reláció"-nak nevezi.

A fogalmak OntoClean szerinti osztályozásában a formális és materiális szerepet az különbözteti meg, hogy a fogalom hordoz-e azonossági feltételt. Az azonossági feltétel definíciója miatt a materiális szerep csak egy típustól örökölheti az azonossági feltételt. Tehát ez a megkülönböztetés a szerep általános terminológiájában pont azt jelenti, hogy azokat az egyedeket, amelyek a materiális szerepet játszhatják, egy fogalom terjedelmében találhatjuk. Így az OntoClean metodológia ezekben az esetekben nemcsak megengedi, de feltételezi is a `játszhatja` reláció helyettesítését a generikus relációval. Erre pont az öröklés miatt van szüksége. Ha mereven szét akarjuk választani ezeket a relációkat, meg kell különböztetni azokat az eseteket, amikor a `játszhatja` reláció szerint öröklődhetnek a fogalom tulajdonságai.

Viszont a formális szerepek esetében a szerepet játszó egyedeknek nincs közös azonossági feltételük, tehát nincs olyan típus, amely előfordulásai közül kerülhetnének ki a szerep előfordulásai. Ebben az esetben egyértelműen csak a `játszhatja`

reláció jelölheti ki azokat a típusokat, amelyek egyedei játszhatják a szerepet – ha erre szükség van.

Nyilvánvalóan a játszhatja reláció is öröklődik, éppúgy, mint a függést kifejező.

3. Az ontológia szerkezetét érintő kérdések összefoglalása

A szerepfogalmak elemzése néhány olyan kérdést vet fel, amelyeket az ontológiában magas szinten, és minél korábban el kell dönteni. Ezek a kérdések felsorakoznak más alapvető kérdések mellé (tulajdonságok kezelése, anyag fogalmak értelmezése stb.).

Az előadásban a következőket vetettük fel:

- elkülönítjük-e a szerepet a rendeltetése szerint azt a szerepet játszó tárgytól? Példánk volt a `kalapács` vs. `kalapács szerep`.
- megkülönböztetjük-e az eseményszerűségekben ténylegesen résztvevő szerepeket azoktól, amelyek rendeltetés szerint résztvesznek(vehetnek) bennük? Ez megfelel Sowa (l. [10]) `participant` vs. `roles` megkülönböztetésnek.
- felvesszük-e a függést kifejező relációkat, kialakítunk-e egy standard relációrendszert erre a célra?
- feltüntetjük-e a származtatott függést kifejező relációkat, és melyeket?
- hogyan kezeljük a „játszhatja” relációt?

Nem véljük úgy, hogy van ezekre a kérdésekre általánosan jó válasz. Az ontológia szerkezetét attól függően kell kialakítani, hogy milyen célra szánjuk. Más követelményei vannak az információkeresésnek, modellezésnek stb. – és a válaszokat a követelmények alapján kell kialakítani.

Természetesen az OntoClean más fogalomtípusai ugyanennyi problémás kérdést vethetnek fel, sőt, többet is.

4. Vízió

Bár az OntoClean metodológiát elsősorban ontológiák vizsgálatára alakították ki, használható (és használt is) ontológiafejlesztésre. Kutatásunk célja a fejlesztés segítése. A végső cél egy ontológiaszerkesztést és -kezelést segítő eszköztár. A generikus reláció ellenőrzése adott – ezt több ontológiaszerkesztőbe be is építették már. Azzal, hogy a metodológiába más relációk kezelését is bevonjuk, további lehetőségek adódnak. Az egyes fogalmakat a hozzájuk kapcsolódó relációkkal együtt vihetjük fel, illetve a rendszer felkínálja ezeket, ellenőrzi az öröklődésre vonatkozó szabályok szerint.

A szerepek elemzése alapján a következő szolgáltatásokat álmodjuk meg:

- ha egy szerepnek kategorizált fogalmat iktatunk be az ontológiába, a szerkesztő felkínálja, hogy határozzuk meg a függést kifejező relációt és annak értelmezési tartományát. Ha a szerep közvetlen neménél már meghatároztuk, azt a relációt, és annak értelmezési tartományát felkínálja. Ugyanígy a játszhatja relációval. Öröklődésüket ellenőrzi.

- ugyanígy jár el olyan gerincbeli fogalmaknál, amelyektől függő szerepek feltehetően vannak – elsősorban az eseményszerűségeknél. Azaz felkínálja az egyes szereprelációkat, és azoknak a tartományát.
- a függést kifejező relációknál generálhatja és felkínálhatja a származtatott függést kifejező relációkat.

Bibliográfia

1. Baader, F., W. Nutt: Basic Description Logics, in **The Description Logic Handbook** eds. F. Baader, D.L. McGuinness, D, Nardi, P.F. Patel-Schneider, Cambridge University Press, 2003
2. Davidson D.: The Logical Form of Action Sentences, in: Roberto Casati & Achille C. Varzi (eds.), **Events**, Ashgate, 1996.
3. Fan, J. & tsai.: Representing Roles and Purpose, *K-CAP'01*, Oct. 22-23, 2001, Victoria, British Columbia, Canada
4. Guarino, N.: Concepts, Attributes and Arbitrary Relations, *Data & Knowledge Engineering* 8 1992
5. Guarino, N., C. Welty: A Formal Ontology of Properties, in R. Dieng ed. *Proceedings of 12th Int. Conf. on Knowledge Engineering and Knowledge Management* 2000 Springer Verlag LNCS
6. Guarino, N., C. Welty: Supporting ontological analysis of taxonomic relationships, *Data & Knowledge Engineering* 39, 2001
7. Guarino, N., C. Welty: An Overview of OntoClean. in S. Staab, R. Studer (eds.): **Handbook of Ontologies** Springer-Verlag, 2004
8. Oltamari A. & tsai.: Restructuring Wordnet's top-level, *AI Magazine* 2003.
9 Parson, T.: **Events in the Semantics of English (A Study in Subatomic Semantics)**, MIT Press, 1990
10. Sowa, J.F.: Thematic Roles URL=<http://www.jfsowa.com/ontology/thematic.htm>
11. Stemainn, F.: On the representation of Roles in Object-Oriented and Conceptual Modelling, *Data and Knowledge Engineering,* 35(1): pp. 83-106, 2000
12. Szőts M.: Az OntoClean metodológia ismertetése, problémái és továbbfejlesztési lehetőségei, URL=<http//ontologia.hu/Members/szots/OntoClean.pdf>
13. Ungváry R.: A fogalom fogalma URL=<http://ontologia.hu/forum/MEO_forum_toplevel_ontology/fogalmi_definicio/569481451164/view>
14. Varasdi K.: Konceptuális reprezentációk a lexikonban, in: Kálmán L., Trón V., Varasdi K. (szerk.) **Lexikalista elméletek a nyelvészetben** Budapest, Tinta Kiadó, 2002. (Segédkönyvek a nyelvészet tanulmányozásához 13.)
15. B. Weatherson: Intrinsic vs. Extrinsic Properties, *The Stanford Encyclopedia of Philosophy* (Fall 2004 Edition), E.N: Zalta (ed.), URL=<http://plato.stanford.edu/archives/fall2004/entries/intrinsic-extrinsic/>

Köszönetnyilvánítás: a kutatatás, amelynek néhány eredményéről itt beszámolunk, a NKFP-2/042/04.sz. MEO (Magyar Egységes Ontológia) projekt keretein belül folyt. A projekten dolgozó kollektívának köszönettel tartozunk. Név szerint ki kell emelnünk Szakadát Istvánt, aki az OntoClean metodológiára irányította figyelmünket, Varasdi Károlyt, aki segített a tematikus szerepek zátonyai közt navigálni, és utoljára, de nem utolsósorban Ungváry Rudolfot éles kritikáiért.

Magyar EuroWordNet projekt: bemutatás és helyzetjelentés

Miháltz Márton

MorphoLogic, Orbánhegyi út 5, 1126 Budapest
mihaltz@morphologic.hu

Kivonat: A tanulmányban bemutatjuk azt a projektet, melynek célja a magyar nyelvű, a EuroWordNet többnyelvű architektúrájába illeszkedő nyelvi ontológia létrehozása. Az ontológia általános része a EuroWordNet-et továbbfejlesztő BalkaNet projekt erőforrásaira épít. Az ontológia kiinduló fogalmi készlete főneveknél és mellékneveknél a BalkaNet Base Concept Set angol nyelvű, Princeton WordNet-ből származó synsetjeinek lefordításával készült, igéknél ezekkel párhuzamosan—a két nyelv igei rendszerének szemantikai különbségei miatt—saját erőforrásokból kiindulva történt. A synsetek lefordítása gépi heurisztikák alkalmazásával, valamint ezek eredményeinek kézi ellenőrzésével történt. A cikkben bemutatjuk az eddigi eredményeket, illetve az ontológia továbbfejlesztésének a projekt során tervezett következő lépéseit.

1 Bevezetés

Természetes nyelvi szövegek gépi feldolgozásában mára vitathatatlanul fontos szerep jutott az ontológiáknak. A legismertebb, nyelvi tudást rendszerező ontológia a WordNet, melyet az 1990-es évektől kezdtek el fejleszteni, először angol nyelvre [9]. A gépi fordítás és az egyéb, több nyelvet kezelő nyelvtechnológiai alkalmazások számára további segítséget jelentenek a többnyelvű, az eredeti angol WordNet anyagához egyéb nyelvű ontológiákat kapcsoló nyelvi erőforrások, melyek első képviselője a EuroWordnet (EWN) projekt volt [14]. Az 1999-ben zárult munka eredménye az angolon kívül 7 európai nyelvre kifejlesztett és összekapcsolt WordNet ontológia volt. A BalkaNet projekt ennek továbbfejlesztése volt 2004-es befejezéssel, további 5 délkelet-európai nyelv bevonásával [13].

Magyar nyelvű, a EWN-hez kapcsolódva többnyelvűséget biztosító, használható méretű és minőségű WordNet ontológia fejlesztésére a GVOP-AKF-2004-3.1.1 pályázati projekt keretei között nyílt lehetőség, három, magyar nyelvtechnológiában vezető intézmény (MorphoLogic Kft., MTA Nyelvtudományi Intézet, Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportja) részvételével, a 2005-2007-ös időszakban. Ebben a tanulmányban ezt a jelenleg is futó projektet szeretnénk bemutatni, az eddigi eredményeket és a további munkát megismertetni.

A továbbiakban először kivonatosan bemutatjuk a WordNet-típusú ontológiák alapfogalmait, a többnyelvű EuroWordNet koncepciót, valamint ennek legutóbbi

megvalósulását, a BalkaNet projektet, melyek kiindulási anyagként szolgáltak a projekthez (2. fejezet). A 3. fejezetben bemutatjuk a magyar EWN ontológia létrehozásának tervezett metodológiáját és a projekt során tervezett lépéseit, majd a 4. fejezetben a cikk elkészültéig megvalósult eredményeit, kitérve az ezen idő alatt felmerült problémákra és a munka közvetlen folytatására.

2 Egy- és többnyelvű wordnetek

2.1 Princeton WordNet

A mentális lexikont, ezen belül az angol nyelv lexikális és fogalmai viszonyait modellező Princeton WordNet (PWN) lexikális szemantikai hálózatot George Miller és munkatársai a Princeton Egyetem Kognitív Tudomány Laboratóriumában, pszicholingvisztikai kísérletek eredményeiből kiindulva fejlesztették ki [9]. A *wordnet* köznév azóta az eredeti, Princeton-ban készült angol nyelvű WordNet felépítését követő nyelvi adatbázisokra utal.

A WordNetben a tartalmas szavak (főnevek, igék, melléknevek, határozószók) különböző értelmeit szójelentéseknek hívják. A szinonimitás jelenségére—egyes szavak bizonyos értelemben, egy adott kontextusban a (denotációs) jelentés megváltoztatása nélkül felcserélhetők—épülnek a synsetek (szinonima-halmazok), a WordNet fogalmi alapegységek. A WordNetben egy fogalom tehát ekvivalens szójelentések halmazával reprezentálható (pl. {léc, deszka}, {fut, szalad, rohan}, {helyes, hibátlan} stb.).

A synsetek között különböző, világismereti, illetve nyelvi kapcsolatot kifejező szemantikai kapcsolatok (relációk) vannak, melyek ezeket a csomópontokat egy összefüggő, irányított körmentes gráfba, fogalmi hálózatba szervezik. A főnévi fogalmak között a legfontosabb reláció a hipernímia (ill. inverze: hiponímia), mely hierarchikus alá-/fölérendeltséget, specifikus/generikus, faj/nem, IS-A öröklődési viszonyt fejez ki (pl. {toll}-{írószer}, {bokor}-{növény}). Speciális altípusa a példánya-hipernímia reláció, mely tulajdonnevekhez kapcsolódó, indivudumoknak megfelelő és általánosabb, osztályoknak megfelelő fogalmak között állhat fent (pl. {Magyarország}-{európai ország}). A hipernímiához hasonló hierarchikus reláció a meronímia (inverzének neve: holonímia), mely rész-egész viszonyt fejez ki. Három fajtája van: egyén-csoport (pl. {fa}-{erdő}), alkotóanyag-tárgy (pl. {cellulóz}-{papír}) és alkatrész-egész (pl. {kerék}-{bicikli}) viszonyt kifejező.

A domain reláció egy tetszőleges fogalom (domain term) és egy témát, fogalmi osztályt (domain) reprezentáló fogalom között áll fenn. Három fajtája van: kategória (szemantikai mező, téma), pl. {teniszütő}-{tenisz}, régió (nyelvhasználók földrajzi helye szerint), pl. {ballup, balls-up}-{ United Kingdom, Great Britain} és használat (nyelvréteg szerinti besorolás), pl. {parázik}-{szleng, argó}. Főnévi fogalmak és más szófajú synsetek között is vannak relációk: tulajdonság mn. és neki megfelelő attribútum fn. között (pl. {piros}-{szín}), morfológiailag rokon (képzett) alakok között (pl. {fekvés}-{fekszik}-{fekvő}).

Főnevek, melléknevek és igék között is értelmezett az antonímia-reláció, mely szembenállást fejez ki valamilyen észszerű denotációs tartományban (pl. {nő}-{férfi}, {megszületik}-{meghal}, {hideg}-{meleg} stb.) Igéknél a hipernímia-

hiponíma relációpárhoz hasonló hierarchikus viszonyt fejez ki a hipernímia-troponímia (pl. {fut, szalad}-{mozog}). Speciális, igei synsetek közötti reláció az előfeltételezést kifejező kapcsolat, pl. {horkol}-{alszik}, illetve az okozás, pl. {meggyújt}-{elég}. Domain, illetve más szófajokhoz kapcsolódó derivációs relációk ezeknél a szófajoknál is vannak. A melléknéveknél a legfontosabb strukturáló reláció az antonímia. A határozószavaknak megfelelő synsetek csak más szófajokhoz kapcsolódnak derivációs morfológiai relációkkal.

A Princeton WordNet jelenleg legfrissebb változata (2.1) mintegy 155.000 angol szót szervez 81.400 különböző főnévi, 13.700 igei, 19.900 melléknévi és 3.700 határozószói synsetbe.

2.2 EuroWordNet

Az 1996-1999 között, az Európai Közösség finanszírozásában megvalósított EuroWordNet (EWN) projekt fő eredménye a WordNet architektúra többnyelvű környezetbe való átültetése volt. A EWN moduláris környezetet biztosít, ahol egy közösen elfogadott közvetítő fogalmi réteghez (Inter-Lingual Idex, ILI) kapcsolódnak a különböző nyelvek (holland, olasz, spanyol és angol, majd német, francia, cseh és észt) wordnetjeinek synsetjei.

Az EWN ILI nagyrészt az angol nyelvű Princeton WordNet 1.5-ös verziójának synsetjeiből állt, a közöttük lévő szemantikai relációk nélkül. Az ún. ekvivalencia-relációk biztosítják az átjárást az ILI-fogalmakon (ún. ILI-rekordokon) keresztül a különböző nyelvek synsetjei között. Ugyanahhoz az ILI-rekordhoz kapcsolt nyelvspecifikus synsetek ekvivalens jelentésűek a nyelvek között. A rugalmas kapcsolat megőrzése érdekében az ekvivalencia-relációnak a pontos azonosság kifejezésén túl több fajtája is van: pl. az adott ILI-fogalomnak egy adott nyelvben csak specifikusabb (vagy általánosabb) megjelenése van stb. Összesen 15-féle ilyen, nyelvek közötti ún. komplex ekvivalencia-relációt definiáltak.

Annak érdekében, hogy a különböző wordnetek fogalmi készlete egységes legyen (általánosságban ugyanazokkal a domain-ekkel vagy fogalmi területekkel foglalkozzanak), a wordneteket egy közösen meghatározott kiinduló fogalmi készlet, a Common Base Concepts (CBC) elemeiből kiindulva építették fel, felülről-lefelé haladva. A CBC fogalmakat a 8 nyelv wordnetjeinek fejlesztői közösen választották ki a PWN synsetjei közül, minden nyelvre lefordították őket, külön-külön kiegészítették egyéb, az adott nyelvben fontosnak ítélt kiinduló fogalommal (Local Base Concepts), és a lokális wordneteket ezekből kiindulva fejlesztették tovább, ahol lehetett, a saját synsteket az ekvivalencia-relációkkal az ILI-fogalmakhoz kapcsolva. A különböző wordnetek tehát egy közös vázra, a Common Base Concept-ekre épülnek, az erre épülő wordnet-struktúrák pedig nyelvenként eltérhetnek.

Noha a teljes ILI strukturálatlan (az angol PWN 1.5 synsetekhez nem vették át a PWN relációit), a részhalmazát képező, CBC 1310 fogalmát egy új, nyelvfüggetlen ontológia, a EuroWordNet Top Ontology (TO) rendszerezi. A TO 63 nyelvfüggetlen fogalom (Top Concept, TC) hierarchiája, melyek fontos szemantikai distinkciókat tükröznek, és meghatározó szemantika-elméletek alapján határozták meg őket. A TO a CBC-ket valójában nem osztályokba szervezi, inkább feature-ök kombinációjaként jellemzi őket, egy CBC-hez több TC is tartozhat. A TC-k a CBC-k ILI-rekordjain keresztül öröklődnek a kapcsolódó nyelvspecifikus jelentésekre.

A EuroWordNet-ben az egyes nyelvek WN-jeit alapvetően két fő különböző módszer egyikével alakították ki:

a) Összevonásos módszer (Merge Model): a lokális alapfogalmakat (BC-k) valamilyen saját erőforrásból kiindulva választották ki, belőlük a synseteket és az azok között lévő relációkat önállóan fejlesztették ki, majd az ekvivalencia-relációkkal leképezték őket az ILI (PWN1.5) synsetekre.

b) Kiterjesztéses módszer (Expand Model): a lokális alapfogalmakat a PWN1.5-ből választották és a PWN1.5 synseteket (kétnyelvű szótárak segítségével) lefordították ekvivalens saját synsetekre. Ebben a megközelítésben a belső relációkat a PWN-ből örökölték, és a továbbiakban, amennyire lehetett, egynyelvű erőforrások segítségével ellenőrizték őket.

Az Összevonásos módszer alkalmazásával a PWN1.5-től független, a nyelvspecifikus tulajdonságokat megőrző wordnetet lehet létrehozni. A Kiterjesztéses módszer a PWN1.5 által erősen determinált wordnetet eredményez. A követendő módszert elsősorban a rendelkezésre álló erőforrások határozták meg.

2.3 BalkaNet

A 2001-2004 között megvalósított, EK finanszírozású BalkaNet (BN) projekt célja a EuroWordNet kiterjesztése volt 5 újabb, délkelet-európai nyelvvel (bolgár, görög, román, szerb és török) [13].

A BN végső változatában az Inter-Lingual Index szerepét a 2.0-ás verziójú Princeton Wordnet synsetjei töltötték be. A BN ILI (BILI) fogalmai fölött egy újabb nyelvfüggetlen ontológiát definiáltak a SUMO felsőszintű ontológia [10] és a PWN fogalmai közötti megfeleltetések átvételével.

A BN-ben a közös kiinduló fogalmi készlet (BalkaNet Concept Set, BCS) 8.516 PWN synsetetből áll: a EWN CBC synsetjein kívül további, az új nyelvek által hozzáadott fogalmakat is tartalmaz.

A projektben az összes erőforrást közös platformra, XML formátumba konvertálták, melyek így a szabadon felhasználható, egyszerre több nyelvi erőforrás böngészését-szerkesztését lehetővé tevő, a BN projekt céljára kifejlesztett VisDic programmal [4] kezelhetők A minőség-ellenőrzőséhez különböző validációs módszereket rendszeresítettek, melyek biztosítják a wordnetek szintaktikai és strukturális helyességét és konzisztenciáját, valamint a nyelvek közötti kapcsolatok érvényességét. [12].

3 A magyar EuroWordNet fejlesztési koncepciója

A bevezetőben említett, 3 éves kutatási-fejlesztési projekt egyik fő terméke a magyar nyelvű EuroWordNet adatbázis lesz. A három intézményből álló konzorcium az ontológia fejlesztéséhez az alábbi stratégiai megfontolásokat fogadta el:

1) A BalkaNet projekt szabadon hozzáférhető erőforrásainak használata.

A magyar wordnet építésének kiindulópontjául nem a EuroWordNet Common Base Concepts, hanem a BalkaNet Concept Set (BCS) synsetjeit választottuk. Utóbbi mellett a következő érvek szóltak:

- A BN BCS a EWN CBC fogalmain felül tartalmaz további 5 európai nyelvben alapvető fontosságúnak tartott fogalmakat (összesen tehát 13 nyelv többnyelvű WN-jének felépítésében hasznosnak tartott információkat, szemben a EWN CBC 8 nyelvével).
- A BCS a Princeton WordNet újabb, 2.0-s verziójára alapul, a EWN CBC a PWN1.5-ösre.
- A BCS 8 516 synsetet tartalmaz, a CBC 1 310-et. A nagyobb mennyiségű synset teljesebb kiindulási alapot ad a létező EWN/BN wordnetek szókincsének magasabb átfedéséhez.
- A BCS fölött rendelkezésre áll két struktúra is (PWN, illetve SUMO hierarchiák), melyek a BN projekt tapasztalatai alapján, rendkívül hasznosak lehetnek az általunk követett kiterjesztéses modell követésekor (ld. lejjebb).

A BCS adaptációjával összhangban az Inter-Lingual Index (ILI) számunkra is a PWN2.0 anyaga lett. Az erőforrásainkat a BN projekt által kialakított XML formátumba konvertáltuk, szerkesztésre és megjelenítésre a VisDic programot választottuk.

2) Ahol lehet, a kiterjesztéses modell, máshol a kiterjesztéses-összevonásos módszerek keverékének alkalmazása.

   Korábbi kísérleteinkben bebizonyosodott, hogy az angol és a magyar főnévi fogalmak rendszere közötti hasonlóság kellő mértékben fennáll ahhoz, hogy a kiterjesztéses módszert követni lehessen [8]. Ennek során a kiindulásul választott ILI (PWN 2.0) synsetjeit automatikus és kézi módszerekkel lefordítjuk és átveszszük a PWN-ben közöttük definiált szemantikai relációkat. Annak érdekében, hogy végeredmény a magyar nyelv fogalmi sajátosságait tükrözze, a lefordított synseteket, illetve az angolból örökölt relációkat alapos kézi munkával, egyenként ellenőrizzük, és ahol szükséges, módosítjuk.

   Amennyiben a nyelvi különbségek miatt ez a módszer tarthatatlan, bizonyos területeken, ill. szófajoknál az összevonásos módszert is alkalmazzuk (magyar synsetek önálló kifejlesztése, beillesztése a magyar ontológiába, majd ILI-rekordhoz kapcsolása).

   Az alapvetően a kiterjesztéses módszert követő megközelítés mellett a megfelelően strukturált erőforrások hiánya, az alacsonyabb fejlesztési költségek, illetve a korábban kifejlesztett és sikerrel alkalmazott, rendelkezésre álló automatikus módszerek szóltak (ld. következő pont).

3) Fél-automatikus módszerek alkalmazása.

   Egy korábbi projekt során olyan módszereket fejlesztettünk ki, melyek segítségével automatikusan lehetett egy magyar-angol alap szótár magyar főnévi címszavait angol (PWN 1.6) synsetekhez hozzárendelni [8]. A 9 különböző heurisztikát alkalmazó algoritmus a kétnyelvű szótárban található strukturális és morfoszemantikai információkon kívül a Magyar Értelmező Kéziszótár [6] egy elektronikus változatából kinyert főnévi definíciókban azonosított szemantikai relációkat használta fel. A különböző heurisztikák eredményeinek legelőnyösebb kombinációja egy kézzel egyértelműsített etalon halmazhoz képest átlagosan kb.

75%-os pontosságot eredményezett (a magyar főnevek és a PWN synsetek között generált kapcsolatokat tekintve).

A legelőnyösebbnek bizonyult automatikus módszereket, újabb erőforrásokkal támogatva (MorphoLogic Tezaurusz szinonimaszótár) alkalmazzuk arra, hogy a BCS, illetve más PWN 2.0 synseteket magyar szinonima-ajánlásokkal lássunk el, melyeket ezután kézi munkával ellenőrzünk.

4) A konzorcium számára rendelkezésre álló szemantikai erőforrások integrációja a készülő ontológiába.

Az automatikusan szinonima-ajánlatokkal ellátott magyar synsetek kézi ellenőrzési fázisa során egyfelől a Magyar Értelmező Kéziszótár (ÉKSz) bejegyzéseit megpróbáljuk megfeleltetni a készülő magyar synsetekkel, másfelől a Nyelvtudományi Intézetben fejlesztett magyar igei vonzatkeret-leíró adatbázis tételeit hozzárendeljük a megfelelő igejelentésekhez.

Az ÉKSz-jelentésekkel létesített leképezés előnye, hogy egyfelől a magyar synsetekhez alkalmas magyar szöveges definíció rendelhető, másfelől az ÉKSz definíciókban feltárt szemantikai relációk alapján lehetőség nyílik az ontológia további kiterjesztésére (ld. 5. fejezet).

5) A BN projekt által kidolgozott minőségbiztosítási metodológia ([12]) adaptációja és rendszeres alkalmazása az eredményeink validációjához. Az ellenőrzésnek a következő kérdéseket kell érintenie:

Synsetek formai ellenőrzése:

- minden synsetben minden literálnak (szinonimának) van jelentés-azonosítója,
- egy synsetben nem lehet két azonos literál (jelentés-azonosítótól függetlenül),
- egy literál ugyanazzal a jelentés-azonosítóval nem fordulhat elő egynél több synsetben (ugyanabban a szófajban),
- egy szó különböző jelentéseihez tartozó jelentés-azonosítók számozása folytonos ,
- literálok automatikus helyesírás-ellenőrzése,
- synset ID ellenőrzés: az azonosítóknak egyedinek kell lennie, minden synset csak a 4 megengedett szófajkód egyikével lehet megjelölve (n, v, a, b)
- synset literáljainak sorrendezése korpuszban megfigyelt gyakoriság alapján.

Belső (lokális) relációk formai ellenőrzése:

- nincsen ugyanaz a reláció ugyanazon 2 synset között többször felvéve,
- a reláció (neve) a standard BN szemantikai relációk (nevének) egyike,
- nincsenek irányított körök,
- a relációkban álló synsetek megfelelő szófajúak,

- egy synsethez kapcsolódó relációknak kompatibiliseknek kell lennie egymással (pl. egy synset nem lehet egyszerre hipernímája és hiponímája ugyanannak a synsetnek),
- nem lehetnek másokkal kapcsolatban nem álló csomópontok,
- minden synsetnek kell, hogy legyen hipernímája, hacsak nem legfelső szintű (gyökér) fogalomnak felel meg,
- nem létező synsetekkel alkotott relációk javítása/törlése.

Synsetekhez tartozó definíciók és használati példák formai ellenőrzése:

- a definíció ne legyen üres,
- a definíció a saját nyelven legyen megfogalmazva,
- definíció szövegének automatikus helyesírás-ellenőrzése,
- a definíció lehetőleg ne tartalmazza a definiált synset literáljait,
- a használati példa a literált a megfelelő szófajjal tartalmazza.

(Az ILI és a magyar synsetek közötti) ekvivalencia-relációk ellenőrzése:

- minden magyar synsetnek legyen(ek) ekvivalense(i) az ILI-ben.

Az ellenőrzések egy része automatikusan elvégezhető, az így felismert hibákat ezután kézzel kell kijavítani.

A magyar WordNet ontológia fejlesztése a következő lépésekben történik: először a kiinduló „mag" részt készítjük el a BN BCS 8.516 synsetjének lefordításával. A fordításhoz a gépi heurisztikákkal minél több angol BCS synsethez automatikus javaslatokat próbálunk tenni. Ezek eredményét emberi munkával, egyenként ellenőrizzük. Az ellenőrzés közben történik az ÉKSz és az Igei Vonzatkeret-adatbázis tételeivel való megfeleltetés, illetve megfelelő magyar synset definíciók és példamondatok megírása. Az összes BCS synset magyar reprezentánsának elkészülte után a PWN-től örökölt szemantikai relációk ellenőrzése-szerkesztése történik egyenként.

A magyarra fordított BCS halmazt ezután kiegészítjük olyan további alapvető fogalmakkal, amelyek nem szerepelnek benne, viszont a kiinduló halmazban fontos szerepük lehet. Ehhez korpuszban (Magyar Nemzeti Szövegtár, illetve ÉKSz definíciós korpusz) megfigyelt gyakorisági értékek alapján keressük meg a potenciális fogalmakat, melyekből synseteket képezünk, azokat beillesztjük a már meglévő ontológiába, valamint megkeressük az ILI ekvivalenseiket. Mindezek után egy teljes validációs ciklussal véglegesítjük a magyar kiinduló fogalmi halmazt (hibák és hibalehetőségek automatikus listázása, kézi javítása).

A második tervezett nagy lépés a mag rész további, felülről-lefelé irányuló kiterjesztése lesz nagy mennyiségű további fogalommal (a végleges magyar wordnet ontológia mintegy 30.000 synsetet fog tartalmazni). A munka részben a BCS létrehozásához hasonlóan fog történni. Automatikus módszerekkel lefordítjuk a Princeton WordNet BCS után fennmaradó synsetjeit, majd az eredményeket a fentiekhez hasonlóan kézzel ellenőrizzük, szerkesztjük és kiegészítjük.

Az angol WN fordítás mellett a már elkészült felső szintek, az ezek és az ÉKSz jelentései közötti megfeleltetések, valamint az ÉKSz definíciókban végzett szemantikai elemezések eredményei alapján lehetőség lesz további magyar hiponímák (troponímák) automatikus hozzáadására [2]. További, fél-automatikus bővítési lehetőséget kínál a nagy mennyiségben rendelkezésre álló tematikus szólisták feldolgozá-

sa (pl. földrajzi nevek, cégnevek, tulajdonnevek stb.) A magyar nyelvben megtalálható derivációs morfológiai relációk a rendelkezésünkre álló morfológiai elemző és generáló eszközök segítségével automatikusan kiegészíthetők. Mindezen munkák eredményét szintén kézzel kell majd ellenőrizni.

Az utolsó lépés egy speciális, üzleti szakkifejezéseket tartalmazó szakontológia kifejlesztése és az általános ontológiához kapcsolása lesz. Az üzleti szakontológia a projekt többi célkitűzéseit (információ-kinyerés többmondatos rövid szöveges üzleti hírekből) fogja támogatni [1]. Elkészítésében támaszkodni fogunk a szabadon hozzáférhető Teknowledge Financial Ontology szakontológiára[15], valamint a MorphoLogic rendelkezésére álló angol-magyar üzleti szótárakra.

4 Jelenlegi eredmények

A munka kezdetén a korábbi hasonló projektek ([2], [3]) eredményei alapján kifejlesztett, automatikus synset-fordító heurisztikákat ([8]) alkalmaztuk a 8.516 db BCS synset magyarra fordításához. Az alábbiakban röviden bemutatjuk a kiválasztott heurisztikák algoritmusait és a támogató nyelvi erőforrásokat:

a) Egyjelentésű angol szavak: ha egy magyar szó valamelyik angol fordítása egyértelmű a WN-ben, vagyis csupán egyetlen synsetbe tartozik, akkor létrehozunk egy kapcsolatot a magyar szó és a synset között.

b) Többjelentésű angol szavak egyértelmű fordítással: ha egy angol szónak csak egyetlen, egyértelmű magyar fordítása van (a magyar szónak csak ez az egyetlen angol fordítása), és az angol szó a WN-ben több synsethez is tartozik, a magyar fordítást hozzárendeljük ezekhez.

c) Variánsok: ha egy WN synset kettő vagy több olyan angol szót tartalmaz, melyeknek csupán egyetlen magyar fordításuk van, és az ugyanaz a magyar szó, akkor a magyar szót hozzárendeljük a közös synsethez.

d) Szinonimák: a magyar szó angol fordításaihoz tartozó synsetek közül azt választjuk ki, amely a legtöbbet tartalmazza a szó szinonimáinak angol fordításai közül (de legalább kettőt). Magyar szinonimák előállításához felhasználtuk egyfelől az ÉKSz definíciókban géppel azonosított szinonimákat ([8]), másfelől a ML Tezaurusz szinonimáit.

e) Latin nevek: ha egy magyar szóhoz rendelkezésre áll latin megfelelő (állat- és növényfajok, rendszertani kategóriák stb.), akkor azt az angol synsetet választjuk, ami az angol fordításon kívül a latin nevet is tartalmazza. Latin ekvivalenseket az ÉKSz-ből, illetve a kétnyelvű szótárakból azonosítottunk magyar címszavakhoz.

f) Minimális fogalmi távolság: amennyiben van egy magyar szó és egy hozzá tartozó magyar hiperníma szó, akkor képezzük ezek fordításainak lehetséges

[15] http://ontology.teknowledge.com/

synsetjeit, majd belőlük megkeressük azt a párt, ami a WN fogalmi hálózatában a legközelebb helyezkedik el egymáshoz. A magyar címszót a minimális távolságú pár megfelelő tagjához rendeljük.

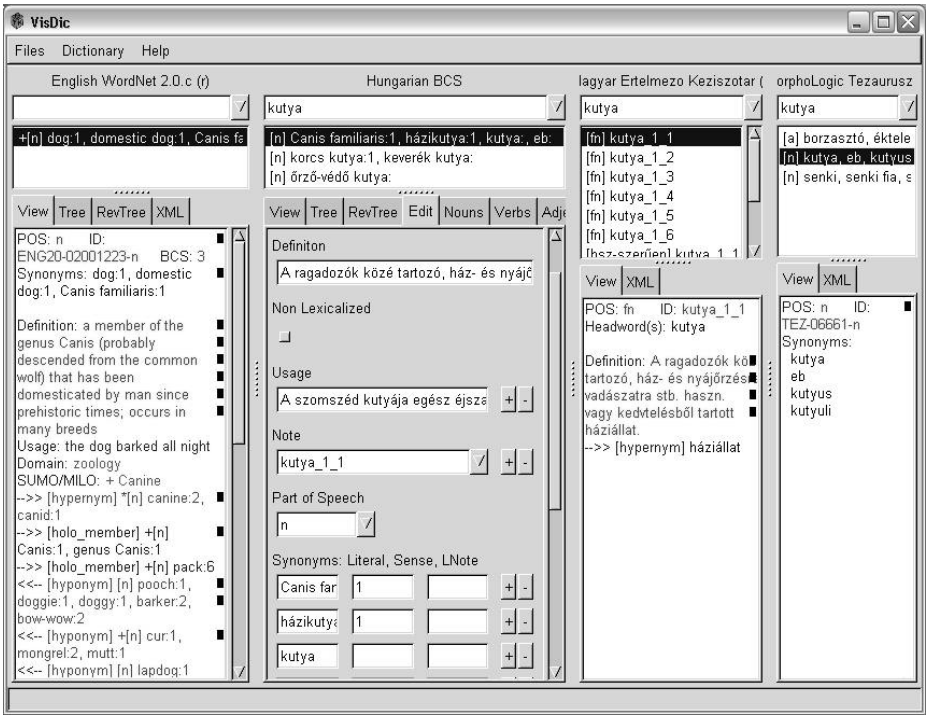
Magyar címszavakhoz hiperníma szavak kétféle forrásból álltak rendelkezésre: egyrészt az ÉKSz definíciók szemantikai elemzésével, másrészt a kétnyelvű szótárak összetett főneveinek morfológiai elemzésével. Ez utóbbi kétféleképpen történhetett: endocentrikus főnévi-főnévi összetételek (egybe írt összetett szavak) utótagjának azonosításával, illetve többszavas lexikalizált főnévi kifejezések esetén a fej azonosításával.

Mivel a BCS synsetek csak kb. 87%-ában volt legalább egy szinonimának magyar fordítása, az automatikus fordítás által elérhető elméleti maximum lefedettség 87% volt. A heurisztikák kombinált eredményei a teljes BCS anyagának kb. felét voltak képesek lefedni (1. Táblázat).

1. Táblázat: a BCS automatikus fordításának eredményei

	BCS	**Automatikusan. lefordítva**	**(%)**
Főnévi synsetek	5 896	3 149	(53,41%)
Igei synsetek	2 318	1 139	(49,14%)
Mn-i synsetek	302	77	(25,50%)
összesen:	*8 516*	*4 365*	*(51,26%)*

Az automatikus fordítást a kézi ellenőrzési-szerkesztési fázis követte: a lefordított synsetek szinonimáit ellenőrizni és/vagy kiegészíteni, a nem lefordított synseteket pedig le kellett fordítani. A munkához segítséget a VisDic-be integrált Értelmező Kéziszótár és a MorphoLogic Tezaurusz nyújtott. A szinonimák ellenőrzése mellett az ÉKSz, illetve a NYTI vonzatkeret-adatbázissal való összekapcsolás is ekkor történt. A munka egy 2 oldalas irányelv betartásával, a VisDic editor felhasználásával folyik (1. Ábra). A cikk megírásának időpontjáig a BCS mintegy háromnegyedét dolgoztuk fel.



1. Ábra: 4 szótár szimultán használata a VisDic program segítségével. Balról jobbra: Princeton WordNet 2.0, synset áttekintő nézet; magyar WordNet, synset szerkesztés nézet; Értelmező Kéziszótár és ML Tezaurusz, szócikk megjelenítés. Az automatikus szinkronizálásnak köszönhetően az angol és a magyar wordnetek ekvivalens fogalmakat jelenítenek meg. A másik két szótárban kézzel kerestük meg a megfelelő szócikkeket.

A manuális munka során számos, előre nem látott nehézséggel és problémával találkoztunk.

Először is, az igei rész fordításának elején nyilvánvalóvá vált, hogy a kiterjesztéses módszer nem tartható teljes mértékben. Egyrészt az eredeti PWN-ben az igék rendszerében megfigyelt hibák és inkonzisztenciák, másrészt a két nyelv igei rendszere közötti morfológiai és szemantikai különbségek felismerése miatt úgy döntöttünk, hogy az igei részt nem teljes mértékben az angol taxonómiára támaszkodva, hanem részben önállóan elindulva készítjük el (részletesen ld. [7]).

A deverbális (igékből képzett) főneveket tartalmazó synsetek feldolgozásakor hasonló okokból szintén problémákat észleltünk. Pl. ezen a fogalmi területen kiugróan magas azon BCS synseteknek a száma, melyek nem, vagy csak igen nehezen, csak nem lexikalizált, ad hoc frázisokkal fordíthatók magyarra. Mindezek miatt célszerűnek látszik a közeljövőben a deverbális főnévi és az igei rész taxonómiáját összehangolni (a deverbális synsetek közötti relációkat a morfológiailag megfelelő igei synsetek között kézzel megállapított relációkhoz igazítani). A derivációs relációkat automatikusan kell majd hozzáadni.

A biológiai taxonómikus fogalmak (állatfajok, -nemek, -rendek stb.) fordítása során előforduló problémák közül kiemelhető, hogy sokszor a WN synsetek pontatlanok voltak, illetve a segítségként felhasznált magyar taxonomikus forrásoktól eltérő

rendszerezéseket mutattak be. Ezen synsetek lektorálásához szeretnénk egy biológus szakfordítót felkérni.

Sokszor előfordult, hogy egy PWN synset lefordításakor nehezen lehetett pontosan azonosítani a fogalmat, mivel a synset hipernímája (hiponímája) nagyon apró, nem anyanyelvi beszélő számára csupán nagyon nehezen észrevehető különbséget tartalmazott a kérdéses synsethez képest. Az ilyen eseteket megjelöltük a relációk kézi ellenőrzési munkaszakaszához. A megoldást a komplex ekvivalencia-relációk alkalmazása jelentheti, melyekkel modellezhetők azok az esetek, amikor a magyar hierarchia nem követi pontosan az angolt.

Részletes, összefoglaló adatokat és felvetéseket a teljes BCS anyag fordításának elkészülése és elemzése után tudunk megadni, amit egy későbbi publikációban tervezünk bemutatni.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas Gy.: Construction of the Hungarian EuroWordNet Ontology And Its Application To Information Extraction. To appear in: Proceedings of the 3rd International WordNet Conference, Jeju Island, Korea (2006)
2. Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997)
3. Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)
4. Horak, A., P. Smrz: New Features of Wordnet Editor VisDic. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
6. Juhász, J., I., Szőke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest (1972)
7. Kuti, J., Vajda P., Varasdi K.: Javaslat a magyar igei WordNet kialakítására. Elbírálás alatt a III. Magyar Számítógépe Konferencián, Szeged (2005)
8. Miháltz, M.: Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, January 20--23 (2004)
9. Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.
10. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19 (2001)
11. Prószéky, G.: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany, pp 149–158 (1996)
12. Smrz, P.: Quality Control and Checking for Wordnets Development: A Case Study of BalkaNet. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
13. Tufiş, D., D. Cristea, S. Stamou: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
14. Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document (Deliverable D032D033/2D014) (1999)

Javaslat a magyar igei WordNet kialakítására

Kuti Judit[1], Vajda Péter[1], Varasdi Károly[1]

[1]MTA Nyelvtudományi Intézet – Korpusznyelvészeti Osztály
1068 Budapest, Benczúr u. 33.
{kutij, vajda, varasdi}@nytud.hu

Kivonat: Előadásunkban a magyar igei WordNet elkészítésének javasolt módszerét mutatjuk be. A Princeton WordNet automatikus fordításának kézi ellenőrzése során elvetettük a WordNet hierarchikus struktúrájának kritika nélküli átvételét, mivel az a jelentések elkülönítésében és a hierarchia kialakításában következetlenségeket tartalmaz, továbbá nem tükrözi a magyar igék lexikális viszonyait. Kiinduló jelentéshalmazunk meghatározásához ugyanakkor megtartjuk fogalmainak egy részét, és ezeket megkíséreljük a magyarban gyakori jelentésű igékkel kibővíteni. Az igék közötti relációkat egy olyan általunk kialakított módszer alapján vesszük fel, amely nyelvészetileg megalapozott szemantikai teszteken alapul, valamint megengedi, hogy egy szó több másikkal is alá-fölérendelt viszonyban legyen. A jelentések, relációk és az igei hierarchia kialakításához a Magyar Értelmező Kéziszótárt és több európai nyelvre elkészült wordnetet is felhasználunk.

1 Bevezetés

A magyar WordNet adatbázis a 2005 tavaszán indult *Magyar Ontológia Építése* projekt keretén belül készül, mely a Szegedi Tudományegyetem, a MorphoLogic Kft. és a Nyelvtudományi Intézet közös projektuma. A Nyelvtudományi Intézet a magyar EuroWordnet igei részének elkészítését vállalta magára. A EWN és a BalkaNet projektek (ld. [8], [2], [9], [10]) lezárulásával két módszer vált nemzetközileg elfogadottá WordNetek kialakítására. Az egyik az ún. *kiterjesztéses módszer (expand model)*, melynek során a készülő WordNet mag-synset halmaza a Princeton WordNet 2.0-ból (a továbbiakban PWN) kerül ki. A kiválasztott PWN synseteknek célnyelvi synset-megfeleléseket keresnek, melyek megöröklik a PWN hierarchikus relációit. Ezt a nagyrészt automatizálható munkafázist manuális ellenőrzés követi, saját nyelvi erőforrások használatával. A másik módszer az ún. *összevonásos módszer (merge model)*, melynek során saját erőforrásokból kiindulva határozzák meg a kiindulási jelentéshalmazt, és a közöttük fennálló relációkat is önállóan alakítják ki. A mostanra szabvánnyá vált PWN-tel való kompatibilitást az ezután következő munkafázis biztosítja, mely megfelelteti a készülő WordNet synsetjeit a PWN aktuális verziójának synsetjeivel. A HuWN kialakítását eredetileg a melléknévi és főnévi részhez hasonlóan, túlnyomó részben a kiterjesztéses módszerrel terveztük. Ez a BalkaNet projektben használt nyelvek közötti közvetítő réteg (Balkanet Concept Set (BCS)),

mely 8516 PWN synsetnek megfeleltetett jelentést tartalmaz, ebből 2318 igei synset) automatikus módszerekkel történő magyar nyelvre képezését, majd manuális ellenőrzését és kiterjesztését jelentette volna. A manuális ellenőrzés kezdeti fázisában felmerülő problémák ill. kérdések, valamint az az igény, hogy saját nyelvi erőforrásaink teljesebb kihasználása érvényesüljön, olyan módszertani változtatásokat, kiegészítéseket tett szükségessé, melyek jobban figyelembe veszik mind az igei szófaj, mind a magyar nyelv lexikalizációs sajátságait.

A következő pontban megvizsgáljuk a Princeton WordNet azon tulajdonságait, melyek fogalmainak és hierarchiájának teljes átvétele ellen szólnak, és megindokoljuk a kiterjesztéses modell elvetését. A 3-ik pontban megemlítjük az igék sajátosságaiból adódó szempontokat, melyeket a HuWN kialakításakor figyelembe veszünk. Bemutatjuk a javasolt munkamódszert, kitérve a magyar WordNetbe felveendő jelentések kiválasztására és a köztük lévő relációk kialakítására. Az utolsó pontban a további munkafázisokról ejtünk néhány szót.

2 A Princeton WordNet átvételének korlátai

A felmerülő problémák részben a kiindulási alapot képező PWN hiányosságaiból (ld. alább, ill. részletesebben [3]), részben pedig a tervezett kiterjesztéses módszer korlátaiból adódtak.

2.1 A jelentések elkülönítése

A kiterjesztéses módszer alkalmazásával a PWN belső hierarchiáját a magyar hierarchia is megörökölné. A PWN igei részének azonban számos olyan jellemzője van, amelyeknek az átvételét az igei HuWN készítésekor el szeretnénk kerülni, ill. amelyet módosítani szeretnénk. Alapvető kérdésként merül fel, hogy a PWN felépítése során mely elv határozta meg a jelentések megkülönböztetését. Ha abból indulunk ki, hogy a PWN *nyelvi ontológia* (ld. [10]), azaz egy bizonyos nyelv *lexikalizálódott* jelentésmegkülönböztetéseit hivatott tükrözni, nem egyértelmű, mi alapján kerültek egy ill. több synsetbe a következő jelentések:

> *{shed:4, molt:1, exuviate:1, slough:1, moult:1}*[16] (def.: Némely állat bizonyos időszakokban leveti kültakarója bizonyos rétegét, pl. szőrét, tollát, agancsát.)
>
> *{feed:2, give:24}* (def.: enni ad valakinek) ill. *{give:19}* (def.: gyógyszert ad valakinek)
>
> *{clean:1, make clean:1}* (def.: valamit a rajta vagy benne levő szennytől egészen megtisztít) ill. *{wash:3, launder:1}* (def.: Vízzel tisztít, általában ruhaneműt.)

Míg a *{shed:4}* synset szinonimái közül a *shed* minden típusú levetett állati külsőre vonatkozhat, a *slough* viszont nem vonatkozhat madártollra, csak kétéltű vagy hüllő bőrére ill. agancsra, addig a *clean:1* és a *{wash:3, launder:1}* között (bár a definíciók

[16] Kapcsos zárójelbe kerül egy synset, amennyiben több elemével utalunk rá.

ezt nem tükrözik egyértelműen) pont a tárgyi vonzat milyensége adja meg a különbséget: a *{clean:1}* bármire vonatkozhat, a *launder:1* azonban csak ruhaneműre. Ennek ellenére ugyanabba a synsetbe tartozik a *shed:4* és *slough:1*, ellenben két külön synsetbe a *clean:1* és a *launder:1*. A *give:24* és *give:19* között fellelhető egyetlen különbség szintén csupán abban áll, hogy a tárgyi vonzat ennivaló vagy gyógyszer-e, de annak ellenére, hogy a köztük levő különbség nem lexikalizálódott, két külön synsetbe kerültek. Ezek a példák azt mutatják, hogy a PWN nem konzekvens a lexikalizálódott vonzatok kezelésének tekintetében. Amennyiben nyelvi ontológiának tekintjük a PWN-t, szintén megkérdőjelezhetővé válik a következő synseteknek, mint lexikalizálódott kifejezéseknek a felvétele a hierarchiába: *create from raw material:1*, *create from raw stuff:1*, *change magnitude:1* ill. *change integrity:1*. Amennyiben azonban a PWN nem „csupán" nyelvi ontológiának tekintendő, hanem legalább részben következtetés alapú ontológiának (ld. [10]), érthetővé válik ezeknek a csomópontoknak a bevezetése, de szembeötlik mesterséges, azaz nem lexikalizált csomópontokként való megjelölésük hiánya.

Szintén inkonzekvensnek hat némely metaforizációra visszavezethető jelentésmegkülönböztetés.

(i)

{take away:1, carry away:1} def.: valamely helyről, környezetből, mentális vagy érzelmi állapotból kimozdít – usage: The car carried us off to the meeting; I got carried away when I saw the dead man and I started to cry.

{sweep:1, brush:4} def.: át- vagy keresztülsöpör – usage.: Her long skirt brushed the floor.; A gasp swept cross the audience.

(ii)

take off:3 def.: eltávolodik a földtől – usage.: The plane took off two hours late.

take off:7 def.: elindul, mozgásba jön *átvitt értelemben* – usage.: the *project* took a long time to get off the ground.

Mivel az (i) alatt szereplő szinonimák használatára vonatkozóan metaforikus és nem metaforikus példák is szerepelnek, indokolatlannak tűnik a (ii) alatti *take off* két külön jelentésének megkülönböztetése figuratív használat ürügyén.

2.2 A fogalmi hierarchia átvételének hátrányai

Az előző pontban említett módszertani következetlenségeken kívül számos olyan eset van, ahol a hierarchia kialakítása mögött álló elv megvalósítása hibásnak tűnik. Ilyen például az ágens - páciens thematikus szerepek alternációjának inkonzekvens kezelése. Például több ágens alanyú ige (*correct, falsify, undo, modify*) a páciens alanyú, és nem az ágens alanyú *change* hiponímájaként (azaz a *változtat* helyett a *változik* csomópont alatt) szerepel.

Ezektől a hibáktól eltekintve is bizonyos negatívumokkal járna az angol fogalmi háló teljes átvétele. Egyrészt elkerülhetetlen, hogy kevésbé fogja tükrözni a magyar lexikalizációs viszonyokat az elkészült HuWN, mintha saját erőforrásokra támaszkodva vennénk fel a relációkat. Másrészt köztudomású, hogy a PWN hierarchia

csomópontjai olyan sűrű fogalmi hálót alkotnak, hogy magyar nyelvre történő leképezésük nemcsak sok esetben manuálisan is erőltetettnek bizonyul, de az alábbi nehézségeket is magában hordozza.

Ha a tervezett módszert követve, először a PWN összes, a BalkaNet Base Concept Set-jét alkotó synsetjét megfeleltetjük magyar synseteknek, majd azokat feleltetjük meg ÉKSz-beli jelentéseknek, óhatatlanul lesznek olyan ÉKSz jelentések, melyek kimaradnak egy adott ige jelentései közül. Ezeket ugyan egy későbbi, a BalkaNet Concept Set-et bővítő munkafázis során bevehetjük a HuWN-be, de nem lesz rá lehetőségünk, hogy esetleg már synsetként felvett jelentésekkel összevonjuk őket. A tapasztalat pedig azt mutatja, hogy az ÉKSz rengeteg olyan kollokációt és idiomatikus kifejezést tartalmaz, amelyeket a HuWN céljainak megfelelően össze lehetne vonni kevesebb synsetbe is.

A PWN fogalmi sűrűségének magyarra való leképezése azért is lenne problematikus, mert a készülő HuWN egyik legfontosabb felhasználása egy információkinyerő rendszer hatásfokának javítása lesz. Amennyiben azonban a HuWN igei része átveszi a PWN fogalmi hálóját, túlságosan sűrű lesz ahhoz, hogy hatékony segítséget nyújtson nyelvfeldolgozási alkalmazásokban. A PWN a *go* lemmának például 28, a *cut* lemmának 21, a *see* lemmának pedig 19 igei jelentését különbözteti meg. Többnyelvű információkinyerés esetében exponenciálisan nő a keresés eredményekor a zaj, ha az eredeti keresett kifejezéshez hozzávesszük a PWN synsetjeiből való szinonimákat. Ha csökkenteni tudjuk a többértelműséget, pontosabb eredményeket fog adni egy-egy keresés (ld. [7]).

Végül meg kell említeni, hogy a PWN készítése során nem alkalmazták a többszörös öröklődés lehetőségét a hierarchiában. Minden fogalomnak fix helye van, amely meghatározza jelentését. Bár ez a módszertani döntés érthető, hiszen a PWN készítése a főnévi szókincs építésével kezdődött, az igei HuWN készítésénél úgy gondoljuk, hasznos lenne bizonyos esetekben a többszörös öröklődés bevezetése.

3 Javasolt módszertani változások

3.1 A kiinduló jelentéshalmaz meghatározása

Az igék, mint eseményszerűségeket leíró egységek, általában nyelvspecifikusabb módon tükröznek lexikalizációs mintákat, mint a főnevek, hiszen ugyanannak az eseménynek több szempontját is megnevezhetik. Pl. az angol *clear:24* (def.: megtisztítani a torkot valamitől, rekedtes hangadás kíséretében) synset a *remove:1* (def.: valami konkrétat eltávolít valahonnan, emeléssel, nyomással stb., vagy valami absztrakt dolgot eltávolít) hiponímája az angolban, míg az ennek a magyarban megfelelő (meg)köszörüli (a torkát) inkább valamilyen hangadást kifejező ige hiponímája lehetne. Ebből adódóan jobban szeretnénk függetleníteni a HuWN igei részét az angol nyelv által meghatározott fogalmi hálótól, mint a főnévi és melléknévi részt.[17] Ennek

[17] A főnévi és melléknévi HuWN kialakítása a következő módszertan szerint történik: a PWN fogalmi csomópontjait automatikus módszerekkel megfeleltetjük magyar synset-kezdeményeknek, és átvesszük a közöttük fennálló relációkat. Az automatikus fordítás munkafázisát alapos manuális ellenőrzés követi.

érdekében három fontos szempontot tartanánk szem előtt az igei HuWN kialakításakor:[18]

(i) Csak PWN fogalmi csomópontokat képezünk le magyarra, a köztük fennálló relációkat nem.

(ii) Korlátozzuk az angolból automatikus leképezéssel átvett igei jelentések számát a taxonómiailag fontos jelentésekre.

(iii) Az automatikus leképezéssel nyert kiinduló fogalmak körét kiegészítjük saját erőforrásokból származó jelentésekkel.

Taxonómiailag fontos jelentésűnek vesszük azokat a PWN-beli jelentéseket, melyek magas hierarchiai pozícióval rendelkeznek. Ezt a feltételt teljesítik a EWN igei base concept-jeiből nyert BCS synsetek (254 igei synset) és a BCS igei felső csomópontjai. Ezeket a synseteket egészítjük ki olyan igei jelentésekkel, melyekről feltételezhető, hogy gyakoriak, ill. hogy releváns szerepet játszanak lexikális ismeretek tárolásakor. Mivel jelentésgyakorisági adatok nem állnak a rendelkezésünkre, az előbbi szempont figyelembevétele érdekében igei lemma gyakorisági lista készült a Magyar Nemzeti Szövegtárból, míg az utóbbi szempont betartásának érdekében két igei lemma gyakorisági lista készült a Magyar Értelmező Kéziszótár elektronikus verziójából: egy igei genus proximum gyakorisági lista az igei definíciókból, és egy igei lemma gyakorisági lista az összes ÉKSZ definícióból.[19] Az MNSZ-ből és az ÉKSz összes igei definíciójából kapott lemma gyakorisági lista első 50 elemének metszetét kiegészítettük a a genus proximumok listájának azon elemeivel, melyek ebben a metszetben még nem szerepeltek. Az így kapott lista 28 elemet tartalmaz, melyeknek összesen 471 ÉKSz-beli jelentés felel meg. A Nyelvtudományi Intézet birtokában levő igei vonzatkeret adatbázis rekordjai közül az említett 28 lemmának 371 vonzatkeret felel meg. Terveink szerint ennek a 471 ÉKSz-beli jelentésnek, és 371 igei vonzatkeretnek azon elemeit fogjuk első lépésben hierarchiába rendezni, amelyek angol megfelelői a BCS részét képezik. Az így kapott magyar fogalmi hálóba fogjuk ezek után második lépésben beleilleszteni a még nem lefedett BCS igei jelentéseit. Ilyen módon, bár a HuWN igei része maximálisan kompatibilis marad mind a Princeton WordNettel, mind a BalkaNettel (hiszen a BCS igei részét megfeleltetjük PWN synseteknek), mégis jobban fogja tükrözni a fogalmi háló a magyar nyelv lexikalizációs sajátosságait, mintha kizárólag angol irányból közelítettük volna meg a mag-wordnet készítését.

Lehetővé válik, hogy indokolt esetben (ld. 2.2.) több ÉKSz-beli jelentést összevonva alakítsunk ki magyar synseteket, illetve, hogy bevezessük a többszörös öröklődést a hierarchiában. Ahogyan utaltunk már rá, a PWN nem alkalmaz sem többszörös öröklődést, sem mesterségesnek megjelölt csomópontokat. Ennek ellenére az igei szófaj sajátosságai miatt tervezzük bevezetni ezeket a lehetőséget is. Bizonyos morfémák a magyarban (igekötők ill. képzők) kifejezhetnek akcióminőséget ill. az igével kifejezett állapot változását (ld. [6]). Ezeket az eseteket célszerű lenne a több-

[18] A következőkben leírtakat a EWN résztvevőinek módszertani döntéseire alapozva alakítottuk ki (ld. [1].

[19] A gyakorisági listák elkészítéséért Miháltz Mártont, Nagy Viktort és Oravecz Csabát illeti köszönet.

szörös öröklődés és a mesterséges csomópontok lehetőségét kihasználva elhelyezni a kialakítandó hierarchiában.[20] Így például az általában *el-*, *fel-* és *meg-* igekötős inchoatív akcióminőséget kifejező igék az alapige jelentése szerint és akcióminőségük szerint is kapnának hipernímát. A *felkacag* ige ugyanúgy lenne a mesterséges *kezd* akcióminőséget jelző csomópont hiponímája, mint a *kifejez*, *kimutat*-é.

3.2 A felveendő relációk típusai

A felveendő relációkat a PWN relációtípusaiból kiindulva határoztuk meg (ld. [4]).

Szemantikai és időbeli relációk

Az alapvető szemantikai reláció igejelentések között az *implikáció*: V_1 akkor implikálja V_2-t, ha mindahányszor, amikor X V_1 igaz, X V_2-nek is igaznak kell lennie, függetlenül X megválasztásától. Például: a *töpreng* implikálja a *gondolkodik*ot, mert ha valaki töpreng, akkor az a valaki gondolkodik is. Hasonlóképpen, a *vásárol* implikálja a *fizet*et, hiszen ha valaki vásárol, akkor fizet is. Ezt a viszonyt általánosan a következő sémával tesztelhetjük:

V_1 implikálja V_2-t, ha abból, hogy „X V_1-zett", következik, hogy „X V_2-zött".

Ha valaki töprengett, akkor gondolkodott (is), illetve, ha valaki vásárolt, akkor fizetett (is). Implikációs viszony fennállhat időben diszjunkt események között is: ha valaki elveszített valamit, akkor – szükségképpen – előzőleg birtokolnia (is) kellett azt a valamit; ez a birtokviszony azonban az elvesztés pillanatában megszűnik, így a két esemény időben diszjunkt.

A PWN felfogása szerint az igék eseményeket írnak le. Pontosítva: minden V_i igéhez (vonzatkeretének kitöltése után) hozzárendelhető egy e_i esemény, amelyet a szóbanforgó kijelentés leír. Mivel minden esemény időben zajlik, tetszőleges e eseményhez hozzárendelhető az esemény *futási ideje*, $\tau(e)$, ami általában egy időintervallum. Két esemény között az alábbi temporális viszonyok relevánsak a WN szempontjából:

- kotemporalitás: e_1 kotemporális e_2-vel, ha $\tau(e_1) = \tau(e_2)$;
- időbeli tartalmazás: e_1 időben tartalmazza e_2-t ha $\tau(e_2) \subseteq \tau(e_1)$ (azaz e_2 minden pontja egyben e_1-nek is pontja);
- szigorú időbeli tartalmazás: e_1 időben szigorúan tartalmazza e_2-t ha $\tau(e_2) \subseteq \tau(e_1)$, de $\tau(e_1) \neq \tau(e_2)$ (azaz e_1-nek van olyan pontja, ami e_2-nek nem pontja);
- időbeli megelőzés: e_1 időben megelőzi e_2-t, ha $\tau(e_1) < \tau(e_2)$ (azaz: e_1 minden pontja korábbi, mint e_2 bármely pontja).

[20] A mesterséges csomópontok rendszerét a GermaNet-re támaszkodva (ld. [5]) alakítjuk ki.

Szinonímia

V_1 **szinonim** V_2-vel, ha

1. V_1 és V_2 kölcsönösen implikálják egymást és
2. szükségképpen kotemporálisak.

Részesemény

V_1 (valódi) **részeseménye** V_2-nek, ha

1. V_1 implikálja V_2-t és
2. e_1 szükségszerűen (szigorú) időbeli része e_2-nek.

Például: *horkol - alszik*, *álmodik - alszik*, *összead - számol*

Troponímia

A troponímia a főnévi hiponímia igék közötti megfelelője. Lényegében azt a viszonyt jelöli, amikor egy esemény egyfajta *módja* egy másik esemény kivitelezésének.

V_1 egy **troponímája** V_2-nek, ha

1. e_1 szükségszerűen időbeli része e_2-nek és
2. „X V_2-zött és közben V_1-zett” mondat *nem* jólformált.

Például: *gitározik - zenél*, *vánszorog - jár*, *prédikál - beszél*

Temporális prekondíció

V_1 **temporális prekondíciója** V_2-nek ha

1. e_2 implikálja e_1-t és
2. e_1 időben szükségképpen megelőzi e_2-t.

Például: *birtokol - elveszít*, *alszik - felébred*, *elalszik - alszik*

Okozás

Az okság alapvetően nem nyelvi kategória. A nyelvészetben azonban elfogadottá vált egy, a filozófiaival többé--kevésbé összeegyeztethető értelmezés, amit a PWN is felhasznál. E (nyelvészeti értelemben vett) oksági viszonynak a magyarban (is) van grammatikalizált megjelenése, a kauzatív alternáció. Ez azonban formailag meglehetősen szabályos és kiszámítható, így most nem foglalkozunk vele külön. A WN-ben releváns okságfogalom a példáik alapján részben a temporális prekondíció „tükörképe”, de vannak még további feltételek is:

V_1 **oka** V_2-nek ha

1. e_1 implikálja e_2-t és
2. e_2 nem előzheti meg e_1-et és
3. e_2 függ e_1-től abban a kontrafaktuális értelemben, hogy ha e_1 nem lenne (nem történne meg/abbamaradna), akkor e_2 sem lenne (nem történne meg / abbamaradna

Így például, ha valaki *forgat* valamit, akkor ez azt okozza, hogy az a valami *forog*. A forgatás implikálja a tárgy forgását, a két esemény kotemporális, és fennáll a kontrafaktuális függés is (ha ugyanis a dolog anélkül is forogna, hogy az ágens ezért bármit tenne, nem mondanánk, hogy forgatja az illető dolgot).

A kontrafaktuális feltétel meglehetősen lényegi, mert ez akadályozza meg, hogy a természetes nyelvi okság tranzitív legyen.

Abból, hogy

(1) Mari belakatolta az ajtót, mert zajt hallott.

és

(2) A zajt Mari vőlegényének közeledése okozta.

nem következtetünk arra, hogy

(3) Mari belakatolta az ajtót, mert közeledett a vőlegénye.

Sajnos a PWN példái azt mutatják, hogy okságon a szerzők elég gyakran pusztán az első két feltétel teljesülését értik. Így pl. az *ad - birtokol* is oksági viszonyban áll, holott valaki anélkül is egy dolog birtokába juthat, hogy azt adták volna neki (és akkor már a *megvesz - birtokol* párt is oksági viszonyban állónak kell tekinteni, és ott is ugyanaz a helyzet).

Az igék közötti okozási viszony legfeltűnőbb sajátossága, hogy az okot leíró ige *alanya* leggyakrabban a másik ige *tárgyának* változását idézi elő (de: *megvesz - birtokol*). Ha pl. X lelövi Y-t, akkor Y az, aki átmegy az ige jelentésében implicite kódolt állapotváltozáson (ő hal meg). Általában: ha V_1 és V_2 oksági viszonyban áll, akkor

1. „X V_1-te Y-t” implikálja, hogy „Y V_2-zött” és
2. X ágens, Y páciens thematikus szerepű.

Például: *fékez - lassul*, *lelő - meghal*, *gyűjt - gyűlik*

Antonímia

Az antoníma reláció nem synsetek, hanem különböző synsetek elemei között fennálló lexikális reláció a PWN-ben. Mivel tapasztalataink szerint az antonímia relációk meglehetős hűséggel átvehetők az angol eredetiből fordítás útján, így ezeket megőrizzük a HuWN-ben.

4 További feladatok

A további munkafázisok során, a HuWN információkinyerő rendszerben történő alkalmazására való tekintettel a következő irányelveket szándékozunk követni: Annak érdekében, hogy az információkinyerő rendszer eseményleíró kereteken alapuló felismerőképessége javuljon, minden további felvett jelentéshez hozzárendeljük a megfelelő igei vonzatkeretet, ami segít az események szereplőinek meghatározásában. Ezenkívül megpróbáljuk integrálni a HuWN igei részébe az EWN-ben kidolgozott szemantikai attribútumok rendszerét, az ún. Top Ontology-t, mely a hierarchiát az igékhez rendelhető szemantikai jegyekkel egészíti ki.

Bibliográfia

1. Alonge, A.: Definition of the links and subsets for verbs. Deliverable D006. Technical Report WP4.1, EuroWordNet, LE2-4003, Amsterdam (1996)
2. Christodoulakis, D.: Balkanet. Deliverable D2.1. Technical Report WP 2. BalkaNet IST-2000-29388 (2000)
3. Cristea, D.: Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets. Romanian Journal of Information Science and Technology 7 (2004) 124–145
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
5. Hamp, B., Feldweg, H.: GermaNet — A Lexical-Semantic Net for German. In Vossen, P., ed.: Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Association for Computational Linguistics, New Brunswick, New Jersey (1997) 9–15
6. Kiefer, F.: Jelentéselmélet. Corvina, Budapest (2000)
7. Kunze, C.: Semantics of Verbs within GermaNet and EuroWordNet (1999)
8. Tufis, D.: Balkanet: Aims, methods, results and perspectives. Romanian Journal of Information Science and Technology 7 (2004) 1–35
9. Vossen, P.: The EuroWordNet Base Concepts and Top Ontology, Deliverable D017, D034, D036. Technical Report EuroWordNet (LE 4003), University of Amsterdam, Amsterdam (1998)
10. Vossen, P.: EuroWordNet General Document. Technical Report EuroWordNet (LE2-4003, LE4-8328) (2005)

Taxonómia felismerése dokumentumszerkezetből

Lendvai Piroska

Tilburg University, Dept. of Language and Information Science
PO Box 90153, 5000 LE, Tilburg, Hollandia

p.lendvai@uvt.nl

Kivonat: Munkánk orvosi enciklopédiák szövegéből kinyerhető taxonomikus kapcsolatok automatikus felfedezésére irányul, melyet szabadszavas, egészségügyi tematikájú kérdések automatikus megválaszolásában használunk fel. Az enciklopédiák szócikkeit különböző szövegszegmensi szinteken a témakörre jellemző szemantikai annotációval láttuk el. Mesterséges intelligencia alapú tanulási kísérleteket írunk le, amelyek során a taxonomikus kapcsolatok automatikus felismerésének betanítása és értékelése történik.

1 Bevezetés

A holland ROLAQUAD projekt keretében fejlesztett intelligens válaszadó rendszer célja, hogy szabadszavas, egészségügyi tematikájú kérdéseket válaszoljon meg. A rendszer alapját két, holland nyelvű orvosi enciklopédia kézzel annotált szócikkei képezik. A rendszer felismeri a felhasználó kérdésében a kérdezett tárgyszót (pl. „agyhártyagyulladás"), s hogy annak mely aspektusára kérdez rá a felhasználó (pl. „tünetei"). Ezután a referenciadokumentumok szemantikai annotációjához illeszti ezeket, majd a legpontosabban illesztett dokumentumrészt adja vissza válaszként.

Előfordulhat azonban, hogy a felhasználó kérdése alulspecifikált, például mert a kérdezőnek nincsenek pontos ismeretei az adott területről. Ilyen kérdés lehet a „Mik az agyhártyagyulladás tünetei?", mert a rendszer referenciaszövegében az „Agyhártyagyulladás" szócikk két szakaszában is előfordul a `'Tünetek'` szemantikai annotáció. A helyes válaszadáshoz szükséges felismerni, hogy a szócikk az agyhártyagyulladás két típusát is körülírja, vagyis taxonomikus kapcsolatokat tartalmaz, és hogy emiatt a felhasználót a kérdése pontosítására kell megkérni.

Ahhoz, hogy a rendszer dinamikus módon tudjon ilyesfajta visszakérdezéseket generálni, szükséges, hogy a referenciadokumentumokból automatikusan ki tudja szűrni azokat, amelyek a címszóban megnevezett entitásnak több altípusával is foglalkoznak. Munkánk erre tesz kísérletet, a dokumentum szerkezetére vonatkozó szemantikai annotáció alapján. Az alkalmazott tanuló algoritmusnak azt kell felismernie, hogy a címszóban megjelölt entitás altípusait tárgyalja-e az adott enciklopédia-szócikk olyan részletesen, hogy a címszóban megjelölt entitás vagy annak egy aspektusa végeredményben az altípusok által definiálódik. Pl.: kortikoszteroidok{alkalmazása külsőleg;alkalmazása belsőleg}, vékonybél-daganat{jóindulatú;rosszindulatú}, steri-

lizálás{férfiaknál;nőknél}, stb. A feladatot kétfajta megközelítésben is elvégezzük. Ezekben az algoritmus a szócikkek különböző jellemzőit használja fel a tanulás során, pl. az abban előforduló szavakat, egészségügyi fogalmakat, statisztikai gyakoriságot, stb.

A javasolt eljárás nem morfológiai/szintaktikai alapú [2], hanem közvetlenül a dokumentumok szerkezete és az azok fölötti szemantikai tartalmak alapján azonosítja a taxonomikus kapcsolatot. Korpuszunk dokumentumai kevesebb strukturális hierarchiát mutatnak, mint a [4] által ontológia létrehozásához felhasznált szövegek, a klasszifikáció pedig nem szegmentálásra [1], hanem előre meghatározott szövegszegmensek közötti taxonomikus kapcsolatok felfedezésére irányul.

A következőkben bemutatjuk a felhasznált korpusz szemantikai annotációjának elvét és a különböző szemantikai címkéket. A 3. szakasz a konkrét gépi tanulási kísérleteket írja le, részletezve az alkalmazott algoritmust, a két tanulási feladatot, a tanulásban felhasznált attribútumokat, és a kapott eredményeket. Az utolsó részben összefoglaljuk és értékeljuk munkánkat.

2 A korpusz szemantikai annotációja

A rendszer által felhasznált referencia-dokumentumgyűjtemény a holland nyelvű Merck orvosi kézikönyv és a Spectrum egészségügyi enciklopédia szócikkeiből áll. Korlátozott számú, a témakörre jellemző szemantikai annotációt kaptak a szavak szintjén a fogalmak, a mondatok szintjén a mondat témája, a szakaszok szintjén pedig a szakasz témája. Pl. az "Agyhártyagyulladás" szócikk fordításának második mondata:

<SZAKASZ: `Definíció;Ok`> ... <MONDAT: `Definiál;Fertőz;Okoz`> *A betegség kórokozói különféle* <FOGALOM: `mikroorganizmus`>*vírusok*</FOGALOM> *és* <FOGALOM: `mikroorganizmus`>*baktériumok*</FOGALOM> *lehetnek*.</MONDAT> ... </SZAKASZ>

A teljes korpusz több, mint 3000 dokumentumból áll, ezek 54%-a azonban nem használható fel a kísérletekben, mert szerkezetileg csak egyetlen szakaszból állnak. A több szakaszból álló dokumentumok között 128 olyan található, amely szemantikailag rekurzív szerkezettel rendelkezik, vagyis a bevezető szakaszt követően legalább két olyan szakasza van, amelyek temájukban megegyeznek, pl. két szakasz is tárgyal `Tünetek`-et vagy `Megelőzés`-t.[21]

A kézi annotálás a következő protokoll alapján történt. Egy dokumentum szakaszához 15 különböző címkét lehet hozzárendelni, pl. `Definíció` ('a szakasz a címszó-entitás definícióját tartalmazza'), `Ok` ('a szakasz egy entitás előfordulásának okát írja le'), `Megelőzés`, stb. A teljes címkelistát az 1. Táblázat első oszlopa mutatja. Egy szakaszhoz természetesen több címke is hozzárendelhető.

[21] A kísérlet annotálatlan szövegeken is elvégezhető, ha azok konzisztens (al-)alcímeket tartalmaznak.

Szakasztípus	Mondattéma	Fogalom
alkalmazás	jellemez	testi funkció
ok	okoz	testrész
következmény	alfaja	betegség
fertőzés	fertőz	betegség jellemzője
definíció	definiál	betegség tünete
diagnózis	diagnosztizál	diagnosztikai eljárás
betegségek	hasonlít	időtartam
elsősegély	szinonima	mikroorganizmus
előfordulás	előfordul	személy
mellékhatások	mellékhatása	személy jellemzője
kezelés	kezel	kezelés
tünetek	tünete	kezelés jellemzője
megelőzés	megelőz	
módozatok		
formák		

1. Táblázat. Szemantikai annotáció címkéi a dokumentum három szintjén.

Mondatszinten 13 témát annotáltunk, egy-egy mondatot szintén több címke is jellemezhet; ezeket lásd az 1. Táblázat középső oszlopában. A szavak illetve a szókapcsolatok szintjén 12 egészségügyi fogalomtípust címkéztünk; lásd az 1. Táblázat harmadik oszlopát.[22]

3 Gépi tanulási kísérletek

A kísérleteket felügyelt tanulási feladatként formalizáljuk, melyeket a TiMBL szoftvercsomag 5.1 verziójának IB1 algoritmusával végzünk el[23]. Az algoritmus a *k*-legközelebbi szomszéd ('*k*-nearest neighbour', '*k*-NN') tanulási módszert használja, lásd pl. [5]. Ennek a felügyelt módszernek a működési elve példányalapú tanulás, vagyis egy feladatot példák attribútum-vektoraként jelenítünk meg, és az algoritmus ezekhez tanul meg osztályokat rendelni. Az algoritmust alapbeállításokkal futtattuk (*k*=1, a példák között euklidészi távolság mérése, 'gain ratio' attribútumsúlyozás). A kísérleteket a „kihagyok egyet" ('leave-one-out') predikció módszerével folytattuk le. Két kísérletsorozatot végeztünk el, amelyekben különbözőképpen közelítettük meg a taxonomikus kapcsolatok feltárását.

[22] A domén-entitások jó része egyszerű, formai jellemzők alapján kinyerhető annotálatlan szövegekből is, erről lásd pl. [3].

[23] Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2004). TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report Series 04-02. http://ilk.uvt.nl/timbl

3.1 Osztályozási feladatok

Az első kísérletsorozatban az algoritmus feladata annak eldöntése, hogy egy dokumentum két adott szakasza taxonomikus testvérpárt ír le vagy sem. Az „Agyhártyagyulladás" szócikk például négy szakaszból áll: (1) bevezetés, (2) „Bakteriális agyhártyagyulladás", (3) „Megelőzés", és (4) „Vírusos agyhártyagyulladás". Ebből a {2,4} szakaszok taxonomikus testvérpárt alkotnak: azonos rangú taxonomikus relációban állnak a szócikk által leírt entitással (agyhártyagyulladás), mivel mind a (2), mind a (4) szakasz témája a szemantikai annotáció szerint `'Okoz'`, `'Tünetek'`, és `'Kezelés'`. A feladat itt annak a felismerése, hogy a két szakasz tartalmi átfedéseket tartalmaz ugyan, de a témák fő argumentuma különbözik egymástól. Az algoritmusnak tehát nemcsak a két szakasz közötti hasonlóságokat, de a különbségeket is számon kell tudni tartani.

A második kísérletsorozatban a feladat olyan pozitív példák felismerése, ahol a szakaszpár egyik tagja a címszóban megjelölt egészségügyi fogalmat általánosságban jellemzi, míg a másik annak altípusát írja le. A példadokumentumban az {1,2} és az {1,4} szakaszpárok írnak le ilyen, alárendeltségi kapcsolatot egy általános fogalom és annak alfaja között, mert a bevezető az általános fogalmat, az agyhártyagyulladást írja körül, míg a 2. illetve a 4. szakasz annak egy specifikus alfaját. Ez a feladat látszólag még nehezebb, mint az első megközelítés, mert egy enciklopédia-szócikk bevezetője tartalmilag szükségszerűen utal az összes következő szakaszra, és a szakaszok is utalhatnak egymásra – az algoritmus dolga itt az, hogy felismerje, hogy az egyik szövegszegmens a másik egy adott elemét részleteiben tárja fel. Bizonyos szempontból a szegmensek közötti kapcsolatot nemcsak alárendeltséginek foghatjuk fel, de anaforikusnak is.

Az algoritmus számára az {1,3} szakaszpár mind a taxonomikus testvérpárnak, mind az alárendeltségi kapcsolatnak negatív példája.

Mivel a dokumentumgyűjteményben két különböző típusú enciklopédia szócikkei szerepelnek, hasznosnak láttuk ezeket egymástól különválasztva feldolgozni. A Spectrum enciklopédia szócikkei igen következetesen strukturáltak, a szakaszok címei konzisztensen visszatérnek, vagyis szerkezetelemzéshez „ideális" anyagot nyújtanak. A Merck kézikönyv dokumentumaiban a szerkezet lazább, az alcímek esetlegesebbek, a szócikkek pedig hosszabbak, mint a Spectrumban, ezért a Merck feldolgozása inkább hasonlítható egy „valós" szemantikai elemzési környezethez.

A két különböző osztályozási feladatban szükségszerűen különbözik a vonatkozó pozitív és negatív példák száma is. A taxonomikus testvérpárok feladathoz 174 pozitív és 523 negatív példát tudunk generálni a Spectrum enciklopédiából, és jóval kevesebbet a Merck kézikönyvből (49 pozitív, 161 negatív példa). Az alárendeltségi kapcsolat meghatározásának feladatához valamivel több pozitív és valamivel kevesebb negatív példa áll rendelkezésre, ami elősegítheti a hatékonyabb osztályozást (Spectrum: 255 pozitív és 442 negatív, Merck: 51 pozitív és 159 negatív példa).

3.2 Felhasznált attribútumok

A szakaszpárokat különbözőképpen jelenítjük meg az egyes kísérletek során. Az attribútumvektor komponensei numerikus elemekből (főként bináris bitekből) állnak, amelyek a következő információt hordozzák: a két szakaszban előforduló

(a) közös szavak (szóhalmazban, 'bag-of-words')
(b) közös szóhármasok (trigram-ok)
(c) dokumentumcím – szakasz alcím(ek) – vizsgált szakasz(ok) közös szavai
(d) közös egészségügyi fogalmak
(e) közös mondattémák.

Fontos tudni, hogy egy-egy attribútumcsoport kódolása nagyságrendekkel különbözhet egymástól: a szóhalmaz vektora pl. 7288 elemből áll, mert ekkora a korpusz lexikonja. Ha egy szó a vizsgált szakaszok mindegyikében előfordul, a szót jelző bit értéke 2, ha csak az egyik szakaszban, a bit értéke 1, ha egyik szövegszegmensben sem fordul elő, a bit értéke 0. A szóhármasok vektora 1155 elemből áll, mert ekkora a korpuszban a három vagy annál nagyobb (jelen esetben: 36-ig terjedő) gyakorisággal előforduló trigramok lexikonja. A dokumentumcím-szakaszalcím(ek) egybeesése viszont mindössze 4 bitből áll, a közös fogalmaké 12, a közös mondattémáké 13 elemből (lásd 1. Táblázat).

3.3 Eredmények

Az algoritmus teljesítményét többféle mérték szerint is értékeltük: globálisan számított mértékek a pontosság ('accuracy', az általános hibaszázalék ellentettje), a mikro-F-pontszám (az összes példa alapján kiszámított F), a makro-F-pontszám (a két osztály alapján kiszámított F), valamint az osztályokra levetített pontosság ('precision'), teljesség ('recall'), és ezek harmonikus középértéke, az F-pontszám (2PreRec/Pre+Rec). Az értékelés során a legnagyobb figyelmet a pozitív példák klasszifikációjára vonatkozó F-pontszámnak szenteljük, mert ez mutatja, mennyire jól képes az algoritmus a fogalmi taxonómia különböző elemeinek (mellérendelt kapcsolatban lévő „testvéreknek", vagy „alá-fölérendelő" hiperonim-hiponim kapcsolatoknak) a felismerésére.

Korpusz	Attribútum				+ osztály			– osztály		
		Acc	Fmik	Fmak	Pre	Rec	F	Pre	Rec	F
Spectrum	szóhalmaz	55	56	44	20	23	20	72	66	69
	szóhármas	*61*	*61*	*48*	*22*	*21*	*21*	*74*	*75*	*74*
	(al-)címek	86	85	78	98	47	64	85	100	92
	fogalmak	*75*	*75*	*67*	*51*	*49*	*50*	*83*	*84*	*84*
	mondattémák	88	88	85	78	76	77	92	93	92
Merck	szóhalmaz	73	74	65	44	57	50	86	78	81
	szóhármas	*71*	*71*	*58*	*37*	*35*	*36*	*80*	*82*	*81*
	(al-)címek	79	75	61	69	22	34	80	97	88
	fogalmak	*69*	*69*	*56*	*33*	*33*	*33*	*80*	*80*	*80*
	mondattémák	79	80	73	55	65	60	89	84	86

2. Táblázat. Mellérendelt viszonyú taxonomikus testvérpárok meghatározása a kétfajta korpuszban, különböző attribútumok alapján.

Az első kísérletsorozat eredményeit a 2. Táblázat tartalmazza. Megállapítható, hogy legjobb eredményt akkor tudtuk elérni taxonomikus testvérek azonosításában, ha a szakaszpárokat az azokban előforduló azonos mondattémákként ábrázoltuk: a szabadabb formátumú Merck szövegeiben 60 F-pontszámot, a Spectrum szövegeiben pedig, amelynek dokumentumai szabályosabb szerkezetbe rendezettek, 77 F-pontszámot értünk el. A Spectrum anyagán a második legmagasabb F-pontszámot (64) a dokumentumcím – szakasz alcímek – közös dokumentumszavak egybeesésének információja alapján zajló kísérletben értük el. Ebből arra következtetünk, hogy dokumentumszerkezet alapján szemantikai tartalmat fel lehet ismerni abban az esetben, ha a szerkezet jelölése következetes. A szakaszokban szereplő egészségügyi fogalmak csak harmadrangú információt nyújtanak arról, hogy adott szócikk két szegmense tartalmilag egymás mellé rendelhető-e.

A Merck kézikönyv szócikkein elért eredményekből kitűnik, hogy ezeknek a dokumentumoknak a felépítése más, mint a Spectrumban, mert az (al-)címek egybeesésének információja a pozitív osztályt nem, a negatív osztályt viszont igen jól képes jellemezni (88 F). Megállapítható, hogy az azonos fogalmi körbe tartozó, de különböző séma alapján felépített dokumentumokban más és más attribútumcsoportok hordoznak taxonómiai információt. A leginformatívabb természetesen az, hogy mely mondattémák esnek egybe a két szegmens között; ezt optimális esetben a témákkal egybeeső alcímek jelzik.

A szóhalmaz, illetve a szóhármasok által hordozott információ a Merck anyagán jobb eredményt ad, mint a Spectrumén, ami valószínűleg azzal magyarázható, hogy a Merck szócikkei hosszabbak és szabadabb megfogalmazással íródtak. Ez utóbbira tanú az is, hogy a Merckben a témaköri fogalmak megléte, illetve valószínűleg inkább azoknak a hiánya, kevesebb információt tud nyújtani, mint maguk a dokumentumban szereplő szavak (33 F, lásd a táblázat utolsó előtti sorát). A szövegekben megjelenő fogalmak statisztikailag tulajdonképpen csak egy esetben adnak jobb eredményt, mint akár a szóhalmaz, akár az alcímek egybeesése: az alá-fölérendeltségi kapcsolat megállapításakor a Spectrum anyagán (51 F). Ezzel rá is tértünk a második kísérletsorozat tárgyalására (lásd: 3. Táblázat).

Korpusz	Attribútum				+ osztály			– osztály		
		Acc	Fmik	Fmak	Pre	Rec	F	Pre	Rec	F
Spectrum	szóhalmaz	55	54	49	36	28	31	63	71	67
	szóhármas	*54*	*53*	*48*	*35*	*28*	*31*	*62*	*69*	*66*
	(al-)címek	69	64	59	70	27	39	69	93	79
	fogalmak	*64*	*64*	*61*	*51*	*51*	*51*	*72*	*71*	*71*
	mondattémák	85	85	84	78	83	80	90	87	88
Merck	szóhalmaz	79	77	68	58	43	49	83	90	84
	szóhármas	*74*	*74*	*64*	*46*	*45*	*45*	*82*	*83*	*83*
	(al-)címek	76	65	43	-	-	-	75	100	86
	fogalmak	*78*	*77*	*69*	*56*	*49*	*52*	*84*	*87*	*86*
	mondattémák	83	82	74	69	53	60	86	92	89

3. Táblázat. Alá-fölérendeltségi viszonyú (hiperonim-hiponim) szakaszpárok meghatározása a kétfajta korpuszban, különböző attribútumok alapján.

Érdekes megfigyelni, hogy bár a hiperonim-hiponim kapcsolat meghatározása nehezebb feladat lehet, többek között mivel a rövidke bevezető szakasz anyagára kell támaszkodni, aminek nincsenek alcímei, de a korábban tárgyalt anaforikus jelleg miatt is, a 3. Táblázat pontszámai mégis némileg magasabbak és kiegyensúlyozottabbak, mint a mellérendeltségi feladaton elértek. Technikai kérdés, hogy ez vajon annak köszönhető-e, hogy ebben a feladatban valamennyivel több pozitív példa raktározható el a memóriában a tanulási fázis során.

Fontos eredmény, hogy legmagasabb pontszámot ebben a modellben szintén a mondattémák közötti átfedés alapján lehet elérni: a Merck szövegekben 60 F-pontszámot (ez megegyezik a taxonómiának testvérpárok alapján történő felismerésével), a Spectrum szövegeiben pedig 80 F-pontszámot értünk el, ami magasabb, mint a testvérpárok alapján történő felismerés esetében.

Természetesen a szakasz alcímek egybeesése ehhez a feladathoz nem adhat plusz információt, mert a bevezető szakasznak, ami mindig a szócikk első szegmense, soha nincs alcíme. A szakaszokban szereplő egészségügyi fogalmak a Spectrum esetében ismét viszonylag jól jellemzik, hogy adott szócikk két szegmense tartalmilag egymásra mutat egy alá-fölérendeltségi kapcsolatban, a Merck anyagán viszont gyakorlatilag nem adnak többletinformációt az egyszerű (bár nagy számú) szóhalmazhoz képes.

4 Értékelés

Munkákban arra tettünk javaslatot, hogyan lehet gépi tanulási kísérleteket felépíteni fogalmi taxonómia elemeinek kinyerésére strukturált, szemantikailag annotált dokumentumokból, jelen esetben holland, egészségügyi témájú enciklopédia-szócikkekből. A kísérleteket az motiválja, hogy olyan általános módszert találjunk, amelyet következetesen felépített, leíró jellegű dokumentumokra – pl. enciklopédiák, wikipédiák, értelmező szótárak – lehet alkalmazni taxonómia kinyerésére. Megállapítottuk, hogy a taxonómia komponenseit legalább kétféle modellel írhatjuk le: kereshetjük az egy dokumentumban előforduló taxonomikus testvérpárokat, illetve közvetlenül az általános fogalmat és annak egy altípusát. A kísérletekhez példaalapú tanuló algoritmust használtunk, amelynek betanítása öt különböző attribútumcsoporton történt. Mindkét módszerrel megközelítőleg azonos eredményt értünk el, a legmagasabb F-pontszámot (80) a Spectrum egészségügyi enciklopédiából generált példákon: az algoritmus egy általános egészségügyi fogalmat és annak egy altípusát leíró szakaszpárokat azonosított be a szakaszok tematikai egybeesése alapján. Ez első hallásra triviálisnak tűnhet, azonban egyáltalán nem kézenfekvő, hogy a tematikai egybeesés éppen alá-fölérendeltségi kapcsolatra utal, hiszen éppúgy utalhat egy általános anafora-katafora vagy rész-egész kapcsolatra is, hiszen egy dokumentum bevezető szakaszának funkciója, hogy a teljes mondanivalót előrevetítse. Ezért a modellben a negatív példák felismerését szintén nagy pontossággal kell megoldani. A táblázatokból látható, hogy a negatív példák osztályozása jó eredménnyel történik.

A tematikai egybeesés attribútumvektort kézzel annotált címkékből generáltuk. A jövőben arra fogunk sort keríteni, hogy ezt az attribútumot gépi tanulással ki tudjuk nyerni a mondatokból, és további, a szövegről magas szintű szemantikai és morfo-szintaktikai információt közvetítő attribútumokkal egészítsük ki.

A viszonylag kevés számú példa és a korpusz „zajossága” – nem szakértők által történt annotálása és szócikk-szegmentálása – valószínűvé teszi, hogy az itt bemutatottaknál egy színvonalasabban felcímkézett korpuszon jobb eredményeket lehetne elérni a javasolt módszerrel. Amennyiben a taxonomikus kapcsolatokat megbízhatóan tudjuk felismerni, a folyamatot beépítjük az orvosi válaszadó rendszerbe, a taxonómia elemeit pedig ontológia létrehozására használjuk fel.

Bibliográfia

1. Cho, P., Taira, R., Kangarloo, H.: Automatic Segmentation of Medical Reports. Proc. of AMIA Symposium (2003) 155-159
2. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: Buitelaar, P., Magnini, B., Cimiano, P. (Eds): Ontology Learning from Text: Methods, Applications, Evaluation. IOS Verlag (2005)
3. Lendvai, P.: Conceptual Taxonomy Identification in Medical Documents. In: Proc. of The Second International Workshop on Knowledge Discovery and Ontologies (2005) 31-38
4. Makagonov, P., Figueroa, A., Sboychakov, K., Gelbukh, A.: Learning a Domain Ontology from Hierarchically Structured Texts. Proc. of ICML workshop on Learning and Extending Lexical Ontologies by using Machine Learning Methods (2005) 50-57
5. Mesterséges Intelligencia. Szerk.: Futó, I. Aula Kiadó (1999)

III. Fordítás és szótár

A MetaMorpho fordítóprogram projekt 2005-ben

Tihanyi László

MorphoLogic
1126 Budapest, Orbánhegyi út 5.
tihanyi@morphologic.hu

Kivonat: Az előadásban felsorolom a MetaMorpho fordítóprogram projekt 2005-ös évben elért eredményeit a MorphoLogicon belül. Röviden ismertetem a MorphoLogic, a Nyelvtudományi Intézet és a Szegedi Egyetem alkotta konzorcium keretében folyó magyar–angol fordítóprojekt első évében történt eseményeket. Beszámolok a MoBiCAT internetes fordítószolgáltatás elsőéves tapasztalatairól. Ez a cikk folytatása a 2003-as és 2004-es MSZNY-konferencián elhangzott projekt beszámolóknak [1],[2], és összefoglalja a 2005-ös év legfontosabb fejlesztéseit.

1 Programfejlesztések

1.1 Január

Megoldottuk a birtokos szerkezetek felhasználói szótárba történő felvételét. Az egyértelműség biztosítása érdekében kötelezővé tettük a genitivus jelölését, és egy külön beállítással biztosítottuk, hogy az eredményben jelöletlen maradhasson.

Formátumkonverzió a mintákhoz. A magyar-angol fordítóprogram fejlesztés keretein belül felmerült az igény a MetaMorpho projektben használt belső mintaformátum (MMD) szabványos XML alakra történő konverziójára.

Karakter-alapú elszámolás. A MoBiCAT szolgáltatás havi előfizetéses rendszerben működik. Erre az ad lehetőséget, hogy a felhasználó egyszerre csak egy-egy mondat fordítását igényelheti. A teljes szöveg fordítását megengedő, böngészőből vagy Wordből működtethető fordítói szolgáltatások díjazásának a fordított szöveg mennyiségével kell arányban állnia. A kifejlesztett karakterszámlálás a forrás vagy a célnyelvi karakterek száma alapján is működtethető.

A RuBi mintabővítő eszközünkben megjelent a szakterület- (vagy domain-) kezelés. Szabadon definiálhatók új szakterületek, és különböző területeken egy adott szóhoz különböző jelentéseket rendelhetők. Fordításkor a megfelelő szakterület választása mellett a szó a kívánt jelentésével fog szerepelni.

Beindult a MetaMorpho tesztablaka a www.metamorpho.hu oldalon. Az MmoText egy olyan tesztfelület, amellyel korlátozott hosszúságú angol szövegek fordíthatók le magyarra. A felület ingyenes, és azonnali kipróbálására ad lehetőséget. A felület

kódszámos védelemmel van ellátva, így a fordítandó szöveggel együtt egy csak emberek által értelmezhető számsorozatot is meg kell adni.

1.2 Február

Elválasztottuk a belső és a termékbe szánt mintabővítő változatokat. Egy jól kitöltött minta létrehozásához szakmai képzettségre volt szükség, ezért ezeken egyszerűsítetünk Ugyanakkor belső használatban ezeknek a tulajdonságoknak a helyes beállítását továbbra is biztosítani akartuk.

Bevezettük a kompatibilitási operátort, amely a minták típusának közvetlen levizsgálhatóságán keresztül a minták számának, és így a nyelvtan méretének csökkenéséhez vezetett.

Megszületett az első magyar–angol minta konverter. A fejlesztésben csak konverternek hívott eszköz, a megírt szabályokat alakítja át egy olyan még mindig a megírt szabályokkal azonos formátumú forrássá. A konverter valósítja meg pl. az öröklési mechanizmust, de a metatulajdonságok segítségével a nyelvtan írója tetszőleges átalakításokat és kiegészítéseket végezhet.

Megvalósítottuk az előfizetők saját webes adminisztrációs felületét. A regisztrált felhasználók ezen a felületen ellenőrizhetik pl. különböző fordítószolgáltatásaik állapotát.

A MorphoWord Pro változatát egy USB-portba helyezhető hardverkulcsos védelemmel láttuk el.

Elértük, hogy a felhasználói minták mindig érvényre jussanak, azaz felülbírálják a rendszerbe épített tudást.

Nyelvileg értékeltük a szolgáltatás közben létrejövő fordításokat, és a gyakori ismeretlen szavakkal kiegészítettük a szótárat.

Beindult a hamarosan nyilvánosan is elérhető lesz a MorphoWeb weblapfordító. A programtól tetszőleges angol internetes oldalt fordítása kérhető. A program követni tudja a linkeket is, így a fordított oldalról továbblépve további fordított oldalakra juthatunk.

1.3 Március

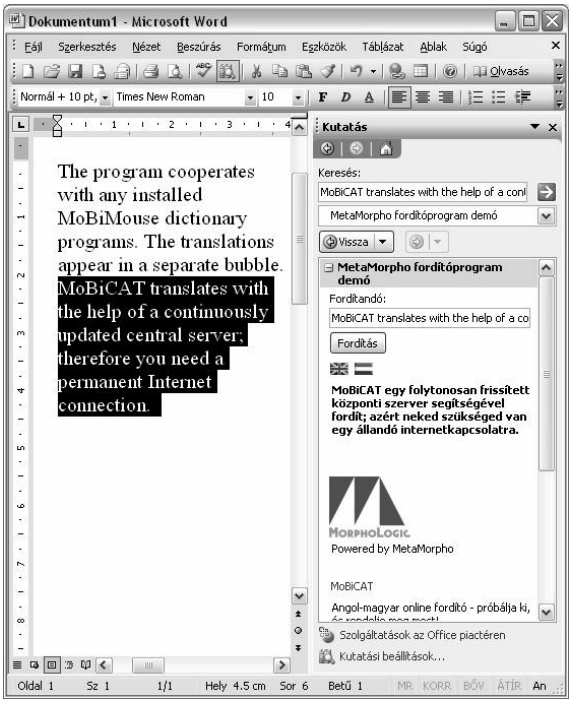
A programot összehasonlítottuk más fordítóprogramokkal. A vizsgálathoz angol-német fordítókat használtunk, úgy hogy az angol szövegekhez három német és három magyar emberi fordítást készítettünk. A vizsgálat alapján megállapítható volt, hogy a legjobb eredményt az orosz PROMT érte el. A MetaMorpho a piacon létező fordítókkal gyakorlatilag azonos minőséget produkált. Kiemelendő, hogy a leghosszabb szó szerinti egyezést (12 szó) a MetaMorpho érte el.

	Systran	SDL	PROMT	Wordlingo	MMO
Bleu (1-4 gram)	0,4206	0,4138	0,4618	0,4158	0,4101
1 gram	0,6519	0,667	0,6848	0,6528	0,6230
max gram	10	7	8	10	12

Megvalósult a Microsoft Office 2003 szövegszerkesztőből a Kutatási felületen keresztül működtethető ingyenes angol-magyar és magyar-angol szótár, valamint az angol-magyar fordítóprogram demó.

Office 2003 Research angol-magyar szótár

Office 2003 Research fordítóprogram-demó



1.4 Április

Megvalósult a vonzatos és vonzat nélküli igék egységes kezelése az adatbázisban.

Az Európai Unió sikeres vizsgálatot folytatott a MetaMorpho programról. A részletes felmérés eredményeként a programot nemcsak az Uniós magyarról és magyarra fordításokhoz javasolják, de alkalmasnak találták további uniós nyelvpárok fejlesztésére is.

1.5 Május

A mySQL mellett már a szabadon termékbe építhető SQLite adatbázis-kezelővel is építhetők a felhasználói szótáraink.

Létrejöttek a MorphoWord termék telepítőprogramjai.

1.6 Június

Összehangoltuk a szintaktikai és morfológiai szótárainkban lévő szavakat, és kiegészítettük az adatbázisokat.

Elkészültek a MorphoWord program dokumentációi.

1.7 Július

A bálna (MoBiDic), az egér (MoBiMouse) és a macska (MoBiCAT) után megszületik a negyedik „MorphoLogic-állat”: a MorphoWord lepke, mely nevét a délamerikai kék Morpho lepkétől (is) kölcsönzi.

Elkészül a MorphoWord aktiválós szoftvervédelme.

A MoBiCAT az Internet Explorer mellett már Mozilla FireFox alatt is működik.

Egy felmérés készül a MoBiCAT-vásárlókról és szokásaikról. A felmérésben a szokatlanul magas érdeklődés mellett hasznos információkhoz jutunk a további fejlesztésekhez.

1.8 Augusztus

Reklámakciót szervezünk a Sziget-fesztivál alatt.

Kiegészül a MorphoWord telepítőprogramja, amely a MorphoLogic-szótárak demó változatai mellett az xPlace intelligens szócserélőt is tartalmazza.

Kialakul a MorphoWord designja.

Tesztelések folynak.

Befejeződik a hardverkulcsos védelem fejlesztése.

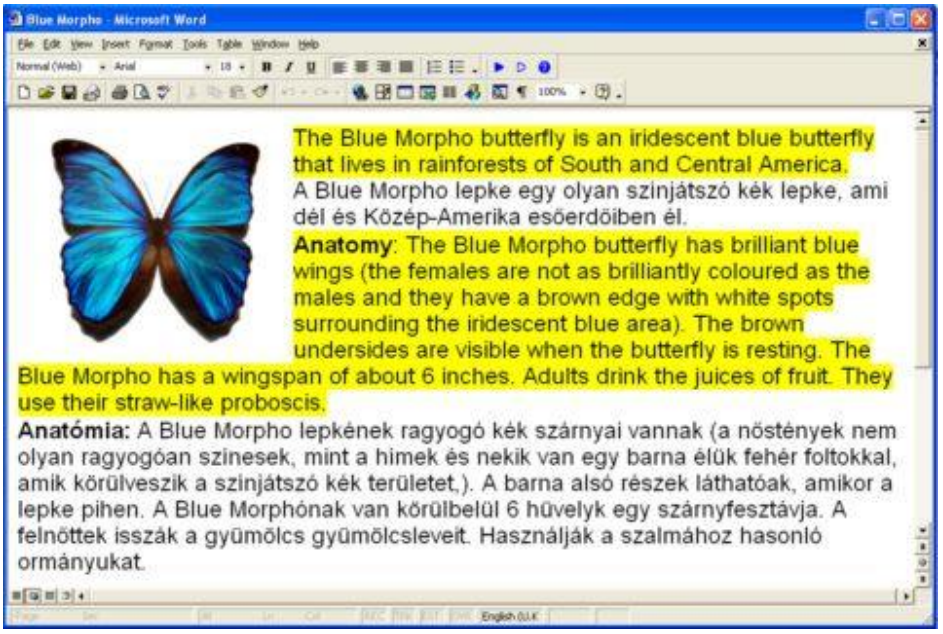
Lezáródik a RuBi (RuleBuilder) szabálybővítő fejlesztése. A felületet az aktuális szabály típusától/szófajától függően (itt főnév) egy másik panel egészíti ki, ahol további lexikális jegyek állíthatók be. A program TMX fájlok importjára is képes

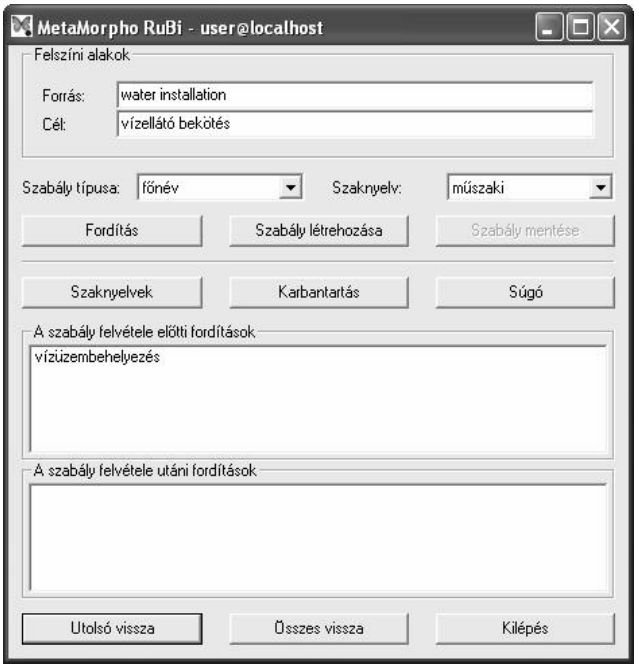
.

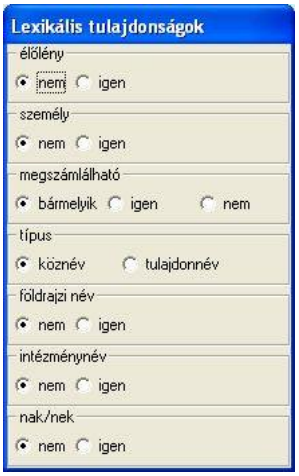
1.9 Szeptember

Megszervezzük a CD-gyártást és az értékesítést. Az interneten publikáljuk a programmal kapcsolatos információkat. Véglegesülnek a szoftvervédelmi megoldások. Elkészül a MorphoWord termék.

A MorphoWord hagyományos CD-n kapható dobozos program a Microsoft Word szövegszerkesztőkbe integrálódik. Már nem szolgáltatásként, hanem hagyományos dobozos termékként jelent meg. Két változatban készült el, az egyfelhasználós alapverzió mellett született egy intézményeknek szánt hálózatos megoldás is, amelyhez bővíthető adatbázis tartozik.

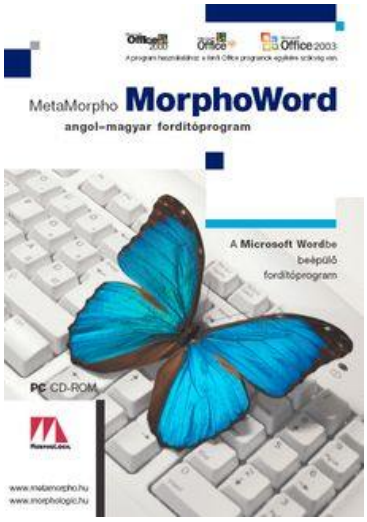

Blue Morpho - Microsoft Word
The Blue Morpho butterfly is an iridescent blue butterfly that lives in rainforests of South and Central America.
A Blue Morpho lepke egy olyan szinjátszó kék lepke, ami dél és Közép-Amerika esöerdöiben él.
Anatomy: The Blue Morpho butterfly has brilliant blue wings (the females are not as brilliantly coloured as the males and they have a brown edge with white spots surrounding the iridescent blue area). The brown undersides are visible when the butterfly is resting. The Blue Morpho has a wingspan of about 6 inches. Adults drink the juices of fruit. They use their straw-like proboscis.
Anatómia: A Blue Morpho lepkének ragyogó kék szárnyai vannak (a nöstények nem olyan ragyogóan szinesek, mint a himek és nekik van egy barna élük fehér foltokkal, amik körülveszik a szinjátszó kék területet,). A barna alsó részek láthatóak, amikor a lepke pihen. A Blue Morphónak van körülbelül 6 hüvelyk egy szárnyfesztávja. A felnöttek isszák a gyümölcs gyümölcsleveit. Használják a szalmához hasonló ormányukat.

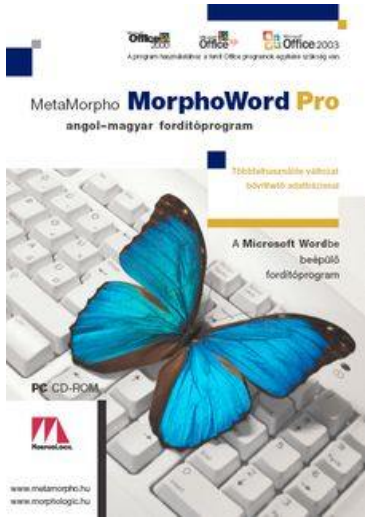

MetaMorpho RuBi - user@localhost
Felszíni alakok
Forrás: water installation
Cél: vízellátó bekötés
Szabály típusa: főnév
Szaknyelv: műszaki
Fordítás
Szabály létrehozása
Szabály mentése
Szaknyelvek
Karbantartás
Súgó
A szabály felvétele előtti fordítások
vízüzembehelyezés
A szabály felvétele utáni fordítások
Utolsó vissza
Összes vissza
Kilépés


Lexikális tulajdonságok
élőlény
nem igen
személy
nem igen
megszámlálható
bármelyik igen nem
típus
köznév tulajdonnév
földrajzi név
nem igen
intézménynév
nem igen
nak/nek
nem igen

10 Október

Október 21-én megkezdődik a MorphoWord fordítóprogram piaci értékesítése. A MorphoWord Pro változat nyelvi tudása az alapváltozattal azonos, de bővíthető, hálózatban is használható és két felhasználói jogot tartalmaz.





2 Nyelvészeti fejlesztések

Az idei évben a nyelvészeti munka elsősorban a magyar-angol fejlesztésekre irányult.2005 elején indult el a magyar-angol fordítóprogram projekt az MTA Nyelvtudományi Intézetével és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjával közös együttműködésben A fejlesztésekre az NKFP-2/008/2004 pályázat biztosítja a megfelelő kereteket. Az év eredménye, hogy meglévő eszközeinket és erőforrásainkat a projektbe importáltuk, és jelentős lépéseket tettünk abba az irányba, hogy ezekből egy valóban jól működő, robosztus fordítót hozzunk létre. A munkamegosztásban a magnyelvtan és a szabályok felcímkézését végző konverter program fejlesztése, valamint a morfológia karbantartása maradt a MorphoLogicnál, az igei vonzatkeretek fordítása és a működtetéséhez szükséges felcímkézése a Nyelvtudományi Intézet, a névszói szerkezetek fordítása és szemantikai jegyekkel történő ellátása az SZTE feladata lett.

A pályázat a következő táblázatban összefoglalt két munkaszakaszt zárta le ebben az évben:

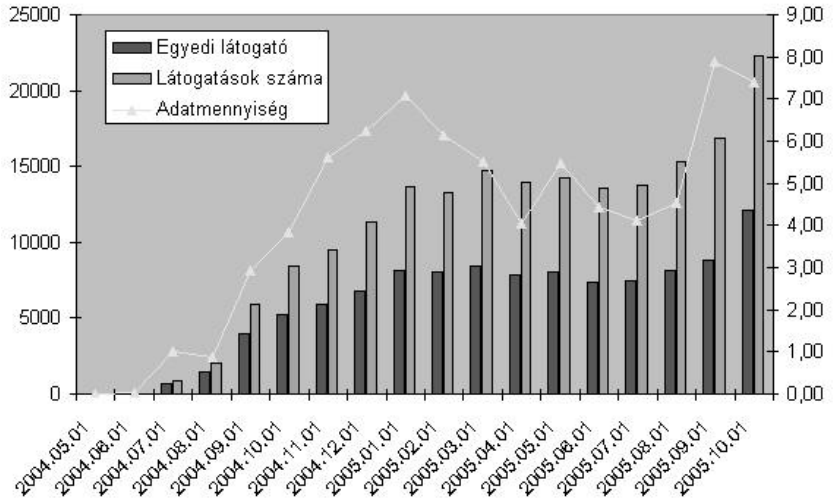
A feladat megnevezése	Közreműködő szervezetek
Tervezés (2005.01.01 – 2005.04.30)	
A rendszer architektúra specifikációja	ML, SZTE, NYT
A nyelvtan reprezentációjának specifikációja	ML, SZTE, NYT
A rendszer kiértékelési módszertanának kidolgozása	SZTE, NYT
A rendszer monotonitásával kapcsolatos kutatás	ML
Az internetes frissítések lehetőségének kidolgozása	ML
A munkaszervezés módszerének kidolgozása	ML
A tesztterv kidolgozása	SZTE
Magnyelvtan megírása (2005.05. 01. – 2005.10.31.)	
A szabályok lekódolása	ML
A nyelvtan pontosságának kimérése	ML, SZTE
Jegyzőkönyv és ennek alapján a szabályok kiegészítése	NYT
Újabb mérés	SZTE

A magyar-angol program nyelvészeti munkálataira ez a dolgozat nem tér ki, erről Merényi Csaba ugyanebben a kötetben található cikkében lehet olvasni. [3].

Az angol-magyar fejlesztések a MoBiCAT és a MorphoWord felhasználóinak visszajelzései alapján folytak.

3 A fordítóprogram első évének értékelése

A MoBiCAT-szolgáltatás első évét a felhasználók számának függvényében és a használati szokásokat kutató kérdőívre adott válaszok alapján értékelhetjük. A MoBiCAT ma egy szűkebb, de a visszajelzések alapján elégedett felhasználói körrel rendelkezik. A kezdeti felfutás az idei év közben stabilizálódott. Az év végén megjelenő MorphoWord a MoBiCAT-eladásokra pozitív hatást gyakorolt:



4. Tervek röviden

Jövőre el akarjuk érni a magyar szövegeken általában érthető angol fordítást produkáló nyelvészeti leírást. Javítani szeretnénk az angol magyar fordítások nyelvi minőségét az elő- és utófeldolgozás lehetőségének kialakításával (ismeretlen szavak, terminológiák) Ezt szolgálná a tervezett angol–magyar jelentésegyértelműsítő adatbázis bővítése a leggyakoribb többértelmű angol szavakkal. Szeretnénk integrálni a fordítóprogramot fordítómemóriába. Továbbfejlesztjük a nyelvi hasonlóság elvén működő intelligens fordítómemóriát, és ezt is fordítómemória-keretprogramba integráljuk.

Referenciák

1. Tihanyi László: A MetaMorpho projekt története. *I. MSzNy, Szeged* (2003)
2. Tihanyi László: A MetaMorpho projekt 2004-ben. *II. MSzNy, Szeged* (2004)
3. Merényi Csaba: A MetaMorpho magyar-angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan *III. MSzNy, Szeged (2005)*

A MetaMorpho magyar-angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan

Merényi Csaba

MorphoLogic kft., 1126 Budapest, Orbánhegyi út 5
merenyi@morphologic.hu

Kivonat: A Morphologic kft., a Szegedi Tudományegyetem és az MTA Nyelvtudományi Intézete által közösen fejlesztett magyar-angol gépi fordító rendszer nyelvtanának az igei vonzatkeretek kezelését végző részét mutatjuk be. A magyar nyelv a környezetfüggetlen nyelvtanok számára általában nehezen kezelhetőnek tartott jelenségek többségét mutatja, ilyenek a szabad szórend, a megszakított összetevők, vagy az üres kategóriák. A vonzatkeretek azonosításához ezekkel a problémákkal mindenképpen meg kell birkózni. A MorphoLogic MetaMorpho rendszerének különleges eszközeivel, és az alkalmazott kiterjesztett argumentum modell segítségével mindezen problémákra megoldást kínálunk.

1 Bevezetés

Jelen cikkben a MorphoLogic kft., a Szegedi Tudományegyetem és az MTA Nyelvtudományi Intézete által közösen fejlesztett magyar-angol gépi fordító rendszer jelenlegi állásáról számolunk be. Az eddigi felesztés során elsődleges feladatunknak azt tekintettük, hogy a Nyeltudományi Intézetben készülő igeivonzatkeret-leírások működőképessé tételéhez megteremtsük a technikai hátteret, illetve kidolgozzuk a kezelhetőnek tekintett szabálytípusokat. Ennek során elkészült a MorphoLogic MetaMorpho rendszerében egy mondatnyelvtan és egy szabálykonverter, melyek már alkalmasak arra, hogy a lefordított vonzatkereteket egyszerű főnévi csoportokat tartalmazó mondatokban működtessék.

Az alábbiakban röviden összefoglaljuk a MetaMorpho rendszer alapjait, majd rátérünk az igeivonzatkeret-leírások bemutatására, és betekintést nyújtunk abba, hogy a rendszer magnyelvtana hogyan képes ezeket az egyszerű, tömör, magasszintű leírásokat felhasználva az adott vonzatkeret tényleges megjelenései közül gyakorlatilag bármelyiket felismerni és lefordítani.

2 A MetaMorpho rendszer általános bemutatása

Ebben a fejezetben a MetaMorpho rendszer alapjait mutatjuk be a teljesség igénye nélkül, csupán tájékoztató jelleggel. A leírás célja, hogy a cikk következő fejezeteiben a rendszer tulajdonságaira tett utalások érthetőek legyenek az olvasó számára. A MetaMorpho rendszerről általában szólnak, illetve egyes részleteket illetően pontosabb leírásokat tartalmaznak [1], [2] és [3].

2.1 mmo: az alacsony szintű alapformalizmus

A MorphoLogic MetaMorpho rendszerének alapja az *mmo* formalizmus. Egy *mmo* szabály egy ún. *elemzősorból* és egy vagy több ún. *generálósorból* áll. Egy MetaMorpho nyelvtan szabályainak elemzősorai együttesen egy a forrásnyelvet leíró környezetfüggetlen frázisstruktúra-nyelvtant alkotnak, míg a generálósorok az elemzősorban leírt nyelvi egységnek megfelelő célnyelvi alakot vagy alakokat adják meg. Álljon itt egy leegyszerűsített példa, mely csak a bemutatást szolgálja:

```
*NP:123456789-01
HU.NP[Head<-NM] = DET(lex="a") + NM(type!=PROPER)
EN.NP = DET[dettype=DEF] + NM[NP.num]
```

A fenti szabály legelső sora csupán egy szabályazonosító. A második sor az elemzősor, melyben egy determinánst és NM nevű kategóriát vonunk össze NP-vé. A harmadik sor a generálósor, melyben azt írjuk le, hogy az adott NP-t hogyan kell a fordítás előállítása során kifejteni. A formlizmus alapvetően kétfajta kategóriát kezel: terminális és nemterminális szimbólumokat. Magától értetődő módon az elemzősorban a terminális szimbólumok közvetlenül a szintaktikai modul inputját képező tokeneknek felelnek meg, míg a nemterminálisok a más szabályok által létrehozott egységekre utalhatnak. A generálósorban szabadon csak terminális szimbólumok szerepelhetnek, ezek mellett csak olyan nemterminálisok állhatnak, melyek az elemzősorban előfordultak[24]. A terminálisok generálása már a szintaktikai modul után következő programelemek feladata, míg a nemterminálisok kifejtése az elemzősorbeli párjukat létrehozó szabály generálósorai által történik.

Mint az a példából is kiderül, az egyes szimbólumok különféle tulajdonságokkal, pontosabban jegyekkel rendelkezhetnek, ennyiben több a rendszer egy egyszerű környezetfüggetlen nyelvtannál. Ezek a jegyek tetszőleges nyelvi (lexikális, morfológiai, szintaktikai), vagy pusztán a rendszer belső működése szempontjából fontos technikai jellegű információt hordozhatnak. Jegyeink három típus valamelyikébe tartoznak. A szimbolikus jegyek egy előre definiált véges értékkészletből vehetnek fel egy értéket. Ilyen a fenti példában az NM *type* jegye, melyet most a PROPER értékkel hasonlítottunk össze. A sztring típus értelemszerűen egy karakterlánc tárolására alkalmas, példánkban a DET *lex* jegye ilyen. A harmadik, redkívüli jelentőségű jegytípus a poin-

[24] Nem teljesen igaz az, hogy minden nemterminálisnak van megfelelője az elemző sorban. Lásd később az ún. Brahma-szabályokat, valamint a pointerből való generálás lehetőségét.

ter, amely részelemzések hatékony tárolására alkalmas. Fent az NP *Head* nevű pointer típusú jegyében tároltuk azt a részfát, melyre az NM nevű szimbólum utal.

Bármelyik típusba tartozó tulajdonságok szabadon beállíthatók, örökölhetők, és tetszőleges kikötést lehet rájuk tenni. A pointer jegyben tárolt szimbólum tulajdonságaira is lehet hivatkozni. A rendszer legújabb változatában a szimbolikus értékekből szabadon definiálható operátorok segítségével kiszámított értéket is tovább lehet adni. A jegyérték-ellenőrzésnek is van egy új módja, mely nem egyszerűen egyenlőséget vagy nemegyenlőséget vizsgál, hanem szintén szabadon definiálható módon egyfajta „kompatibilitást", amely különösképpen mondatrészek közötti egyeztetések, illetve szemantikai megötések ellenőrzésekor hasznos.

A pointerek továbbadhatósága és a generálás során a pointerből való létrehozás lehetősége együttesen megoldják azt a problémát, hogy bizonyos összetevőket ne azon a ponton generáljunk, ahol az elemzésbe bekerültek. Ilyen „mozgatások" segítségével továbbra is jól strukturált, könnyen karbantartható nyelvtant írhatunk olyan nyelvi jelenségek kezelésére is, melyeknél a magyar illetve az angol szórend eltérő, sőt valamelyik nyelvben az összetevő egyes részei akár egymástól távol is megjelenhetnek.

2.2 mmc: Brahma szabályok

Korábban említettük, hogy a generálósorban csak olyan nemterminális szimbólumok szerepelhetnek, amelyeknek van az elemzősorban párja. Ez alól a szabály alól két kivétel is van. Az egyik az a fent említett lehetőség, hogy egy nemterminális szimbólumot egy pointerben továbbadott (másik szabályban szereplő) elemzősori szimbólumból hozunk létre. A másik lehetőség első hallásra ennél jóval meglepőbb: a semmiből is előteremthetjük. Ha a rendszer egy olyan nemterminálist talál generáló oldalon, melynek nincs párja az elemzősorban, akkor mielőtt hibát jelezne, egy különleges szabályhalmazban, az ún. *Brahma fájlban* megnézi, hogy a nyelvtanírók nem rendelkeztek-e az adott kategória elemzősori megfelelő nélküli létrehozásának miként jéről. Ha az adott kategória szerepel ebben a fájlban, akkor innen folytatódik a generálás; a csak generálósorokból álló *Brahma szabály* a jegyértékek alapján hozza létre a megfelelő alakot. Ezek a speciális elemzősor nélküli *mmo* szabályok másnéven az *mmc* szabályok.

Az *mmc* szabályok bementetét képező jegyértékek között természetesen szerepelhetnek pointerek is, így a Brahma szabályok nem minden esetben a „semmiből teremtenek", néha csak bonyolult feltételrendszerek többszöri leírását spóroljuk meg velük. Egy pointerben tárolt szimbólum különböző körülmények közötti más-más formában való generálását több *mmo* szabályból is kényelmesen mintegy „szubrutinszerűen meghívhatjuk" így. Akárcsak a programozásnál, a többször előforduló összetett feladatokat célszerű ilyen jól karbantartható egységes módon megoldani.

2.3 mmd: a könnyen kezelhető magasszintű leírás

A MetaMorpho rendszer egyik nagy erényének tartjuk az egységes leírási módot. Legyen szó mondatszerkezeti szabályról, lexikálisan csak részben specifikált igei mintáról, vagy a legegyszerűbb szótári tételről, gyakorlatilag minden nyelvi informá-

ció *mmo* formalizmusban van tárolva. Ugyanakkor az *mmo* olyan alacson szintű leírás, amelynek minden technikai részletet tartalmaznia kell. Különösképpen az adott esetben rendkívül nagy számban előforduló jegyértéköröklések teljesen olvashatatlanná tehetik a rendszer számára használható *mmo* szabályt. A lexikális erőforrások hatékony előállítása érdekében kidolgoztuk azt a technológiát, amely lehetővé teszi, hogy a szabályírók egy egyszerűsített formalizmusban dolgozhassanak, és csak a nem redundáns információt kelljen látniuk, illetve szerkeszteniük. Ezt az egyszerűsített formalizmust nevezzük *mmd*-nek. Az *mmd* szabályokból egy konverterprogram állítja elő a megfelelő *mmo* szabályt (vagy szabályokat) a rendszer igényeinek megfelelően. A konverzió a programba előre bekódolt algoritmusok alapján történik, de nem teljesen rugalmatlanul, ugyanis az *mmd* formalizmus, szemben az *mmo*-val, tartalmaz ún. metajegyeket is, amelyek segítségével a szabályírók magasszintű utasításokkal befolyásolhatják, hogy pontosan milyen szabályok jöjjenek létre. Jó példa erre az igei vonzatkeretek leírásánál használt :*opt* metajegy, amely azt mondja meg, hogy egy adott vonzat opcionális-e. Ha igen, akkor a konverter két *mmo* szabályt is előállít az egy *mmd* szabályból: az egyikben szerepel az adott vonzat, míg a másikból kimarad.

3 Magyar-angol vonzatkeret-leírások

Az igei vonzatkeretek leírásánál természetesen a magasszintű *mmd* nyelvet használjuk. Pillanatnyilag a nyelvtan több mint 17000 VP-s szabályt tartalmaz, és a hiányok pótlása, valamint a lexikális és szemantikai megkötések alapján történő további finomítás után ennek többszöröse is lehet a végleges szám. Ilyen nagy munkánál mindenképpen törekedni kell arra, hogy a leírás a lehető legegyszerűbb és legtömörebb legyen.

Magyar nyelvű mondatok elemzésekor a vonzatkeretek azonosítását számos a nyelvre jellemző jelenség nehezíti meg. Első helyen említendő ezek közül a szabad szórend. Figyelembe véve a topikalizáció és a fókuszba emelés lehetőségeit, a vonzatok gyakorlatilag bármilyen sorrendben előfordulhatnak, illetve az ige és a szabad módosítók ezek között kevés megkötéssel bárhol elhelyezkedhetnek. További gondot okoz az, hogy bizonyos vonzatok nem egy megszakítatlan összetevőként jelennek meg a mondatban. A birtokos szerkezetekből kiemelhető a birtokos, 'hogy'-os alárendelt mondatok extrapozíció után elválhatnak az utalószótól, a vonatkozó mellékmondatok szintén gyakran az igei csoport végére mozgatva jelennek meg. A vonzatkeret azonosítása és a fordítás egyaránt megkövetelik, hogy ezeket a távoli összetevőket egymáshoz rendeljük és egy egységként kezeljük. Végül megoldást kellett találni arra is, hogy egyes névmási vonzatok elhagyhatók, pontosabban a felszínen nem jelennek meg, de a vonzatkeretek azonosításához, illetve fordításukhoz fel kell tételeznünk jelenlétüket.

Az általunk alkalmazott leírás a fenti bonyodalmak ellenére az egyszerűség, tömörség és kezelhetőség követelményeinek jól megfelel. A szabályírónak gyakorlatilag csak annyi a feladata, hogy egy semleges magyar szórendben felsorolja a vonzatokat és az igét, és azok azonosító jegyeit (pl. eset, névutó, lexikális alak) leírja. A vonzatok között esetleg fennálló kapcsolatokra magasszintű meta-jegyekkel tehetnek megkötést. A fordítást egy semleges szórendű angol vonzatkeret formájában kell megadni, melynek elemei a magyar vonzatoknak felelnek meg. A használható szimbólumok a magnyelvtan által egyébként nem használt, intuitív kategóriákba tartoz-

nak: SUBJ, OBJ, COMPL. Példaként tekintsük a „vmi beleharap vmibe” vonzatkeret leírását:

```
*VP=bele|harap:7
HU.VP = SUBJ(animate=YES) + TV(:lex="bele|harap") +
COMPL#1(pos=N, case=ILL)
EN.VP = SUBJ + TV[lex="bite"] + COMPL#1[prep="into"]
```

Ennek az egy mmd szabálynak a segítségével a rendszer képes lefordítani az alábbi mondatok mindegyikét (ezek valódi példák a működő programból):

Moose>a mókus beleharapott a körtébe. (2)
1: [the squirrel bit into the pear.]

Moose>beleharapott a mókus a körtébe? (3)
1: [did the squirrel bite into the pear?]

Moose>a körtébe a mókus harapott bele. (4)
1: [the squirrel bit into the pear.]

Moose>beleharaptam a körtébe. (5)
1: [I bit into the pear.]

Moose>beleharaptak? (6)
1: [did they bite into it?]

Moose>belém harapott a mókus. (7)
1: [the squirrel bit into me.]

Moose>a mókusnak ki harapott bele a körtéjébe? (8)
1: [who bit into the squirrel's pear?]

Amint azt (1)-(3) mutatják, az ige, az igekötő és a vonzatok tényleges megjelenési sorrendjétől függetlenül azonosítani tudjuk a vonzatkeretet. (4) és (5) azt illusztrálják, hogy egy vagy több vonzat is megvalósulhat zéró felszíni alakú névmásként. Ezeket az eseteket is kezelni tudjuk a fenti *mmd* szabállyal, annak ellenére, hogy abban mind az alany, mind az *ILLATIVUS* esetű vonzat kötelező vonzatként szerepelnek. (6) a vonzatkeret „ragozott igekötős” változatát mutatja be, amelynek létezése kikövetkeztethető a „vmi beleharap vmibe” minta létezéséből, és amelyet ezért a szabályíróknak nem kell külön lekódolniuk. Végül (7) egy olyan esetet mutat be, amelyben az *mmd* szabályban egy COMPL#1 nevű szimbólummal jelölt vonzatot valójában két különböző helyen megjelenő nyelvi egység – egy elmozgatott birtokos és a birtok – alkotja.

A következő fejezetben betekintést adunk annak technikai részleteibe, hogy az egyszerű *mmd* leírásból a konverter által előállított jóval bonyolultabb *mmo* szabályok hogyan tudják a magnyelvtan segítségével gazdaságosan és áttekinthető módon lefedni a példákban bemutatott jelenségeket.

4 Magnyelvtan

4.1 A Kiterjesztett Argumentum fogalma

Az ige vonzatai a felszínen sok különböző formában jelenhetnek meg, de sok szempontból azonos módon kell őket kezelni. Ezért célszerű volt mindenféle vonzatot egy közös kategória, az ARG alá rendelni. Azonban egy környezetfüggetlen nyelvtanban egy egyszerű nemterminális kategória önmagában nem elegendő nem összefüggő vagy a felszínen nem megjelenő összetevők ábrázolásához. Emiatt a magnyelvtanban a vonzatokat egy absztrakt kategóriának tekintjük, ezt nevezzük *Kiterjesztett Argumentum*nak vagy röviden *xARG*-nak. A kiterjesztett argumentum tényleges ábrázolása nem egy nemterminális szimbólum, hanem egy jegyhalmaz, amely a teljes vonzatkeretet reprezentáló nemterminális szimbólum (a VPP) jegyei között szerepel.

Az xARG-ot számos jegy alkotja, amelyek a vonzat lexikális, morfológiai és szintaktikai tulajdonságait kódolják, valamint több pointer típusú jegyből is áll. Ezek a pointerek tárolhatják a fejet tartalmazó ARG összetevőt (ha volt ilyen a felszínen), valamint a kimozgatott elemeket, melyek egységesen az ARG-hoz hasonló közös kategória, a LINK alatt jelennek meg. LINK lehet például az NP birtokosa, vagy az NP-t módosító vonatkozó mellékmondat. A pointerek szerepe az elemzés során gyakorlatilag elhanyagolható (néha azokon keresztül hivatkozunk olyan tulajdonságokra, melyek nincsenek kivezetve az xARG-ba), a vonzatot alapvetően az egyéb jegyértékek reprezentálják. Generálásnál viszont, amennyiben nem zéró névmásról volt szó, az ARG és a LINK pointerek segítségével állíthatjuk elő a fordítást.

A kiterjesztett argumentumok véglegesen csak a teljes vonzatkerettel együtt állnak össze. A vonzatkereteket reprezentáló VPP kategória egy rekurzív szabályhalmaz segítségével jön létre, amely az ige előtt és után álló minden mondatrészt egyenként bekebelez, amíg azok el nem fogynak. Az egyes beelemzett mondatrészeket és tulajdonságaikat ezek a rekurzív VPP szabályok a megfelelő xARG jegyekben helyezik el. A VPP öt xARG számára tartalmaz jegyeket, amelyeket a vonzatok a beelemzés sorrendjében foglalnak el. Ezekre a jegyhalmazokra a továbbiakban xARG1...xARG5 néven fogunk utalni.

A konverter által létrehozott *mmo* szabályok tehát valójában nem maguk vonják össze az *mmd*-ben leírt mondatrészeket egy VP-vé, hanem a magnyelvtan fent vázolt mechanizmusa által már összerakott VPP-k fölötti szűrőként működnek olymódon, hogy azok xARG-jaira tesznek megkötéseket. A továbbiakban azt fogjuk felvázolni, hogy pontosabban hogyan működnek ezek a szűrők, illetve milyen szűrőkre van szükség a szabad szórend, a zéró névmások, illetve más korábban említett jelenségek kezeléséhez.

4.2 A szabad szórend kezelése

A fent elmondottak szerint a VPP az argumentumokat az xARG1 ... xARG5 jegyhalmazokban olyan sorrendben tartalmazza, amilyen sorrendben azokat beelemezte. Az igei vonzatkeretet azonosító szűrők két kölünböző módon ismerhetik fel az adott igei mintát tetszőleges szórenddel. A legkézenfekvőbb eljárás az, hogy minden lehet-

séges sorrendhez előállít a konverter egy külön *mmo* szabályt. Pl. a korábban bemutatott „vmi beleharap vmibe” vonzatkerethez lehetne két *mmo*-t generálni, melyek elemző sorai vázlatosan így nézhetnének ki:

```
HU.VP = VPP(lex=”harap”, ik=”bele”, argnum=ARG2, Arg1-
>case=NOM, Arg2->case=ILL, ...)

HU.VP = VPP(lex=”harap”, ik=”bele”, argnum=ARG2, Arg1-
>case=ILL, Arg2->case=NOM, ...)
```

A fejlesztés kezdeti szakaszában ténylegesen ezt az utat választottuk, mert a C++ nyelven írt konverter program egyszerűbben elő tudta állítani a vonzatok összes permutációját. Ennek a megközelítésnek azonban az a hátránya, hogy feleslegesen megsokszorozza az előállítandó *mmo* szabályok számát. Mivel egyéb okoknál fogva is szükséges lehet a szabályok többszörözése, a létrehozott nyelvtan mérete akár a rendszer teljesítményét veszélyeztető módon is felduzzadhatott.

Szerencsére az operátorokkal kiterjesztett *mmo* formalizmus elég erős eszköznek bizonyult ahhoz, hogy a magnyelvtanban hatékonyan megoldjuk az összegyűjtött xARG-ok permutálását. Így akármilyen sorrendben is szerepeljenek eredetileg a vonzatok, az összes lehetséges permutációt tartalmazó VPP-k előállítása után már egyetlen szűrő is elgendő a vonzatkeret azonosításához. A felesleges VPP-k nem terhelik a rendszert, mert bár ezen az elemzési ponton kissé felszaporodik a létrejött elemzési tények száma, azok közül csak egynek lesznek felmenői, amelyek tovább kombinálódhatnak más részelemzésekkel. Így kombinatorikus robbanástól nem kell tartanunk.

4.3 A zéró névmások kezelése

A zéró névmások kezelése pillanatnyilag a konverter segítségével történik. Az *mmd* leírás szerint alanynak vagy tárgynak minősülő vonzatokról tudja a konverter, hogy azok a felszínen esetleg nem jelennek meg, ezért olyan szűrőket is generál, amelyek ezeket a vonzatokat nem kérik számon a VPP-n:

```
HU.VP = VPP(lex=”harap”, ik=”bele”, argnum=ARG2, Arg1-
>case=NOM, Arg2->case=ILL, ...)

HU.VP[xs=YES, ...] = VPP(lex=”harap”, ik=”bele”,
argnum=ARG1, Arg1->case=ILL, ...)
```

A VP-nek a vonzatokra vonatkozó jegyeit viszont úgy tölti ki a második szabály, hogy abban jelzi a zéró névmás jelenlétét, gyakorlatilag maga a szűrő létrehozza a megfelelő xARG-ot, és jegyeinek értékét többek között az ige ragozásából következteti ki. A generálósorban az ilyen vonzatot nem tudjuk a megfelelő xARG pointer típusú jegyeiből létrehozni, hiszen azok nem tartalmaznak semmit, ehelyett a 2.2 alatt említett Brahma szabályok segítségével generáljuk őket. A Brahma generáláshoz szükséges jegyértékeket a konverter tölti ki a konkrét szabály tulajdonságaitól függően. Például az élőre vagy élettelenre utaló névmás generálása közötti választást befo-

lyásolja az, hogy az mmd-ben volt-e, illetve milyen megkötés volt a megfelelő vonzat *human* jegyére.

4.4 A megszakított összetevők kezelése

A megszakított összetevők kezelésében a vonzatkeretet azonosító szűrőknek kevés szerep jut. Ezek a vonzat xARG-gal való modellezésének köszönhetően tulajdonképpen maguktól működnek. A VPP összeállítása során alkalmazott bonyolult, sok nyelvi tudást hordozó szabályok a szűrőknek átadott vonzatkeret-jelöltek xARG-jaiban már minden nyelvtanilag lehetséges módon egymáshoz rendelték az ARG-okat és a különféle LINK-eket. Így pédául az előző fejezetben leírt (7)-ben az *[a mókusnak]* távoli birtokos LINK-et és az *[a körtéjébe]* birtokos személyragot tartalmazó ARG-ot a személyrag meglétéből kiindulva, majd a szám és személy szerinti sikeres egyeztetés után, egyazon xARG-ban tárolva küldjük a szűrőknek, melyek csak közvetve szólhatnak bele a megszakított összetevők felismerésébe. Amennyiben több lehetséges egymáshoz rendelés is létezik, a szűrők az egyes vonzatokra, vagy azoknak például éppen a birtokosára tett megkötéseikkel egyértelműsíthetik az elemzést.

Az xARG generálásakor minden LINK-jét átadjuk neki, melyeket azok típusától függően saját szerkezetén belül a megfelelő helyre fog elhelyezni. Az adott példánál maradva, minden célnyelvi NP kategóriájú argumentum számít arra a lehetőségre, hogy kap felülről egy birtokos LINK-et, amelynek felhasználásával birtokos szerkezetet generál. Így lesz a magyar mondatban egymástól elválasztva megjelenő *[a mókusnak]* és *[a körtéje]* kifejezésekből az angolban egy NP: *[$_{ARG}$[$_{LINK}$the squirrel]'s pear].*

5 Összegzés

Az előző fejezetekben röviden vázoltuk, hogy a MorphoLogic MetaMorpho rendszere elsősorban a pointer típusú jegyek, a Brahma szabályok és a Kiterjesztett Argumentum modell segítségével hogyan tudja kezelni a környezetfüggetlen nyelvtanok számára általában nehézséget jelentő szabad szórendet, zéró elemeket, valamint a megszakított összetevőket. A fejlesztés alatt álló magyar-angol gépifordítórendszerünkben ezeket a technológiákat a gyakorlatban is sikerrel alkalmazzuk.

Bibliográfia

1. Prószéky, Gábor & Tihanyi, László: MetaMorpho: A Pattern-based Machine Translation Project. In: Proceedings of the 24th 'Translating and the Computer' Conference. London, United Kingdom, 19–24 (2002)
2. Tihanyi László: A MetaMorpho projekt története. Magyar Számítógépes Nyelvészeti Konferencia 2003, Szeged.
3. Gábor Prószéky, László Tihanyi, Gábor Ugray: Moose: A Robust High-Performance Parser and Generator. EAMT Workshop, Malta, 2004

Fordítómemóriák és minta alapú fordítórendszerek kiértékelésének módszerei

Hodász Gábor

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
1083 Budapest, Práter u. 50/a.
hodasz@itk.ppke.hu

Kivonat: Cikkünkben áttekintést adunk a fordítómemóriák és a hasonló működésű minta alapú fordítórendszerek kiértékelésének kérdéseiről. Megmutatjuk a lényegi különbségeket, amelyek megkülönböztetik ezeket a módszereket a gépi fordítás kiértékelésének már jól kidolgozott módszereitől. Kitérünk a tanító- és tesztkorpusz tulajdonságaira, valamint a szinkronizáló és a hasonlósági kereső modul önálló kiértékelésére. Bemutatjuk mind az automatikus mind az emberi közreműködésű metódusokat. Az egyes módszereket értékeljük, és javaslatot teszünk egy saját kiértékelő módszerre, amely figyelembe veszi a tanító és tesztkorpusz tulajdonságait és a rendszer célját: a fordítási munka minél hatékonyabb támogatását.

1 Bevezetés

A gépi fordítás irodalma kimerítően taglalja a fordítórendszerek értékelésének módszereit és a fordítás minőségének mérőszámait. Számtalan mérőszám és metódus született az elmúlt két évtizedben, hiszen a gépi fordítás a kezdetektől a tudományos érdeklődés fókuszában volt. Korántsem ilyen kidolgozott a fordítómemóriák és a hasonló módon működő minta alapú gépi fordítórendszerek (Example Based Machine Translation, EBMT) kiértékelésének kérdése [8].

A gépi fordítás esetében a kiértékelés célja jellemzően az, hogy meghatározzuk, hogy a rendszer által adott fordítás „milyen távol van" a mintának tekintett „tökéletes" emberi fordítástól. Ezzel szemben a memória alapú rendszerek esetében nem a rendszer konkrét válasza a mérvadó, hanem sokkal inkább a memóriában levő minták felhasználásának hatékonysága és a felkínált fordítás „hasznossága" a fordítói munka szempontjából. A fordítómemória célja nem a tökéletes fordítás előállítása, sokkal inkább a profi fordító minél hatékonyabb segítése a rendelkezésre álló (eltárolt) korábbi fordítások újrahasznosításával. Ezen megközelítésből következik az is, hogy a kiértékelés támaszkodhat az emberi fordító interaktivitására is, azaz megvalósítható a működés közbeni, mondatról mondatra való emberi kiértékelés is.

Cikkünkben áttekintést adunk a legfontosabb módszerekről, előnyeikről és hátrányaikról, majd bemutatjuk a saját módszereinket a MetaMorpho TM rendszer kiérté-

kelésére. Külön tárgyaljuk az egyes modulok, valamint az egész rendszer kiértékelését, bemutatva mind az automatikus, mind az emberi közreműködésű metódusokat.

2 Kiértékelési stratégiák

A memória-alapú fordítórendszerek kiértékelésének szemlélete erősen függ attól, hogy ki és milyen célra végzi a kiértékelést. A rendszer viselkedéséről, eredményeiről tanúskodó számok érdekesek nemcsak a rendszert használó fordítónak, a beszerzésről döntő vezetőnek, a fordítást megrendelő ügyfélnek, de a rendszert fejlesztőknek is.

2.1 Kiértékelési módszerek osztályozása

Két alapvető stratégia létezik a rendszer szemléletét illetően:

1. fekete doboz (black box): a rendszert csupán a bemenet-kimenet összefüggésében vizsgáljuk, a közbenső lépések, modulok rejtve vannak a vizsgáló elől. Elsősorban a felhasználók, fordítók számára megfelelő kiértékelési módszer.
2. üvegdoboz (glass box): a rendszert felépítő modulok működését vizsgáljuk, valamint hatásaikat az egész rendszer működésére. Fejlesztők és kutatók számára hasznos kiértékelési módszer.

Cikkünkben mindkét megközelítést alkalmazzuk: bemutatjuk a fejlesztői szemléletű-, valamint a teljes rendszer felhasználói szemléletű kiértékelésének lehetőségét is.

Az alábbiak szerint különböztethetjük meg a kiértékelési módszereket:

1. automatikus (gépi, objektív): rövid idő alatt nagy mennyiségű szövegen elvégezhető metódus, amelynek „objektivitása” mindazonáltal függ a definiált mérőszámoktól.
2. manuális (kézi, szubjektív): hosszabb idő alatt, kisebb korpuszon elvégezhető módszer, erősen függ a kiértékelő személytől, de közelebb áll a tényleges felhasználási körülményekhez.

Jelen cikkben mindkétféle metódusra teszünk javaslatokat.

A kiértékelés kiterjedhet számtalan, a fejlesztés vagy a felhasználás szempontjából lényegesnek ítélt elem meglétének, vagy működésének vizsgálatára. Ilyenek lehetnek a felhasználói felület használhatósága, az import/export funkciók, TMX támogatás, statisztikai modul stb. Ennek a cikknek nem témája a szorosan vett memória-funkciókon (tárolás és keresés) kívüli szolgáltatások vizsgálata.

3 A MetaMorpho TM rendszer

A MorphoTM rendszer olyan fordítástámogató eszköz, amelynek célja, hogy a hagyományos fordítómemória-funkciókat nyelvi intelligenciával kiegészítve a jelenlegi rendszereknél többször ajánljon fordítást, és azok jobban közelítsék a kívánt minőségű fordítást [3]. A fordítás egységei a mondatnál kisebb szegmensek (főnévi szerkezetek és az ezeket tartalmazó mondatvázak), amelyeket a forrás- és célnyelvi elemzők állítanak elő. Az adott bemeneti mondathoz hasonló szegmenseket „nyelvi

intelligencián" alapuló távolság segítségével keressük, és a megszülető új fordításokat nyelvi elemzéssel együtt tároljuk.
A nyelvi elemzés lépései:

1. a szegmensek morfológiai elemzése mind a forrás-, mind a célnyelvi oldalon
2. a forrásnyelvi oldalon sekély szintaktikai elemzéssel előállított főnévi csoportok (noun phrases, NPs)
3. a mondat alatti egységeket szinkronizáló modul előállítja a célnyelvi oldalon a forrásnyelvi NP-k párjait (nyelvtani elemzés vagy heurisztika segítségével)

A sikeres elemzés eredménye mindkét oldalon a főnévi csoportok és az ezeket tartalmazó mondatvázak lesznek, az egyes szavak morfológiai elemzéseivel. A főnévi csoportokban jelöljük a fejet, ennek a morfológiai tulajdonságai határozzák meg az egész NP tulajdonságait. Mivel a főnévi csoportok a nyelvi sajátosságok miatt nem minden esetben fordulnak le, vagy nem főnévi csoportra fordulnak, vagy a szinkronizáló nem tudja egyértelműen megfeleltetni őket egymásnak, ezért előfordul, hogy a mondatváz ilyen nem-szinkronizált főnévi szerkezeteket is tartalmaz [6].

A mondatváz az eredeti főnévi csoportok helyett csak ú.n. slotokat tartalmaz, amely jelöli az eredeti NP tulajdonságait, de bármilyen más NP-vel behelyettesíthető, amely megfelel a megszorító jegyeknek. Az NP-slot morfológiai tulajdonságait a célnyelvi mondatban (jelen esetben ez a magyar nyelvű mondat) levő NP határozza meg, mivel a keresés során a célnyelvi mondatba való behelyettesítés a cél.

A új mondat fordítása esetén a forrásnyelvi elemző előállítja a kereső-mondat morfológiai elemzését, a főnévi csoportjait és mondatvázát. A keresés egységei és a felajánlott fordítások az elemzés lépéseiből adódóan a következők lehetnek:

1. teljes mondatok: a nyelvi hasonlóságon alapuló mérték szerint hasonló teljes mondatok a memóriából
2. mozaikok: a keresőmondattal megegyező mondatvázzal rendelkező mondatok, amibe a kereső mondat NP-jeihez hasonló NP-k töltik ki a slotokat a megszorításoknak megfelelően
3. főnévi csoportok: amennyiben nincs a memóriában se hasonló mondat, se hasonló mondatváz, úgy a rendszer a keresőmondat NP-it helyettesíti be hasonló NP-kel a memóriából.

A mozaik fordítások összeállításánál a rendszer figyelembe veszi az NP-slotok tulajdonságait: egyrészt csak az érvényes megszorításoknak megfelelő NP-k helyettesíthetők a slotba, másrészt a mondatvázba beillesztett NP-k felveszik a slot eredeti morfológiai tulajdonságait: morfológiai generátor alakítja megfelelő felszíni alakúvá a beillesztett főnévi csoportot.

Fordítandó mondat:

Microsoft Windows 2000 makes it possible to configure hard disk drives in a variety of ways.

A fordítómemóriában talált mondatváz:

[01] make it possible to configure [02] in a variety of ways.
$[01]_{NOM}$ sokféle módon lehetővé teszi a beállítását $[02]_{DAT}$.

A mondatban levő főnévi csoportok találatai: (1. példa)

Ssz.	*Fordítómemória*	*Tárolt fordítás*
[01]	Microsoft Windows 2000	Microsoft Windows 2000
[02]	hard disk drive	merevlemez

A rendszer által felajánlott fordítás:

[Microsoft Windows 2000] sokféle módon lehetővé teszi a beállítását [merevlemez][ek][nek].

1. példa: mozaik fordítás a memóriában levő mondatváz és főnévi csoportok segítségével

Az 1. példában bemutatunk egy mondatvázból és főnévi csoportokból összerakott fordítás-javaslatot. A memóriában talált mondatváz megegyezik a fordítandó mondat vázával, így a rendszer a keresőmondat NP-jeihez hasonló NP-ket keres a memóriában, amelyeket behelyettesít a mondatvázba a slotok tulajdonságainak megfelelő felszíni alakban. Bár az angol mondatban van még egy főnévi csoport: [in a variety of ways], azonban ez magyarra olyan formában fordul, hogy a szinkronizáló nem jelölte meg párként, így a mondatváz részét fogja képezni.

4 A memória tartalma

A leglényegesebb különbség a gépi fordítás és a memória-alapú fordítástámogatás között az utóbbi függése a memóriában levő szöveg mennyiségétől és minőségétől. Nyilvánvalóan a gépi fordítórendszereknek is lényeges tulajdonsága a szabálybázisuk, vagy a statisztikai alapon feldolgozott korpusz nagysága. Azonban, míg a gépi fordítás esetében az elkészült rendszer részét képezi ez a „tudás”, és a cél a minél tökéletesebb fordítás előállítása, addig a memória-alapú rendszereknek nem része a memória tartalma, a cél pedig az, hogy bármilyen memória-tartalom esetén a lehető leghatékonyabban használjuk fel újra azt.

4.1 A korpusz

A fordítómemóriák kiértékelésének egy elterjedt metódusa, hogy egy adott korpusz bizonyos százalékát tanító korpuszként használva, feltöltjük vele a memóriát, és a korpusz maradék részével teszteljük. A korpusznak párhuzamosnak és szinkronizáltnak kell lennie.

A korpusz összeállításakor két lehetőség közül választhat a kiértékelést végző:

1. speciális széria: gondosan szerkesztett példa-halmaz, amely különböző nyelvi és/vagy fordítói problémát reprezentál, illetve a különböző modulok működésének sajátosságait teszteli. Elsősorban a kézi kiértékeléshez alkalmazható.
2. általános korpusz: megfelelően sokféle valódi szövegből összeállított korpusz, amely sokkal inkább modellezi a valós működést.

A MetaMorpho TM kiértékelésénél mindkét fajta korpuszt alkalmazzuk: egy válogatott mondatokat tartalmazó szériát a modulok kiértékelésére, és egy nagy méretű tanítókorpuszt (a Szak Korpusz egy részét) az automatikus evaluációhoz.

4.2 A korpusz koherenciája

A kiértékeléshez használt korpusz leglényegesebb tulajdonsága a mérete. Azonban van egy olyan tulajdonság is, amelyet a korábbi kiértékeléssel foglalkozó munkák nem taglalnak: a tanításra és tesztelésre használt korpuszok hasonlósága, vagy más szavakkal a szövegek koherenciája. Nyilvánvaló ugyanis, hogy a fordítómemória potenciális hasznossága függ a benne tárolt szöveg és a keresőmondat (tágabban a tesztkorpusz) közötti hasonlóságtól. Amennyiben a memória nem, vagy csak kis számban tartalmaz hasonló szavakat, szókapcsolatokat, mondatokat, úgy a rendszer teljesítménye kézenfekvő módon alacsonyabb lesz, mintha a memória és a tesztkorpusz között nagy az átfedés. Szükség van tehát egy mértékre, amely jellemzi a memória tartalma és a tesztkorpusz viszonyát, hiszen ennek megfelelően lehet kiértékelni a rendszer tényleges válaszát. Az általunk ismert eddigi munkák nem tesznek utalást a kiértékelésnek erre az aspektusára.

A jelenleg forgalomban kapható fordítómemóriáknak van ehhez hasonló célú modulja, amely összehasonlítja a fordítandó szöveget a memória tartalmával, és a kettő hasonlóságát vizsgálja. Ez a megközelítés azonban 3 okból sem felel meg a céljainknak. Egyrészt a hasonlítás alapja a mondat, azaz a fordítandó mondatokhoz keres hasonló mondatokat a memóriában. Ez az eredmény nem árul el semmit a mondatnál kisebb egységek újrahasznosíthatóságát illetően. Másrészt az összehasonlítás eredménye nem egy érték lesz, hanem a különböző hasonlóságú mondatok száma: pl. 100%-os egyezések száma, 90-99%-os egyezések száma stb. Ez a kiértékelés szempontjából nehezen kezelhető eredményt ad. Harmadrészt a hasonlóság meghatározásánál azt a fuzzy hasonlósági mértéket használja, amit kiértékelni szeretnénk. Így nem ad megbízható mértéket a kiértékeléshez.

A következőkben javaslatot teszünk egy olyan mérték definiálására, amely egy súlyozott átlagban ad jellemzést a szövegek koherenciájára, az összehasonlítás alapja nem a mondat, és független a rendszer saját hasonlósági algoritmusától.

4.3 N-gram alapú koherencia mérték

A fordítandó tesztkorpusz és a memória tartalmának hasonlóságát jól jellemezheti a bennük előforduló közös szavak, szó-kettősök, szó-hármasok stb. száma, azaz általánosan a szó n-gramok ismétlődései. Vegyük tehát a tesztkorpusz minden szó-n-esét, és számoljuk meg, hogy hányszor fordul elő a memóriában. Ezt a számot osszuk el a memória szó-n-eseinek számával, így megkapjuk a relatív gyakoriságot. Összegezzük súlyozva ezeket a relatív gyakoriságokat minden n-re. Így egy súlyozott átlag jellegű

mértéket kapunk, ami százalék-jelleggel megmutatja a tesztkorpusz és a memória közötti hasonlóságot.

$$coherence(T_{TC}, T_{DB}) = \sum_n w_n \cdot \frac{\sum_{g_n \in TC} count_{DB}(g_n)}{|g_n|_{DB}} \quad (9)$$

ahol $count_{DB}(g_n)$ egy adott szó-n-es (word n-gram) előfordulásainak száma a memóriában, $g_n \in T_{TC}$ a tesztkorpusz szó-n-esei, $|g_n|_{DB}$ az összes szó-n-esek száma a memóriában, w_n egy n-hez rendelt súlytényező, hogy $\sum w_n = 1$.

A paraméterek értékére a következő javaslattal élünk:

$N=max(n) = 6$, $w_n = \frac{n}{\sum_{i=1}^{N} i}$, így maximum 6 hosszú egyezéseket vizsgálunk, és minél nagyobb az n (azaz minél hosszabb szövegrészlet egyezik), annál nagyobb súllyal számítjuk be a koherencia mértékbe.

A koherencia mértéket több tényező is torzítja:

1. a hosszabb n-esek több rövidebb n-eseket tartalmaznak, amelyeket így többször számolunk. Erre megoldást jelent, ha a számlálást a nagyobb n-ektől kezdjük, és a már megszámlált n-esben levő (vagy átlapolódó) rövidebb n-eseket nem számláljuk.
2. mind a tesztkorpusz, mind a memória nagy számban tartalmaz olyan szavakat, melyek bár gyakran fordulnak elő, mégsem adnak semmilyen támpontot a koherencia vizsgálatban. Ezek a gyakran előforduló, nem önálló jelentéssel bíró szavak (stopword-ök) azonban egyenletes eloszlásúak és adott nyelvre jellemző gyakoriságúak. Így szűrésüket nem látjuk indokoltnak.
3. a ragozó nyelvekben, így a magyarban is, a karakter alapú egyezés nem ad kellőképpen pontos képet a szavak ismétlődéséről, hiszen a különböző toldalékkal ellátott szavak ismétlődése ugyanúgy jelentős a koherencia szempontjából, a számításnál azonban nem lesznek egyezőek. Erre egy lehetséges megoldás, ha kellően nagy szövegen nézzük a koherenciát, így nagyobb az esélye az azonos alakú ismétlődésnek (de különböző nyelvek között így sem összehasonlíthatóak a kapott eredmények). Egy másik megoldás, ha az összehasonlítás során megengedünk a szó végén bizonyos mennyiségű karakteres különbséget. Ez nyilván csak közelíti az ideális megoldást, de megfelelő pontosságot adhat. Mivel a MetaMorpho TM rendszer kiértékelésénél az angol a forrásnyelv, így ez a probléma nem okoz jelentős hibát.

A további kutatásunk célja lesz, hogy megállapítsuk a fentiek szerint definiált koherencia mérték alkalmazhatóságát a rendszer kiértékelésekor kapott eredmények vizsgálatában.

5 A MetaMorpho TM moduljainak kiértékelése

5.1 A főnévi csoport szinkronizáló modul

A főnévi csoport szinkronizáló modul végzi a mondatok főnévi csoport elemzését és párosítását. Kiértékelésére legmegfelelőbb a pontosság/lefedettség alapú módszer, amely százalékosan adja meg a helyesen szinkronizált (pontosság) és a nem szinkronizált (lefedettség) főnévi csoportok arányait az összes főnévi-csoport párhoz képest. A kiértékelést lehet a modul eredményeinek kézi elemzésével, vagy egy korábban kézzel annotált referencia-korpusszal való összehasonlítással végezni. Az eredményekről bővebben lásd [6] munkáját.

A többi modul önálló értékelésénél a kézzel annotált korpuszt használjuk a memóriába töltve, így kiküszöböljük e modul hibáinak halmozódását a többi modul eredményében.

5.2 A hasonlósági kereső modul

A fordító memória lényegi funkcióját végző hasonlósági kereső modul kiértékelésénél szükségesnek érzünk két alapvetést.

Egyrészt nem vesszük számításba a visszaadott találatok számát, csak a legelső találatot értékeljük. A hasonlósági kereső akkor hatékony segítség a fordítónak, ha a lehető leghasznosabb fordítást ajánlja fel, és ezen túl nem kell a fordítónak listákban böngésznie, más lehetőségek után keresve. A minél több találat sokkal inkább akadályozza a gyors munkát, mint segíti azt. Így preferáljuk a kevés, de releváns találatot.

Másrészt kizárjuk az értékelésből a teljes mondat-egyezéses találatokat. Ha egy mondat teljes egészében megtalálható a memóriában, akkor ennek megtalálása triviális megoldás. Ha ezeket is belevennénk a kiértékelésbe (bár a számokat impozánsan felfelé húznák), nem adnának valós képet a minták újrafelhasználásának hatékonyságáról.

A lentebb kifejtett kiértékelő metódusok mindegyikében az általánosan elterjedt módszert alkalmazzuk a tanító- és tesztkorpusz kiválasztását illetően: ugyanannak a szövegnek egy nagyobb részét (jellemzően 9/10-t) használjuk a memória feltöltésére, a maradék részt (1/10-t) pedig a tesztelésre. Ez legtöbbször biztosít egy megfelelő koherenciát a memória és a tesztkorpusz között.

5.3 Kézi kiértékelés

Több korábbi munka foglalkozik a rendszer kimenetének kézi értékelésével. [7] és [1] is 4 fokú skálát használ, amelyek a tökéletes egyezéstől a „semmi haszna sincs” eredményig terjed. Ennek a metódusnak előnye, hogy a valós felhasználást modellezi: a felhasználó közvetlenül a rendszer eredményeinek hasznosságát értékeli. Hátránya, hogy korlátozott korpusz értékelhető ki kézzel, valamint egy (vagy néhány) felhasználó döntése nem elég objektív: egy adott fordítás lehet nagy segítség annak, aki egyáltalán nem tudja lefordítani a mondatot, és gyakorlatilag használhatatlan

annak, aki igényes fordítást ad. Ezen segíthet, ha a felhasználók a fordítás megkezdése előtt értékelik a forrásnyelvi szöveget is: megjelölik azokat a mondatokat, amelyeket nem tudnak részben vagy egészben lefordítani. Ezzel súlyozni lehet az értékelésüket.

Egy másik módszer a szükséges utószerkesztések számának vizsgálata. A felhasználó átalakítja a fordítómemória válaszát az általa elfogadhatónak ítélt fordítássá. A kiértékelő rendszer pedig számlálja az ehhez szükséges lépések számát. Hogy mit tekintünk szerkesztésnek, és milyen szerkesztés hány lépésből áll, az igen különbözővé teheti az ilyen módszerrel végzett kiértékeléseket. Például: formázási módosítások (kis/nagybetű, betűméret, behúzások stb.), sorrendcsere („fogd meg és húzd” módszer hány lépésnek számít?), vagy számít-e a szükséges egérmozgatások száma stb. Evvel a módszerrel bővebben [9] foglalkozik.

5.4 Automatikus kiértékelés

Az automatikus kiértékelés olyan objektívnek tekinthető mérőszám alkalmazását célozza, amely emberi beavatkozás nélkül teszi lehetővé az értékelést. Így a kapott eredmények nem függnek egy adott felhasználó vagy fejlesztő képességeitől vagy véleményétől, és nagyságrendekkel nagyobb korpuszon végezhető el a kiértékelés azonos idő alatt, ezért a statisztikai megbízhatósága is jobbnak tekinthető. Hátránya, hogy ezeknek a módszereknek szükségük van egy referencia-fordításra, amelynek meghatározása nemcsak szubjektív, de adott esetben az emberi felhasználó más fordítást is elfogadna. Ezen kívül a definiált mérték általában sokkal inkább jellemző a mértéket meghatározó céljaira, mint a rendszer tényleges használatának tulajdonságaira.

Egy lehetőség a gépi kiértékelésre az előzőkben leírt utószerkesztés automatikus megvalósítása: a kiértékelő modul kiszámolja a válasz és a referencia-fordítás közötti szerkesztési távolságot, és ez lesz a kiértékelés eredménye. Ennek egy megvalósítása például [2] munkája. A módszer előnye, hogy közel áll a felhasználás-központú kiértékeléshez, hátránya, hogy nem feltétlenül van összefüggés az így definiált szerkesztési távolság és az emberi fordítónak ténylegesen szükséges utószerkesztési munka között.

Egy másik lehetőség a gépi fordításban is elterjedt n-gram alapú BLEU kiértékelés megvalósítása [5], vagy az ehhez igen hasonló NIST index alkalmazása [4]. Az egyszeri kiértékelés helyett azonban pontosabb képet kapunk a rendszer működéséről, ha tízszeres keresztellenőrzéssel számítjuk ki a BLEU indexet: az adott korpusz 9/10-t használjuk tanításra, 1/10-t tesztelésre, és ezt tízszer megismételjük különböző részekkel. A módszer előnye, hogy a gépi fordítás eredményeivel is összevethető eredményt kapunk, amely könnyen automatizálható, azonban ez a módszer áll legtávolabb a felhasználó-központú kiértékeléstől.

6 Konklúzió

Cikkünkben áttekintést adtunk a fordítómemóriák és a példa-alapú fordítórendszerek kiértékelésének lehetőségeiről. Bemutattuk mind a manuális, mind az automatikus kiértékelés több módszerét, ezek előnyeit és hátrányait. Megmutattuk, hogy a memó-

ria alapú rendszerek kiértékelésének célja különbözik a gépi fordító rendszerekétől, hiszen itt a memória tartalma, valamint a memória és a tesztkorpusz viszonya befolyásolja az eredményt. A kiértékelés célja tehát nem az, hogy a konkrét eredmény hasznosságát vizsgáljuk, hanem a memóriában levő korpusz újrafelhasználhatóságának hatékonyságát értékeljük. Ennek érdekében bevezettünk egy súlyozott átlagot, amely jellemzi a memória és a tesztkorpusz koherenciáját. Ez irányadó számként alkalmazható az eredmények vizsgálatánál.

7 További munkák

A fent leírt módszerek megvalósításával a MetaMorpho TM rendszer kiértékelése. Az egyes módszerek összehasonlítása, valamint a fent definiált koherencia mérték vizsgálata az egyes eredményekre.

Bibliográfia

1. Cranias, L., H. Papageorgiou and S. Piperidis: A Matching Technique in Example-Based Machine Translation. Coling (1994), 100–104.
2. Frederking, R., Nirenburg, S.: Three Heads are Better than One. 4th Conference on Applied Natural Language Processing, Stuttgart, Germany, 95–100. (1994)
3. Hodász G., Pohl G.: MetaMorpho TM: a linguistically enriched translation memory. In: International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan) ISBN 954-90906-9-8, Borovets, Bulgaria, (2005)
4. NIST. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/mt2001/resource/ (2002)
5. Papineni, Kishore & Roukos, Salim et al.: BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics, (2002).
6. Pohl, G.: Angol-magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás, Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005)
7. Sato, S. and M. Nagao. 1990. Toward Memory-Based Translation. Coling (1990), Vol. 3, 247–252.
8. Somers, H.: An Overview of EBMT In M. Carl. and A. Way. (eds.) Recent Advances in Example-based Machine Translation, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3—57. (2003)
9. Whyman, E. K., Somers, H. L.: Evaluation Metrics for a Translation Memory System. Software–Practice and Experience 29, 1265–1284. (1999)

Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás

Pohl Gábor

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

Kivonat: A minta-alapú gépi fordítás (EBMT) alapfeltétele, hogy forrásnyelvi és ezeknek megfelelő célnyelvi mintamondatok mondatnál kisebb szerkezeti egységeit automatikusan egymáshoz tudjuk rendelni. Cikkünkben egy EBMT alapú angol-magyar fordítómemóriához (MetaMorpho TM) kidolgozott főnévicsoport-szinkronizáló algoritmust, valamint egy magyar főnévi csoportok angol megfelelőik alapján történő meghatározására kifejlesztett módszert mutatunk be. A főnévi csoportok szinkronizálása során módszerünk tövesített szótári keresést alkalmazva, hasonló alakú szavakat (*cognate*), illetve szófaji egyezéseket keresve minden lehetséges főnévicsoport-párhoz kiszámít egy heurisztikus hasonlósági értéket, majd ez alapján dönt az egyes főnévi csoportok egymáshoz rendeléséről. A szintaktikai elemzővel meghatározott angol főnévi csoportok magyar megfelelőinek meghatározására kidolgozott módszerünk magyar szintaktikai elemzőt nem igényel, az angol főnévi csoportok szavait szótár segítségével képezi le a magyar mondat szavaira, majd a lehetséges fedések közül a magyar mondatra legrövidebben illeszkedőt teljes magyar főnévi csoporttá bővíti (a szótárral meg nem feleltetett szavak szófaját is figyelembe véve a bővítés során). Cikkünkben végül az első szinkronizációs eredményeinket is ismertetjük.

1 Bevezetés

A minta-alapú gépi fordítás (EBMT, [10]) alapfeltétele, hogy a rendelkezésünkre álló forrásnyelvi és az ezeknek megfelelő célnyelvi mintamondatok mondatnál kisebb szerkezeti egységeit automatikusan egymáshoz tudjuk rendelni. Ezt a folyamatot, nevezzük szinkronizációnak vagy párhuzamosításnak[25].

Ebben a cikkben a MorphoLogicnál fejlesztett EBMT alapú, főnévi csoportokat és mondatvázakat kezelni képes MetaMorpho TM fordítómemória rendszerben [2] alkalmazott főnévicsoport-szinkronizáló modul [4] fejlesztése során kidolgozott módszereket, illetve az eddig elért új eredményeket mutatjuk be.

A MetaMorpho TM rendszer a hagyományos, karakteralapú hasonlósági keresést alkalmazó fordítómemóriákkal szemben a Hodász Gábor által kidolgozott nyelvi

[25] Angolul *alignment*, amelyen nem csak a folyamatot, hanem annak eredményét is értjük.

hasonlósági mértéket [3] alkalmazva képes a lefordítandó mondathoz; ennek főnévi csoportjaihoz, illetve a mondatból a főnévi csoportokat kiemelve kapott mondatvázhoz hasonló, ismert fordítású mintákat keresni. A megfelelő morfológiai alakok generálásával a rendszer képes a megtalált mondatvázakba az eredetiektől különböző, de ismert fordítású főnévi csoportokat beilleszteni, így a csak teljes mondatokat kezelő fordítómemóriákénál nagyobb fedés (*recall*) érhető el, miközben a pontosság csak akkor csökken, ha a mondatváz vagy a benne található főnévi csoport fordítása a többi mondatrésztől függ. Ugyanakkor a részekből összerakott mondat még az utóbbi estekben is megkönnyítheti a fordító munkáját, hiszen néhány szó vagy szóalak változtatásával feltehetően gyorsabban tud jó fordítást találni, mintha az egész mondatot kellene lefordítania.

A mondatvázból a jövőben a főnévi csoportokon kívül más mondatrészek is kiemelhetők lehetnek majd, így tovább növelve a fedést; ugyanakkor más frázisokat megkötések nélkül kiemelve a mondatvázból a mondatváz fordításának pontossága jelentősen csökkenhet is, ezért a további vizsgálatok elvégzéséig, egyelőre csak főnévi csoportokat kezel a MetaMorpho TM.

Ahhoz, hogy ismert fordítású főnévi csoportokat tudjunk keresni a memóriában, a lefordított mondatpárok forrásnyelvi és célnyelvi főnévi csoportjait egymáshoz kell rendelni. Ha ezt a szinkronizálási feladatot a fordítóra bíznánk, az – egy csak teljes mondatokat tároló fordítómemória egyszerűségéhez képest – túl sok plusz munkát követelne tőle. A plusz munkára fordított idő pedig lehet, hogy hosszú távon se térülne meg a több keresési találat által megtakarított fordítási idővel. Ezért a főnévicsoport-szinkronizáció automatizálása mellett döntöttünk. Az automatikus főnévicsoport-szinkronizáció, teljes pontosságot biztosító módszer hiányában hibaforrásként jelenik meg a MetaMorpho TM rendszerben, cikkünkben azonban egy olyan egyszerű, szótáralapú módszert fogunk bemutatni, amelyről első eredményeink alapján azt állíthatjuk, hogy megfelelő pontosságot biztosít, abban az esetben is, ha a magyar mondatok főnévi csoportjait szintaktikai elemző nélkül, az angol megfelelőik ismeretében, egyszerű heurisztikával határozzuk meg.

2 Automatikus főnévicsoport-szinkronizáció

Ebben a szakaszban pontosítjuk a főnévi csoportok automatikus szinkronizációjának részben már ismertetett fogalmát; rámutatunk az automatikus megvalósítás eredendő nehézségeire; bemutatjuk a korábbi hasonló módszerek fő jellemzőit, illetve új módszerünk kidolgozásának okait; végül pedig részletesen ismertetjük az újonnan kidolgozott módszerünket.

2.1 Az automatikus főnévicsoport-szinkronizáció feladata

Az automatikus főnévicsoport-szinkronizáció során egymás fordításának tekinthető mondatpárok főnévi csoportjait algoritmikus módszerekkel rendeljük egymáshoz. Az összerendelés során egyes főnévi csoportok pár nélkül maradhatnak. A pár nélkül maradás oka lehet a nyelvek szintaktikai különbözősége (1. példa), állandósultnak tekinthető lexikai különbség (2. példa), vagy a fordító döntése, hogy a természetesebb hangzás érdekében átfogalmazza a mondatot (3. példa). Utóbbi esetben a mondatpár

főnévi csoportjai között lehetnek olyanok, amelyek csak részben tekinthetők egymás fordításának, az ilyen párok szinkronizálása nem lehetséges. (Az automatikus módszer hibájának kell tekinteni, ha mégis rögzít egy csak részben megfeleltethető párt.)

[I] have read [his new book on bread baking].
Elolvastam [a kenyérsütésről szóló új könyvét]. (1. példa)

1. példa: Az angol *I* személyes névmáshoz nem található a magyar fordításban neki megfeleltethető főnévi csoport. (A példákban a maximális méretű főnévi csoportokat szögletes zárójelek határolják.)

[Lolek] had [a huge breakfast].
[Lolek] jól megreggelizett. (2. példa)

2. példa: A *have a huge breakfast* angol kifejezésen belüli főnévi csoportnak nincs párja a magyar fordításban.

[Csabi] ate [ice-cream].
[Csabi] [fagyit] evett. (3. példa)
[Csabi] fagyizott.

3. példa: Ha a fordító az alsó sorban látható fordítást választja, akkor az *ice-cream–fagyi* pár nem határozható meg.

2.2 A főnévi csoportok azonosításának nehézségei

A főnévi csoportok szinkronizálásának veszélyeire már az előző egyszerű példák is rámutattak, nehézségek azonban már a szinkronizálás előtt, a főnévi csoportok automatikus azonosításakor is jelentkeznek. A főnévi csoportok határai sok esetben bizonytalanok (4. és 5. példa), így automatikus meghatározásuk nehéz.

[Ez a királypingvin] éhes.
[Ez] [a királypingvin], [az] pedig [a császárpingvin]. (4. példa)

4. példa: Az *ez* szó a második mondatban külön főnévi csoport.

I saw [the man] in [the garden].
I know [the man in the garden]. (5. példa)

5. példa: Az első mondatban az *in the garden* szabad mondatbővítmény, míg a második esetben a főnévi csoport módosítója.

Ha az alkalmazott szintaktikai elemző különbözőképp határozza meg a főnévi csoportok határait a tárolt mondatpárok forrás- és fordításoldalán, a helyes szinkronizáció elérése akadályba ütközik. Felmerül a kérdés, hogy ilyen esetekben miért nem választjuk ki a helyes elemzést a fordítás alapján. Ha az egyik szintaktikai elemzőt megbízhatónak tekintenénk, akkor a szövegpár másik oldalán korlátozott

mértékben lehetőség lehetne az elemzés egyértelműsítésére, azonban ez nem kívánt mellékhatásokkal is járhatna, hiszen, mint a 2. és 3. példák mutatják, egyáltalán nem biztos, hogy egy adott főnévi csoport fordítása módosítások nélkül jelenik meg a fordításban.

2.3 Korábbi módszerek

Egyszerű, rövid főnévi csoportok („*noun phrase chunk*") szinkronizálására Julian Kupiec 1993-ban ismertetett egy korpuszalapú módszert [6], amely jó ötleteken alapult, azonban mai szemmel nézve viszonylag alacsony pontossága, illetve a lassú (offline) feldolgozás szükségessége nem felelt meg a MetaMorpho TM rendszerbe való integrálás követelményeinek.

A főnévi csoportok szinkronizálásához hasonló fordításkeresési problémákkal foglalkoznak a statisztikai gépi fordító (SMT) rendszerek szinkronizáló algoritmusai, ugyanakkor ezek a módszerek a tanulási fázisban Kupiec módszerénél is nagyobb párhuzamos korpuszt és komoly számítási kapacitást igényelnek, így a MetaMorpho TM rendszerben kezdetben nem kívántuk alkalmazni őket. A statisztikai módszerek másik jelentős gondja, hogy szintaktikai ismeretek híján a főnévi csoportok határait nem tudják pontosan meghatározni. Utóbbira azonban külön szintaktikai elemző (esetleg sekély elemező) alkalmazásával lehetőség lenne.

A mondatnál kisebb egységek szinkronizációs módszereinek másik fő csoportját az elemzésifa-szinkronizáló (*parse-tree alignment*) módszerek alkotják, amelyekkel a közelmúltban bíztató eredményeket értek el [1], de sajnos csak nagyon hasonló elemzési fák esetén, így az angol-magyar nyelvpár kezelésére más módszert kellett kidolgoznunk.

2.4 Angol-magyar szótáralapú főnévicsoport-szinkronizáció

Új főnévicsoport-szinkronizáló módszerünk kidolgozásakor célunk a sebesség és a pontosság (*precision*) maximalizálása volt. Nagy sebességre azért van szükség, mert a tárolt párok főnévi csoportjainak fordításait jogosan várhatja a felhasználó akár már a következő mondat fordításakor is, hiszen a hagyományos fordítómemóriák is gyorsan tárolják, és azonnal elérhetővé is teszik a fordításokat. A pontosság igénye a bevezetés után talán nem szorul részletes magyarázatra, a hibás párok később hibás fordítási ajánlatokhoz vezetnek, ezért a pontosság növelése akár a fedés (*recall*) csökkenése árán is elfogadható.

A kellő gyorsaság elérése érdekében a korpuszalapú módszerek lassú (*offline*) elemzési lépései helyett gyors, tövesített szótári keresést, hasonló alakú szó (*cognate*) [9] keresést és szófaji egyezés keresést alkalmazó megoldás mellett döntöttünk.

A szinkronizáció során minden lehetséges főnévicsoport-párhoz kiszámítunk egy heurisztikus hasonlósági értéket, majd azokat a párokat jelöljük meg összetartozóként, ahol a hasonlósági érték egy küszöbértéknél nagyobb, és a pár mindkét főnévi csoportja a párbeli társára hasonlít leginkább. Utóbbi kitétel azt jelenti, hogy a lehetséges jó párok közül a legjobbat választjuk. Választásra a gyakorlatban csak akkor kényszerülünk, ha egy mondatban legalább két nagyon hasonló (vagy azonos, azaz ismétlődő) főnévi csoportot találunk.

2.4.1 Főnévi csoportok hasonlósága

A különböző nyelvű főnévi csoportok hasonlóságának vizsgálatakor célunk egyetlen, mostantól hasonlósági értéknek nevezett skalár meghatározása. A hasonlósági vizsgálat során az összehasonlított két főnévi csoport tokenjeit (~szavait) egymás után többféle módon is megpróbáljuk egymásnak megfeleltetni, majd az egyes módszerek által lefedett tokenek számából számítjuk ki (heurisztikusan) a hasonlósági értéket az 1. képletben meghatározott módon.

Először tövesített szótári keresést alkalmazunk: a forrásnyelvi főnévi csoport szavainak lehetséges töveit keressük egy speciális, tövesített indexet és találatlistát tartalmazó szótárban, majd a találatok közül csak azokat hagyjuk meg, amelyek a forrásoldalra illeszthetők és fordításuk minden szavának legalább egy lehetséges töve megtalálható a fordításbeli főnévi csoportban. A tövesített index egy keresett tőhöz tetszőleges számú, a forrásoldalon egyszavas kifejezésre tárol mutatót, viszont csak maximalizált számú többszavas kifejezéspárt tesz megtalálhatóvá. Utóbbi azt eredményezi, hogy a gyakori szavak kifejezései csak a többi, kisebb gyakoriságú kifejezésalkotó szót keresve találhatók meg, viszont a kifejezéskeresés tere stopword lista nélkül is jelentősen csökken. A szótári keresés után a forrásoldali főnévi csoport minden tokenjéhez hozzárendeljük a környezetére leghosszabban illeszthető találatokat, ezzel szűrve az elfedett rövidebb kifejezéseket (6. példa).

This is a {hard disk drive}.
In the first {drive} slot there is a {hard disk drive}. (6. példa)

6. példa: Az első mondatban a *hard disk drive* kifejezés elfedi a lehetséges *hard disk, disk drive, disk, drive, hard* találatokat. A *hard disk drive–merevlemez meghajtó*, illetve *drive–meghajtó* szótári találatokat feltételezve az utóbbi pár felvétele hibás lenne. A második példában viszont a *drive* – független előfordulása miatt – mégis szerepelhet a szótári találtatok között, ha a fordításban is megtalálható.

A szótári megfeleltetés után, a főnévicsoport-pár le nem fedett szavai között hasonló alakúakat (*cognate*, pl. az angol *parliament* és a magyar *parlament* szavak) keresünk a Simard és társai által kidolgozotthoz [9] nagyon hasonló algoritmust alkalmazva. Megvalósításunkban két szót akkor tekintünk hasonló alakúnak, ha egy karakternél hosszabbak, tartalmaznak legalább egy nagybetűt, számot vagy valami más speciális karaktert, és legfeljebb az ötödik karaktertől különböznek. (Az ismertetettnél kevésbé hatékonyan számítható, de nagyobb fedésű algoritmus alkotható a szavak közötti Levenshtein-távolság mérésével.)

A korábban le nem fedett szavakat ezután szófajaik alapján próbáljuk egymáshoz rendelni. Ha egy szóhoz több lehetséges szófajt is rendelt a morfológiai elemző, bármelyikkel való egyezést elfogadunk.

A lefedetlen szavak közül a pusztán grammatikai funkciót betöltőket egy kis büntetőpontszámot alkalmazva kiemeljük, így téve lehetővé a csak a grammatikai funkciót betöltő szavaikban különböző rövid főnévicsoport-párok egymáshoz rendelését (7. példa).

Where is [**my** book]?
Hol [**a** könyvem]? (7. példa)

7. példa: A *my book* megfelelője az *a könyvem* főnévi csoport, ugyanakkor szinkronizálásukkor gondot okozhatna, hogy a 4 szó közül 2 nem feleltethető meg egymásnak, ha nem tekintünk el a pusztán grammatikai funkciót hordozó, egymásnak meg nem feleltethető szavaktól.

Az előzőekben ismertetett illesztések után a *h* hasonlósági értéket az alábbi 1. képlet alapján számítjuk ki.

$$h = \frac{a \cdot Dict + b \cdot Cogn + c \cdot POS - d \cdot GF}{T - GF} \qquad \text{(1. képlet)}$$

1. képlet: A hasonlósági érték számításának módja. A képletben *Dict* a szótárral megfeleltetett tokenek száma, *Cogn* a hasonló szavakat keresve lefedett tokene száma, *POS* a szófaji illesztéssel lefedett tokenek száma, *GF* a le nem fedett pusztán grammatikai funkciójú tokenek száma, *T* pedig a két főnévi csoport tokenszámának összege. Az a=1, b=0.9, c=0.3, d=0.1 konstans együtthatók, empirikusan meghatározott értékekkel.

Az 1. képlet kísérletezéssel, de csak kevés mintán meghatározott konstans együtthatóit (vagy akár magát a képletet), a jövőben a főnévi csoportok közötti kapcsolatokat is tartalmazó párhuzamos korpusz vizsgálatával kívánjuk finomítani.

3 Magyar főnévi csoportok meghatározása angol megfelelőik alapján

A MetaMorpho TM rendszerben a tárolt mondatpárok angol oldalán a MetaMorpho elemzőt és a hozzá készített angol (valójában angol-magyar) nyelvtant [11] használjuk a főnévi csoportok meghatározására. A magyar oldal megfelelő pontosságú elemzésre azonban még nem alkalmas az elemzőhöz fejlesztett magyar-angol nyelvtan [8], így megpróbáltuk az angol elemzővel automatikusan meghatározott főnévi csoportokhoz rendelhető magyar főnévi csoportokat az angol főnévi csoportok szavait és kifejezéseit a magyar szövegre leképezve meghatározni.

Heurisztikus megoldásunkban a 2.4.1. pontban bemutatott módszerekkel minden angol főnévi csoport szavaihoz szótári egyezéseket és hasonló alakú szavakat keresünk. A keresés során különbség a 2.4.1. pontban ismertetett módszerhez képest, hogy a grammatikai funkciót betöltő szavakat nem keressük a szótárban, mivel ezek fordítása a magyar mondatban a keresett főnévi csoporttól függetlenül bárhol előfordulhat. Azokat a szavakat tekintjük grammatikai funkciót betöltőnek, amelyek morfológiai elemzéssel meghatározott lehetséges szófajai között csak néhány, előre meghatározott szófaj (névmás, névelő stb.) szerepel.

Mivel egy szó akár többször is előfordulhat a mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a magyar mondatra. Természetesen a találatok között más szavakat is tartalmazhat a kije-

lölt illeszkedés. A legrövidebb illeszkedés kiszámítása költséges művelet, érdemes a fedéshossz korlátozásával redukálni a keresési teret.[26]

Az illeszkedést ezek után egyszerű szabályok és az angol főnévi csoport le nem fedett szavainak figyelembe vételével teljes magyar főnévi csoporttá bővítjük. A lényegesebb szabályokat a következőképp foglalhatjuk össze. Először a találatok közötti szavak szófaját próbáljuk az angol főnévi csoport meg nem talált szavainak szófajával egyeztetni, majd ha pár nélküli melléknév vagy főnév szerepel az angol főnévi csoportban, akkor baloldalról próbáljuk bővíteni a magyar főnévi csoportot. A bővítés során az angol főnévi csoport pár nélküli főneveinél, illetve mellékneveinél maximum eggyel többet engedünk meg, illetve nem folytatjuk a bővítést, ha igéhez vagy más a főnévi csoportba nem illő szóhoz, írásjelhez érünk. Jobbra csak akkor bővítjük a főnévi csoportot, ha a baloldali bővítési kísérlet után is maradt páratlan főnév az angol főnévi csoportban. A főnévi csoportot mindig kibővítjük a tőle balra található névelővel. (A módszer a megvalósításban kicsit bonyolultabb, mivel szófaji egyértelműsítés hiányában az egyes szavak szófajai esetében több lehetőség közül kell választanunk.)

A módszer egyszerűségéből és a szótári találatok bizonytalanságából adódóan néha jelentős hibákat ejt, ezeket azonban a 2.4.1. pontban ismertetett módszerrel könnyen szűrni lehet: ha az angol párja alapján meghatározott magyar főnévi csoport nem hasonló eléggé angol párjára, akkor a párt nem rögzítjük.

A módszer jelenlegi formájában egyesével, egymástól függetlenül választja ki az angol főnévi csoportokhoz rendelt párokat, így hibázás esetén akár átfedő párok is kialakulhatnak, bár átfedés esetén nagyon kicsi az esélye annak, hogy mindkét pár elérje a szükséges hasonlósági pontszámot. Ha ez mégis megtörténne, akkor a jelenlegi megoldásban mindkét párt elvetjük. A jövőben meg fogjuk vizsgálni, hogy jobb megoldás lenne-e, ha – balról jobbra haladva – a szükséges hasonlósági pontszámot megszerző főnévi csoportok által lefedett szavakat foglaltnak jelölnénk, és nem használnánk őket más főnévi csoportban; illetve megpróbálunk módszert kidolgozni arra, hogy még a főnévi csoportok bővítése előtt feltérképezzük viszonyaikat a magyar mondatban.

4 Eredmények

Első kísérleteinket az informatikai témájú szövegeket tartalmazó SZAK-korpusz [7] 40 viszonylag hosszú mondatpárt (átlagosan 23 szó/mondat) tartalmazó kis részletén végeztük. A kísérlethez viszonylag kis méretű, 116 000 szó- és kifejezéspárt tartalmazó szótárt használtunk.

Az automatikusan meghatározott angol főnévi csoportok 56 százalékának volt csak meghatározható magyar fordítása. (Az angol mondatok alanya gyakran személyes névmás volt, a fordító néha igei szerkezetre cserélte a főnévi csoportot, néhány esetben pedig az angol elemző hibázott.) Módszerünk pontossága 84 % volt, azaz a fordítómemóriába felvett párok kevesebb, mint 1/6 része hibás.

---

[26] Megjegyzés: A paraméteres bonyolultságelmélet rámutat arra, hogy a bonyolult (NP-teljes) feladatok között több olyan is van, amelyeknél bizonyos paramétereik rögzítésével redukálható a keresési tér, így könnyű (polinomiális) feladattá vezethetők vissza (természetesen csak rögzített paraméterek mellett).

Az eredetitől jelentősen különbözően fordított mondatokat elhagyva azt tapasztaltuk, hogy a pontosság 91 százalékra nőtt, azaz ha a fordítómemória felhasználója a felhasználói felületen (egyetlen kattintással, vagy billentyűkombinációval) megjelölhetné a részekre nem bontható fordításokat, akkor a pontosság jelentősen növelhető lenne. A párosítható angol főnévi csoportokhoz mérve 65 százalékos fedést (*recall*) sikerült elérni, ami a szótár bővítésével, reményeink szerint még növelhető. (A 65%-os fedés azt jelenti, hogy az angol főnévi csoportok kicsit több, mint 1/3 részéhez rendelt fordítást a módszerünk.)

A 2.4. és 3. pontokban ismertetett módszerek sebessége megfelelő, együttes futásidejük a próbák során legfeljebb néhány ezredmásodperc volt egy átlagos számítógépen (igaz a módszereket a hatékonyságra ügyelve implementáltuk), ez az angol főnévi csoportok megállapítására használt szintaktikai elemzés idejéhez képest elhanyagolható.

5 További tervek

Az eddiginél nagyobb mintán végzett mérésekhez kézzel címkézett tesztanyag, főnévicsoport-szinten párhuzamosított korpusz építése szükséges, így ez rövidtávú terveink között szerepel. A tesztanyag méretének növekedtével annak egy részét az 1. képlet konstans paramétereinek behangolására szeretnénk fordítani.

A szinkronizáló módszer önálló mérése mellett azt is szeretnénk megvizsgálni, hogy egyes hibái hogyan hatnak a teljes MetaMorpho TM rendszerre. A teljes rendszer tesztelését a Hodász Gábor által kidolgozott módszerekkel [5] végezzük. Az eredményeket elemezve kell majd módszert adnunk a memóriába került hibás párok szűrésére, illetve az esetlegesen kritikus hatású hibák elkerülésére.

A hatékonysági mérésekkel párhuzamosan, korpuszalapú, statisztikai szótárbővítő módszereket alkalmazva gyarapítjuk majd a szinkronizációhoz használt szótárat; illetve a magyar főnévi csoportok meghatározására kidolgozott, 3. pontban ismertetett módszer helyett a MetaMorpho elemzőhöz készülő magyar nyelvtant is ki fogjuk próbálni.

Hivatkozások

1. Groves, D.; Hearne, M.; Way, A.: Robust Sub-Sentential Alignment of Phrase-Structure Trees. COLING '04, Geneva, Switzerland, 2004
2. Grőbler Tamás, Hodász Gábor, Kis Balázs: MetaMorpho TM: A Rule-Based Translation Corpus. International Conference on Language Resources and Evaluation, Lisszabon, 2004.
3. Hodász Gábor: Nyelvi hasonlóságon alapuló intelligens keresés fordítómemóriában. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged
4. Hodász Gábor, Pohl Gábor: MetaMorpho TM: a linguistically enriched translation memory. In: International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan, ISBN 954-90906-9-8), Borovets, Bulgaria, 24 Sept. 2005.
5. Hodász Gábor: Fordítómemóriák és mintaalapú fordító rendszerek kiértékelésének módszerei. ugyanebben a kötetben

6. Julian Kupiec: An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora. In: Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 17-22, 1993
7. Kis Ádám; Kis Balázs: A Prescriptive Corpus-based Technical Dictionary. In: Papers in Computational Lexicography: Proceedings of COMPLEX 2003. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2003.
8. Merényi Csaba: A MetaMorpho magyar-angol gépi fordító rendszer igei vonzatkereteit működtető nyelvtan. ugyanebben a kötetben
9. Simard, M., Foster, G. & Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation, (TMI92), Montreal, pp. 67-81, 1992
10. Somers, H.: An Overview of EBMT. In M. Carl. and A. Way. (eds.) Recent Advances in Example-based Machine Translation, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3-57. 2003.
11. Tihanyi, L.: A MetaMorpho projekt 2004-ben. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2004.

A Hunglish korpusz és szótár

Halácsy Péter[1], Kornai András[1], Németh László[1], Sass Bálint[2],
Varga Dániel[1], Váradi Tamás[2], Vonyó Attila

[1] BME – Média Oktató és Kutató Központ
{hp, kornai, nemeth, daniel}@mokk.bme.hu
[2] MTA Nyelvtudományi Intézet,
{joker, varadi}@nytud.hu

Kivonat: Cikkünkben a Budapest Műszaki Egyetem és az MTA Nyelvtudományi Intézet által épített angol–magyar Hunglish korpuszt és Hunglish szótárat mutatjuk be.

Bevezetés

A Budapesti Műszaki Egyetemen 2004 nyarán indult el a Hunglish projekt [5], melynek fő célja egy statisztikai elven működő gépi nyersfordító rendszer kifejlesztése volt. A feladat megoldásához létre kellett hoznunk egy mondatszinten illesztett, magyar–angol párhuzamos korpuszt. A mondatszintű illesztés a korpusz minden forrásnyelvi mondatához hozzárendeli annak célnyelvi fordítását, amely esetenként akár több mondatból is állhat. A mondatszinten illesztett korpusz legalapvetőbb felhasználási területe a gépi fordítás, amely a fordítás statisztikai modelljének paramétereit képes beállítani a korpusz alapján.

A Hunglish projekt eredményeként tehát — az angol-magyar nyersfordító prototípus mellett — elkészült egy automatikus mondatillesztő program, létrejött egy angol-magyar párhuzamos korpusz, illetve kialakult egy teljes párhuzamos korpusz építésére alkalmas eszközkészlet és módszertan. A projekt minden eredményét a Creative Commons “nevezd meg” licensz alatt tettük elérhetővé, vagyis minden termékünk bárki számára szabadon hozzáférhető, felhasználható, átdolgozható. A korpusz felhasználási lehetőségei változatosak: felhasználható nyelvtechnológiai, számítógépes nyelvészeti alkalmazásokban (lásd például ebben a kötetben Miháltz és Pohl 2005), fordítástámogatásban, kétnyelvű terminológiai adatbázis építésben, sőt talán még a nyelvoktatásban és fordítóképzésben is. Nyelvfüggetlen, pontosabban nyelvpárfüggetlen eszközkészletünk egyszerűvé teszi további párhuzamos korpuszok építését.

Már az adatbázis építése során is — de a gépi nyersfordító fejlesztésekor különösen — felhasználtuk Vonyó Attila magyar–angol szótárát. (Ez az anyag képezi a sokak által használt magyar–angol Sztaki-szótár[27] alapját is.) A szótárt más forrásokból gyűjtött terminológiai és egyéb kétnyelvű adatbázisokkal összefésültük, s a pár-

[27] http://dict.sztaki.hu

huzamos mondatok felhasználásával a szótár minőségét javítottuk. Ettől is azt reméljük, hogy sokan sokféle célra használni tudják majd, és a jövőben több tőlünk független kutatás vagy szolgáltatás részeként meg fog jelenni.

1. A korpusz alapanyaga

A webes források automatikus felkutatása és feldolgozása [7] óta igen gyakran alkalmazott módszer párhuzamos korpusz építésére. Az eljárás angol és másik világnyelv esetén ígéretes eredményeket adott [2,8]. Chen és Nie angol–arab nyelvpárra 2000 párhuzamos weboldalról mindösszesen 2,3 millió szövegszót, Resnik és Smith angol–francia nyelvpár esetén 2491 párhuzamos oldalt tudott felkutatni automatikus módszerekkel.

A mi célunk ennél legalább egy nagyságrenddel nagyobb korpusz építése volt. Tapasztalatunk szerint azonban a magyar weben egyszerűen nincs elegendő automatikusan felkutatható párhuzamos weboldal.

Ezért egy manuális, de sokkal hatékonyabb módszert választottunk. Ugyan fő forrásunk továbbra is a web, de nem próbáltunk automatikusan fordításpárokat találni. Bár így a dokumentumok felkutatása komoly munkabefektetést jelentett, a dokumentumok manuális letöltése és párosítása után a további lépéseket már automatizáltan végeztük. A korpusz az alábbi fő forrásokból épül fel:

Irodalmi szövegek

Az irodalmi szövegek fő forrása a *Project Gutenberg*[28] és a *Magyar Elektronikus Könyvtár*[29] , amelyek adatbázisait összevetve kikerestük azokat a műveket, amelyek egymás fordításai. Innen majdnem száz klasszikus irodalmi mű tölthető le Jane Austentól Tolsztojig. Ugyanilyen fontosak a weben szép számmal fellelhető még szerzői jog védelme alatt álló modern művek is.

A szövegek között tehát voltak olyanok (a modern irodalmi anyagon kívül a később említett filmfeliratok is), amelyeknek a változatlan formában történő újrapublikálása jogi problémákba ütközne. Itt említjük meg, hogy ezeket a szövegeket a párhuzamosítás után összefűztük, majd angol–magyar mondatpárok véletlenszerűen (pontosabban ábécérendbe) rendezett halmazává alakítva publikáltuk. Ezzel az eredeti szövegek rekonstruálását lehetetlenné tettük, megvédve így a szövegek jogtulajdonosainak szerzői jogait. Ugyanakkor a korpusz legfontosabb általunk megcélzott felhasználásai a mondatok sorrendjét nem veszik figyelembe, tehát ezen célokra a keverési művelet a korpusz értékét nem csökkenti.

Jogi adatbázis

Legnagyobb forrásunk az EU közösségi jogszabályok *CELEX adatbázisa* és az *Európai Alkotmány*. [30]

[28] http://www.gutenberg.org

[29] http://mek.oszk.hu/indexeng.phtml

[30] A forrás érdekessége, hogy az adatbázis elérhető a közösség minden hivatalos nyelvén. A soknyelvű mondatszintű párhuzamosítás elvégzését tervbe vettük.

Nyílt forráskódú szoftverek dokumentációi

A nyílt forráskódú szoftverek honosításainak eredményeit tudomásunk szerint először [9] használta párhuzamos korpusz alapanyagaként. A Hunglish korpuszba a KDE, Gnome, OpenOffice, Mozilla és GNU eszközök dokumentációit építettük be.

Filmfeliratok

Az internetről letölthető *filmfeliratoknak* csak egy részét (mintegy 400 film) vettük be az adatbázisunkba, főleg kísérleti jelleggel. Ezek, bár sok esetben elég rossz minőségű fordítások, bizonyos nyelvhasználatra és szókészletre (pl. szleng és káromkodás) olyan kiváló forrásanyagot tartalmaznak, amely a formálisabb forrásokból nem lenne kinyerhető.

Hírek, magazinok

Jó minőségű, de az internetről nem letölthető, ezért nehezen beszerezhető anyagok származhatnak *kétnyelvű magazinokból*, illetve magazinok magyar nyelvre fordított kiadásaiból. Mi a National Geographic és a Diplomacy and Trade magazin néhány magyarra fordított számát dolgoztuk fel.

Sajtófigyelés

A Magyar Telekom Rt. szabad felhasználásra a rendelkezésünkre bocsátott nagy mennyiségű távközlési témájú sajtóanyagot, amelyet fordítóik ültettek át angol nyelvre.

További, még fel nem dolgozott források

A fentieken kívül megkezdtük további források korpuszba építését is. *Tőzsdei cégek nyilvános éves jelentései* sokszor elérhetőek angol nyelven is a vállalat weboldalán. Mi három vállalat 19 éves jelentését töltöttük le. *Vallási szövegek*: a katolikus egyház által sok nyelven publikált pápai ediktumok feldolgozását már megkezdtük.

1. Táblázat: A korpusz összetétele szövegtípusok szerint

forrás	Angol tokenek (millió)	Magyar tokenek (millió)
irodalom	14,6	11,5
jog	24,1	18,3
filmfelirat	2,5	1,9
szoftver	0,8	0,7
magazinok	0,3	0,3
sajtó	2,1	1,7
összesen	**44,5**	**34,5**

2. A korpusz feldolgozása

Első lépésben pontosan párosítottuk a letöltött vagy más úton szerzett nyers dokumentumokat. Ezután kinyertük belőle a nyers szöveget, formátumuk, karakterkészletük konvertálásával. Ez az egyszerűnek hangzó lépéssor a korpuszépítés nagy mennyiségű manuális munkát és szakértelmet igénylő fázisa.

A forrásként szolgáló állományok formátuma és karakterkódolása igen változatos, így például PDF, Postscript, DOC, RTF, HTML, SXW, TEX, valamint különböző szöveges formátumok automatikus konvertálását kellett megoldanunk, amihez nyílt forráskódú programkönyvtárakat és segédprogramokat használtunk.

A táblázatokat és szigorú tördelést tartalmazó, Quark Express formátumban lévő, szigorúan a nyomtatás előtti fázisból hozzánk került magazinok feldolgozása alig automatizálható (sokszor csak az OCR programok jönnek számításba). Szerencsére a szövegek nagy része nem ilyen problematikus: az `antiword`, `catword`, `html2text` és más hasonló nyílt forráskódú programok megfelelően alkalmazhatóak.

A karakterkódolás meghatározó jelentőséggel bír a szöveges adatok tárolásában. Tapasztalataink szerint érdemes a nehezebben kezelhető, de veszteségmentes Unicode karakterkódolást választani. A Hunglish esetében mégis 8 bites karakterkódolást alkalmaztunk, mert egyes eszközeink (például a mondatra szegmentáló) csak ezt támogatják, és a veszteség a korpusz legfontosabb várható alkalmazásait (szótárépítés, gépi fordító tanítása) véleményünk szerint nem hátráltatja. A korpusz magyar oldalán ISO 8859-2 kódolást, az angol oldalán ISO 8859-1 kódolást alkalmaztunk. Emiatt egyes speciális szimbólumokat le kellett cserélnünk, de ezek az esetek a korpuszban rendkívül ritkán fordultak elő.

3. Mondatszintű párhuzamosítás

Ha a dokumentumszinten párosított nyers szövegek már elkészültek, mindössze néhány perc elkészíteni egy párhuzamos dokumentumpár mondatszintű illesztését.

Ehhez első lépés a mondathatár-azonosítás, amit a szabályalapú `huntoken` [6] programmal végeztünk, magyar és angol kivételszótárakkal.

A szóhatárolás, amelyre szintén szükség van az illesztés előtt, tapasztalatunk szerint nem kritikus lépés. A komplex `huntoken` az illesztés szempontjából feleslegesen jelöl meg nyílt tokenosztályokat, címeket, kifejezéseket egy tokennek. Helyette egy egyszerű háromsoros programot használtunk, ami tulajdonképpen nem tesz mást, mint a szavak végéről leválasztja az írásjeleket.

A mondatszintű párhuzamosításhoz fejlesztettük ki a `hunalign` programot [11]. A `hunalign` legfőbb előnye, hogy ún. zajos szövegeken is megbízhatóan dolgozik. Mondatok kiesését és összeolvadását akár nagyobb számú mondat esetében is kezeli; ilyen például a csak az egyik oldalon jelen levő utószó, vagy a nagy számú lábjegyzet. A mondatok sorrendjének felcserélődését a `hunalign` nem kezeli. A tanítókorpuszainkban és szúrópróbaszerű korpuszelemzéseink során azt tapasztaltuk, hogy ez a jelenség igen ritka. Magyar–angol tesztkorpuszunkon a `hunalign` pontossága 99.34%, ami jelentősen meghaladja a standard statisztikus módszerekét, köszönhetően – többek között – a kétnyelvű szótár felhasználásának.

A `hunalign` működési folyamata nagyon vázlatosan a következő:

1. Elkészíti a nyers, tokenizált szöveg magyar mondatainak nyersfordítását: a bemeneti magyar–angol szótár alapján lecseréli a magyar szavakat a célszövegben leggyakoribb angol megfelelőjükre, a szótárban nem szereplő szavakat pedig meghagyja eredeti alakjukban (így például a számokat, jogszabályban a paragrafus-hivatkozásokat, email címeket stb. is).
2. A nyersfordítás és a célszöveg hasonlósága, valamint a mondathosszarány alapján hasonlósági mértéket számol a forrásszöveg és a célszöveg mondatai között. Ez alapján megkeresi a legjobbnak vélt illesztést egy dinamikus programozási feladat megoldásával.
3. A megtalált illeszkedő mondatpárok alapján statisztikai módszerekkel szótári tételeket azonosít, és ezekkel kiegészíti a kiinduló szótárat.
4. A kiegészített szótárat felhasználva újra elvégzi a szöveg illesztését az első két pont szerint. (Tapasztalataink szerint ennek a ciklusnak a további iterációja már nem javítja az illesztés minőségét.)

Az algoritmus leírásából látható, hogy a `hunalign`-nak az első nyersfordítás elkészítéséhez szüksége van egy kiinduló szótárra. Ehhez mi a Vonyó szótárat alkalmaztuk. Ahhoz viszont, hogy egy toldalékolt szót megtaláljunk a szótárban, szótövezést kell végeznünk. Ehhez a `hunmorph` programot [10] használtuk: a magyar szövegeknél a `morphdb.hu` [12], angol szövegeknél a szintén saját fejlesztésű `morphdb.en` nyelvi erőforrást alkalmaztuk. Nem foglalkoztunk azokkal az esetekkel, amikor a szótövezés nem egyértelműen adja meg a szó lemmáját. Ilyenkor egyszerűen a legkevesebb toldalékot leválasztó elemzést választjuk.

Vegyük észre azonban, hogy `hunalign` képes kiinduló szótár nélkül is működni. Ilyenkor az első lépésben a nyersfordítás nem változtat semmit a kiinduló mondaton, és a második lépésben használt hasonlósági függvény elsődlegesen a mondathosszak arányától fog függeni [4]. Ebben az esetben a `hunalign` működése hasonlít a sokak által használt vanilla [1] illesztőéhez. Az automatikus szótárépítési fázis után újból elvégzett második párhuzamosítás azonban már jóval nagyobb pontosságot ér el, mint az első.

A 2. táblázat a `hunalign` pontosságát és fedését mutatja be különböző erőforrásokkal a MULTEXT-East 1984 magyar–angol párhuzamos korpuszon [3] mérve.

Látható, hogy a kiinduló szótár növeli a pontosságot, különösen akkor, ha szótövezést is végzünk.

2. Táblázat: A hunalign pontossága és fedése az 1984 korpuszon különböző beállításokkal: szótár: kiinduló szótár használatával, tövez: tövező használatával, iter: automatikus szótárépítés, kiinduló szótár nélkül, id: szótár nélküli, csak azonos szótokeneket illesztő mód, hossz: illesztés a mondat karakterszámának alapján

módszer	**pontosság**	**Fedés**
hossz	97.58	97.55
hossz+id	97.65	97.42
szótár	97.30	97.08
hossz+szótár	98.86	98.88
hossz+szótár+tövez	99.34	99.34
hossz+tövez	98.63	98.74
hossz+iter+tövez	99.12	99.18

A 2. táblázatból az is kiolvasható, hogy az illesztő bármiféle nyelvi erőforrás nélkül is jó eredményt ér el. Megvizsgáltuk, hogy ez más nyelvpárokra is igaz-e. A 3. táblázat mutatja, hogy hogyan teljesít a `hunalign` nyelvi erőforrás nélkül a MULTEXT-East 1984 korpusz más nyelvű párhuzamos anyagain mérve. Megjegyezzük, hogy az SGML formátumú korpusz karakterkonverzióját nem minden esetben végeztük el, ami ront a mondathossz alapú heurisztika pontosságán, és feltehetőleg felelős a román–angol nyelvpáron elért kiugróan alacsony eredményért.

3. Táblázat: A hunalign pontossága és fedése a MULTEXT-East 1984 korpuszon különböző angol–X nyelvpárokra, szótári erőforrás használata nélkül.

nyelv	pontosság	fedés
észt	99.34	99.53
cseh	98.60	98.75
román	97.10	97.98
szlovén	99.44	99.61

4. Kézi illesztés

Az illesztőalgoritmus teszteléséhez szükségünk volt manuálisan illesztett párhuzamos szövegre. Ehhez a fent már tárgyalt MULTEXT-East 1984 párhuzamos korpusz mellett felhasználtuk John Steinbeck Egy marék arany című művének általunk elkészített manuális illesztését is.

A manuálisan végzett munka három részből állt:

1. automatikusan mondatokra bontott és mondatszinten automatikusan illesztett korpusz illesztésének kézi javítása
2. az eredeti automatikus mondatszegmentálás hibáinak kézi javítása
3. kézzel javított szegmentálású, automatikusan illesztett korpusz illesztésének kézi javítása

Az automatikus mondatszegmentálás a `huntoken`, az automatikus mondatszintű illesztés a `hunalign` programmal történt. Az első szakasz végén létrejött korpusz felhasználható párhuzamosító algoritmusok pontosságának kiértékelésére abban a tipikus helyzetben, amikor a bemeneti mondatra szegmentálás automatikus, tehát hibákat tartalmazhat. A második és harmadik munkafázis eredményeképpen hibátlannak tekinthető párhuzamos korpuszt kaptunk. Ez sokféle célra hasznosítható, de az elsődleges célja párhuzamosító algoritmusok pontosságának értékelése azon feltétel mellett, hogy a korpuszban a mondathatárokat hibátlanul ismerjük.

A korpuszt felhasználtuk a `hunalign` tesztelésére. Természetesen az algoritmus egy korábbi változatának kimenete befolyásolta a végeredményt, tehát az algoritmusunk ezen korpuszokon való értékelése megkérdőjelezhető. De egyrészt a végső manuális párhuzamosítás elegendően függetlennek tekinthető a gépi párhuzamosítási lépésektől, másrészt az algoritmus inkrementális paraméterbeállítására a korpusz mindenképpen alkalmazható.

A manuális mondatszintű illesztés munkaigényének becslését segítheti, ha közöljük a következő adatokat: A regény terjedelme 230 oldal, 57,000 szó. Ezen a szövegen a fent leírt manuális munkát négy ember végezte, a teljes ráfordított munkaidő körülbelül 240 órát, azaz 6 emberhetet tett ki.

5. Korpuszjavítás

Egy gépi tanulási szoftverimplementáció sebességét általában joggal tekintik kevésbé fontosnak a pontosságához képest. A Hunglish korpusz építése során azonban sokféle módon előnyünkre tudtuk fordítani azt a tényt, hogy a C++ nyelven írt `hunalign` legalább egy nagyságrenddel gyorsabb, mint más hasonló célú implementációk. A teljes, több tízezer dokumentumból álló korpuszunk párhuzamosítása így néhány nap helyett néhány órán belül elvégezhető volt.

Az algoritmus gyorsasága lehetővé tesz egy iteratív munkafolyamatot: A `hunalign` által legalacsonyabb konfidenciaszintűnek ítélt párhuzamosítások tipikusan valamilyen dokumentumszintű illeszkedési hibát, vagy súlyosabb szövegnormalizációs hibát tartalmaznak. (Példák az előbbire: több kötet illesztése egyhez, nagyszámú egynyelvű lábjegyzet, novelláskötet más novella-sorrenddel, vagy akár gyermekek számára átdolgozott kiadás.) Az alacsony konfidenciaszintű részek megvizsgálásával az ilyen problémák feltárhatók és orvosolhatóak, vagy a menthetetlen szöveg eliminálható. Előfordulhat, hogy a javítást a szövegkinyerő, tokenizáló, mondatra szegmentáló vagy párhuzamosító algoritmusainkon kell megtennünk. Egy-egy ilyen javítás a párhuzamosított dokumentumok ezreit érintheti, tehát a teljes korpuszépítési ciklus újbóli elvégzése után a legalacsonyabb konfidenciaszintű mondatok listája lényegesen megváltozhat. Ezt a folyamatot addig ismételtük, amíg már a legalacsonyabb konfidenciaszintű párhuzamosítások is elfogadhatóak voltak.

Egy másik példa iteratívan végezhető javításra a mondatra szegmentáló kivételszótárának bővítése. Ehhez felhasználtuk azon szavak listáját, amelyek a két mondatot egy mondathoz rendelő szegmentumokban nagy gyakorisággal az elválasztó írásjel előtt állnak.

6. Szótárépítés

A kiinduló kétnyelvű Vonyó szótárat a korpusz alapján javítottuk. Először a korpuszban nem megtalálható rekordokat törölve egy kisebb, de jobb minőségű szótárat kaptunk. Ezután új rekordokat vettünk be, amelyeket a párhuzamos mondatpárokon futó statisztikus alapú automatikus szótárépítő algoritmusunk azonosított.

A kétnyelvű szótárak általában rejtett, de fontos tulajdonsága, hogy a különböző jelentéseket, fordítási alternatívákat gyakoriság szerint súlyozva mutatják. Az elkészült Hunglish szótárba ezek a gyakorisági adatok a korpusz alapján kerültek be.

7. Keresőfelület

A korpuszhoz és szótárhoz készült kereső szolgáltatás kiegészítője lehet a jelenlegi webes szótár szolgáltatásoknak. A kereső találati listáján a párhuzamos mondatok jelennek meg. A beépített magyar és angol szótövezőnek, illetve a nyílt forráskódú Lucene programkönyvtár keresőalgoritmusának köszönhetően nem csak lemmák, hanem teljes kifejezések, idiómák is kényelmesen és hatékonyan kereshetők.

8. Köszönetnyilvánítás

A Hunglish projekt az Informatikai és Hírközlési Minisztérium ITEM pályázatán nyert támogatással vált lehetővé (IHM-ITEM 2003/76/6/2004). A projekthez való hozzájárulásukért köszönettel tartozunk Gyarmati Ágnesnek, Héja Enikőnek, Mészáros Ágnesnek, Balogh Attilának, Kornai Andrásnak és Trón Viktornak. Köszönetet mondunk a Magyar Telekom Rt.-nek a Sajtófigyelő korpusz publikálhatóvá tételéért és a projekt infrastrukturális támogatásáért.

Bibliográfia

[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] Jiang Chen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[3] Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 315–319, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.

[5] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, Viktor Trón, and Dániel Varga. Hunglish: nyílt statisztikai magyar–angol gépi nyersfordító. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 81–84. Szegedi Tudományegyetem, 2004.

[6] András Miháczi, László Németh, and Miklós Rácz. Magyar szövegek természetes nyelvi előfeldolgozása. In *Proceedings of the 1st Hungarian Computational Linguistics Conference*. Szegedi Tudományegyetem, 2003.

[7] Philip Resnik. Parallel strands: A preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber, and E. Hovy, editors, *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas*, Langhorne, PA, 1998. Springer.

[8] Philip Resnik and Noah Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

[9] Jörg Tiedemann and Lars Nygaard. The opus corpus - parallel and free. In *Proceedings of LREC'04*, volume IV, pages 1183–1186, Lisbon, 2004.

[10] Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*, 2005.

[11] Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590-596., 2005.

[12] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, and Vajda Péter. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III Magyar Számítógépes Nyelvészeti Konferencia*, 2005. megjelenés alatt.

A MoBiMouse Plus szótárak készítése közben szerzett tapasztalatokról

(Egy szótárfeldolgozó programcsomag elvi lehetőségei, emberi segítséggel)

Vörösss Ferenc, Trepák Mónika

MorphoLogic
1126 Budapest Orbánhegyi út 5.
{voross,trepak}@morphologic.hu

Kivonat: Dolgozatunkban tipografizált szótári adatbázisok jelölőnyelvi adatbázissá alakítására tett, különböző feldolgozási elméletekhez kapcsolódó kísérleteiről szólunk. Felvázoljuk, hogy a szótári adatbázisok kialakítása, illetve az ezek intelligens keresését lehetővé tevő munkafolyamatok során milyen tapasztalatokkal gazdagodtunk a szótárírási szokásokat illetően. Cikkünk tanácsokkal és javaslatokkal szeretné segíteni azokat a lexikográfusokat, akik szeretnék szótáraikat egyszerűbben feldolgozható, átláthatóbb, egységesebb, felhasználóbarátabb szerkezetben szerkeszteni, és nem szeretnének az általunk tapasztaltakhoz hasonló hibákat elkövetni. Javaslatainkat példákkal is illusztráljuk, melyekhez az elmúlt években általunk feldolgozott több, mint 50 szótár szolgál anyaggal. Kitérünk a papírszótárak szerkezeti-tipográfiai egységének és egyértelműségének fontosságára, valamint felvázoljuk, miért jelenthet egyedülálló és eredményes módszert a papíralapú szótárak feldolgozásában, XML-adatbázissá alakításában az általunk kidolgozott elképzelés és a rá épülő alkalmazások.

1 Bevezetés

Cégünk, a MorphoLogic, 1994 óta foglalkozik elektronikus szótárak kiadásával. Kezdetben szószedeteket, és strukturált nyelvi forrásanyagokból előállított, egyszerűsített szerkezetű szótárakat jelentettünk meg, a cégen belül kialakított egyedi formátumban. Szótáraink versenyképességét a szótári tartalom intelligens keresésére[31] alkalmas hozzáadott nyelvi háttér-információkkal és -adatbázisokkal teremtettük meg. Ez a szótárpiacon egyedülálló nyelvtechnológiai módszer jogosan ébresztette azt a reményt, hogy a szótári tartalomhoz hozzáadott nyelvészeti tudás felkelti majd na-

[31] A MorphoLogic szótárprogramjai a kezdetektől támogatják a szótári tartalom nyelvi elemzésen és szóalaktani háttérinformációkon alapuló keresését, a magyar és néhány kelet-európai nyelv mellett többek között angolul és németül is. Az intelligens keresésről illetve definíciójáról bővebben lásd: Prószéky – Kis: Számítógéppel emberi nyelven. Szak Kiadó, Bicske (1999) 187-193; 207-213 p.

gyobb (esetleg nemzetközi) szótárkiadók érdeklődését is, és lehetőségünk lesz szótáraik közös, elektronikus kiadására.

A nemzetközi szótárkészítési módszerek vizsgálata azt mutatta, hogy az elismert szótáríró műhelyek egyre inkább a nyelvi adatbázisok jelölőnyelveken[32] való leilletve átírásával alkotják meg szótáraikat. Így ha együttműködésre törekszünk velük, szótári rendszerünket és szótárkészítési módszereinket olyan irányban érdemes fejlesztenünk, mely lehetővé teszi jelölőnyelvi kódban tárolt nyelvi adatbázisok fogadását és kezelését.

A jelölőnyelvek alkalmazásáról szóló döntés eredményeként szótári rendszereinket és szótárszöveg-feldolgozási módszereinket is át kellett alakítanunk. Döntésünket megkönnyítette, hogy a jelölőnyelvi adatbázisok alkalmazásának számos előnye van:

- a jelölőnyelvvel leírt adatbázisok megkönnyítik a szótári adatbázis egyben tartását, egységes kezelését, karbantartását, az anyag bővítését, fejlesztését;
- a szótári adatbázisok egységes, könnyen reprodukálható szerkezetben (SGML, XML) való tárolása lehetőséget teremt arra, hogy munkánkba más is bekapcsolódhasson;
- az egységesen kódolt adattípusok (jelölőnyelvi címkék, elemek) egyszerűbbé teszik új programfunkciók kialakítását (például új keresési, megjelenítési feladatok megoldását);
- az egységes szerkezettel, illetve annak kialakításával könnyebben felismerhetővé és megoldhatóvá válnak a szerkezeti hibák, tévesztések, hiányok, elírások stb.;
- az egységes szerkezetű jelölőnyelvi adatbázis az elektronikus szótárkiadás mellett papírszótárak szerkesztésére, egységesítésére, kiadására is alkalmas, jól definiált szerkezeti egységeinek egyszerű kezelhetőségével (keresés, legyűjtés, stb.) forrásául szolgálhat különböző nyelvi kutatásoknak;

Az új technológia megteremtette új szótárkészítő- és feldolgozó programok, új módszerek létrehozásának igényét is. Legfontosabbnak egy olyan eszköz megteremtése látszott, mely emberi segítséggel képes egy nem jelölőnyelvben kódolt szótáradatbázist gyorsan, olcsón jelölőnyelvi adatbázissá alakítani, szerkezetileg pontosan reprezentálva a létrejövő dokumentumban az adott szótár szerkezetét. Erre azért volt szükség, mert olyan szótárak feldolgozásának és szótárprogram alá illesztésének lehetőségét is szerettük volna megteremteni, melyek még csak hagyományos, tipografikus formájukban, ún. papírszótárként léteztek, illetve amelyeknek elektronikus formája és szerkezete nem, vagy nem sokban tért el egy tipografizált szótárétól. Az általunk feldolgozásra és megjelentetésre érdemesnek tartott magyarországi szótárak többsége még csak ilyen formátumban létezett.

A jelölőnyelvek alkalmazására való áttérés még egy megoldandó feladatot adott. Meg kellett határoznunk a szótárleíráshoz használt nyelvi formalizmust. Mivel az elektronikusan tárolt szövegek, ezen belül is a szótárszerkezetek leírására már létezett

---

[32] Jelölőnyelv: markup language; számítógépes tartalmak leírására szolgáló ún. tartalomleíró nyelvek, melyek nemzetközi formátumszabványok alapján egységes jelekkel írják le egy adott tartalomról az önmagán túlmutató információkat: a kódolt tartalom típusát adott osztályozási rendszerben (adatszerkezetben), helyét, szerepét egy nagyobb tartalmi egységben, a tartalom formázási információit, stb. A világhálón szereplő elektronikus tartalmak egységes leíró nyelve, a HTML mellett ide tartoznak – többek között – az általunk szótári adatbázisok kódolására alkalmazott SGML (Standard Generalized Markup Language) és XML (eXtensible Markup Language) nyelvek is. Cikkünkben a továbbiakban a jelölőnyelvekre való hivatkozással az általunk alkalmazott SGML és XML nyelvekre utalunk.

jelölőnyelvi ajánlás, elhatároztuk, hogy amennyire ez lehetséges, a szótárszerkezetek kialakításában igazodunk a TEI[33] által kidolgozott javaslatokhoz.

A cikkben néhány, általunk feldolgozott szótárból[34] közlünk részleteket, példákat, illetve hivatkozunk a bennük tapasztaltakra.

2. Kísérletek a szótári szöveg jelölőnyelvi adatbázissá alakítására

2.1 A lineáris feldolgozási elmélet

Egy szótár sajátos szerkezete a szótárt alkotó szócikkekre vonatkozó szabályosságok elemzésével mérhető fel és alkotható meg, ezért a szócikket határoztuk meg leendő feldolgozóprogramunk legnagyobb felismerendő és feldolgozandó egységeként. Ennél nagyobb adatszerkezeti egységeket már nem kellett felismernie a programnak illetve tükröznie a létrehozandó jelölőnyelvi adatbázisnak.

Az elmélet, melyre leendő eszközünk működését felépítettük, annak felismerésére épült, hogy bármely jól szerkesztett papír- illetve tipografikusan leírt elektronikus szótár már önmagában hordozza szerkezetét, melyet a jelölőnyelv alkalmazásával tulajdonképpen csak láthatóvá teszünk. Ezt a szerkezetet a szótár tipográfiai megformálásával (betűtípusok, grafikai jelek, központozás, szövegtagolás stb.), valamint kötött és szabad tartalmainak[35] felismerhető sorrend – rendszer – szerinti szerepeltetésével jeleníti meg.

A szótárak szerkezetét tehát – adott szótár szócikkeire nézve – kötött elemsorrendű struktúrának tekintettük, melyben a tartalommal bíró elemi alkotórészek, egységek sorrendisége határoz meg nagyobb szerkezeti egységeket. Ezek a nagyobb egységek alakítják ki a szótári szócikk általános szerkezetét. Ezt az általános struktúrát kell az embernek leírni úgy, hogy a szerkezet adott pontjaihoz rendelt tipográfiai információk segítségével egy program felépíthesse a kívánt szövegstruktúrát. Feldolgozási elméletünk szerint egy szótár általános szerkezete alkalmas arra, hogy tartalmát a

---

[33] A TEI – Text Encoding Initiative – egy nemzetközi tudományos projekt, melynek célja megteremteni egy egységes elektronikus kódolási rendszert összetett szerkezetű szövegstruktúrák szerkezetének interpretálására. Ennek az átfogó munkának egyik fejezete a szótári struktúrák elektronikus kódolására ad ajánlásokat. A TEI-ről bővebben a http://www.tei-c.org, a szótári ajánlásokról a http://www.tei-c.org/P4X/DI.html webcímen olvashatunk.

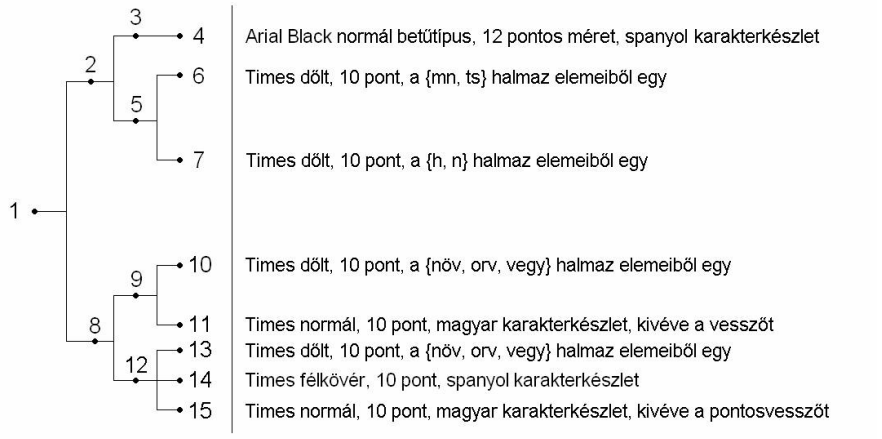
[34] Pálfy Miklós: Francia-magyar kéziszótár. Grimm Kiadó, Szeged; első kiadás: 1999
Hessky Regina: Német-magyar kéziszótár. Nemzeti Tankönyvkiadó–Grimm Kiadó; 1. kiad.: 2000
Forgács Tamás: Magyar szólások és közmondások szótára. Tinta Kiadó, Budapest, 2003
Dorogman György: Spanyol-magyar kéziszótár. Grimm Kiadó, Szeged; 2., jav. kiadás: 2002
Német-magyar Üzleti Nagyszótár (Hamblock/Wessels/Futász). Tudex Kiadó, 2001-2004
Angol-magyar és Magyar-angol Nagyszótárak. Akadémiai Kiadó, Budapest, 1998
Német-magyar és Magyar-német Nagyszótárak. Akadémiai Kiadó, Budapest, 1998

[35] Kötött tartalmúnak nevezem egy szótárszöveg szerkezeti elemei közül azt, melynek lehetséges konkrét, teljes tartalmai, vagy a teljes tartalmat lefedő résztartalmai egy jól definiálható, zárt karakterlánc-halmaz elemeiként meghatározhatóak, ilyen például a szófajkészlet, a nyelvtani nem jelölése, vagy a szakjelzetek. Szabad tartalmú az elem, ha tartalmaira konkrétan nem, csak karakterosztályokkal, karakterkészletekkel hivatkozhatunk.

szócikk első tartalommal bíró elemétől – általában a címszótól – kezdve, elemről elemre, sorrendben, kihagyások nélkül feldolgozhassuk úgy, hogy közben az elemek és elemhatárok tipográfiai kódjait használjuk fel a szerkezetkialakító program vezérlésére[36]. Az egyes szerkezeti elemekhez rendelt, előre felmért és definiált tipográfiai tulajdonságokat az elemzendő szövegrész tulajdonságaival összevetve ismeri fel a program, hogy a szótári szócikk elemzésének adott pontján milyen elemek kialakítására van lehetősége.

Rendszerünk adatleíró modelljét a következő, tipografizált szócikkek alatti faszerkezetben ábrázoljuk, feltüntetve, hogy elemző és szerkezetkialakító rendszerünknek milyen információkat kell ebből a szövegből kinyernie és tárolnia:

cloro *h, vegy* klór
clorofila *n, növ* klorofill, levélzöld
coactivo *mn* kényszerítő
coartar *ts* akadályoz, korlátoz
cobalto *h, vegy* kobalt; **azul de ~** kobaltkék; *orv* **bomba de ~** kobaltágyú



1. szócikk	5. nyelvtani csoport	9. jelentéscsoport	13. szakjelzet
2. szócikkfej	6. szófaj	10. szakjelzet	14. kifejezés
3. címszóforma	7. nyelvtani nem	11. jelentés	15. kifejezésjelentés
4. címszó	8. szócikktest	12. kifejezéscsoport	

1.ábra. A példaszócikkek mindegyikére érvényes szótári adatszerkezet fastruktúrában ábrázolva, a levélszintű elemekhez tartozó, a rendszer által tárolandó információkkal

A szótári szócikk szerkezetét megjelenítő faszerkezetben a szövegtartalom a fastruktúra legutolsó elemein, levelein helyezkedik el. A csomópontok képviselik a kisebb, hasonló szerepű egységeket átfogó szerkezeti elemeket. A szótár tagolóelemeit – a központozást, grafikus szimbólumokat, jeleket stb. – a fabejárás döntési pontjainak, elágazásainak (ÉS/VAGY ágak, elemismétlődések, opcionális

[36] Erre a szótári tartalmakat sorrendben, kihagyások nélkül feldolgozó módszerre hivatkozom lineáris feldolgozási elméletként.

elemek) reprezentánsaként képzeltük el. Ezek jelölik a szótárban az egyes szerkezeti egységek közötti váltást. Tehát a fa elemei (a levelek és az elágazások, melyek sokszor reprezentálnak csoporthatárokat is) modellünkben nemcsak az adott, konkrét egyedi tartalmakat kell hordozzák, hanem minden egyéb olyan általános információt, szabályosságot is, mely a szótárban megjelenített szövegrészek sajátja. Például jól látható, hogy bár a fában minden lehetséges szerkezeti elemet ábrázoltunk, a levélelemek nem mindegyike kötelező. Ennek az információnak is meg kell jelennie a rendszerben. Erre elképzelésünk szerint külön szerkezetleíró állomány szolgált.

Első lépésben a részletes szövegvizsgálat után ebben a szerkezetleíró fájlban határozható meg a dokumentum általános szerkezete. Ezután minden levélelemhez meg kell határozni a szótárban hozzá rendelt tipográfiai és tartalmi információkat – a betűtípust, a kötött tartalmakat, a karakterkészletet vagy karakterosztályokat, illetve bizonyos, az elem tartalmából kizárható karaktereket, karaktercsoportokat. A szerkezetleírásban kell meghatározni a különböző szövegelemek szomszédsági viszonyainak szabályosságait is, például a központozás, szimbólumkészlet azon elemeit, melyek két vagy több elem illetve elemtípus határán kötelezőek.

A szótárt ezek után a szócikkszerkezethez kialakított faszerkezetben fentről lefelé haladva kell feldolgozni úgy, hogy a szerkezetelemzés a fa minden levelét érintse. Az elemző pedig a szerkezet adott helyén eldönti, a szöveg következő feldolgozandó része megfelel-e a szerkezetben soron következő lehetséges elem (levél), illetve szomszédsági viszony formai követelményeinek[37]. Olyan eszközt kellett találnunk, mely képes felismerni, azonosítani egy szöveg tipográfiai megjelenését, és azt képes a tartalommal együtt tárolni. Ezután egy szövegelemzési és szerkezetkialakítási szakaszban a szerkezeti elemekhez előre definiált tipográfiai feltételek, valamint a feldolgozandó szöveg tartalmi tulajdonságai alapján az elemzendő szövegen végigfut, és megalkotja a kívánt szerkezetű jelölőnyelvi adatbázist. Mindezt úgy, hogy lehetőségünk legyen a munka bármely fázisában az előre definiált szerkezet, a tipográfiai feltételrendszer vagy az elemzendő dokumentum módosítására.

2.2 A Yacc-Lex programpáros

Erre a célra megfelelőnek látszott egy tipográfiakivonatoló segédprogram és a Yacc-Lex programpáros együttese. Első lépésben a teljes tipografikus adatállományból egy olyan dokumentum készült, melyben .txt formátumban együtt szerepeltek a tartalmak és a tartalmakhoz tartozó stílusinformációk, tipográfiai tulajdonságok is. Így lehetővé vált, hogy a csak szövegformátumú állományokon dolgozó Lex és Yacc számára feldolgozható anyagot szolgáltassunk.

A Lex program számára a szótár tartalmát karakterosztály-definíciókkal, illetve kötött tartalmú szerkezeti elemek meghatározásával ún. tokenekre bontottuk, melyek alapján a Yacc programmal és saját kódrészletekkel elértük, hogy a létrejövő C++ programkód adott szerkezetű XML-lé kódolja a szövegben a neki megfelelő formalizmusban szereplő szócikkeket. Hamar kiderült azonban, hogy ezzel a módszerrel sokszor nehézkes, sőt, néha további szövegátalakítás nélkül lehetetlen XML-

37 Más szavakkal: az egyes tartalmas elemek azonosított, egy adott elemre nézve elméletben egységes tipográfiai, sorrendi, szomszédsági, opcionalitási jellemzőit információként használjuk arra, hogy egy dokumentumot feltérképező program a szótárszerkezet felderítésében éppen adott szócikk mely elemének felismerésénél tart.

adatbázis létrehozása. A szerkezetelemző a feldolgozandó szövegegységben (szócikkben, entryben) ugyanis mindig csak a feldolgozási folyamatban éppen soron következő feldolgozandó elemi szövegegységet látja. Ha a dokumentum faszerkezetében az elemző döntési ponthoz jut, ott a soron következő szövegegységet (karaktereket) mindaddig feldolgozza, míg annak tulajdonságai megfelelnek a faszerkezetben egy adott döntési pont után levélszinten várt elem tartalmi és formai megkötéseinek. Arra azonban képtelen, hogy az elemzendő szövegrész környezetéből további következtetéseket vonjon le. Egy ilyen elemzőt pusztán a szótárban eredendően meglévő tipográfiai és tartalmi megkötések alaposan félrevezethetik. Gondoljunk csak egy szótár sokszor igen bonyolult szerkezetére, az azonos tipográfiájú, de eltérő tartalomtípusú elemekre, melyek a szótár legkülönbözőbb helyein megjelenhetnek. A különböző nagyobb szerkezeti egységekben megjelenő ismétlődő elemek illetve részszerkezetek ugyanígy a programozott szerkezetértelmezés többértelműségének forrásai lehetnek. Adott esetben olyan szócikkszerkezet is létrejöhet az ilyen, csak a következő elemzendő egységet látó és feldolgozó elemzővel, mely megfelel ugyan az elemzési sorrendben várt összes tipográfiai információnak, a szerkezet mégis hibás lesz: nem a szótár valódi tartalmát és szerkezeti egységeit tükrözi. Az eszköz legfőbb korlátja az volt, hogy az elemzés azonnal megakadt, amint a soron következő adatsor nem illeszkedett ahhoz a formai feltételrendszerhez, melyet az elemző az elemzési fában általa éppen bejárt helyen várt. Ilyenkor a program már nem próbálkozott a szerkezet más módon történő kialakításával.

A problémára megoldást kínált a szótár *egészét* ellátni vezérlőkarakterekkel, mégpedig *minden* olyan döntési ponton, ahol a tartalmi és tipográfiai feltételek önmagukban nem voltak elegendőek az elemző kiszolgálásához. De ez igen nehézkessé és lassúvá tette a munkát. Olyan rendszerre volt szükségünk, mely a lehető legkevesebb emberi előfeldolgozással képes egy szótár jelölőnyelvi adatbázissá alakítására.

Egy új szerkezetelemző és -kialakító program megalkotásának első lépéseként összefoglaltuk a YACC-LEX rendszer számunkra meghatározó korlátait:

- csak a következő feldolgozandó tipográfiai egységig lát előre: amennyiben az megfelel a várt formai követelményeknek, azonnal és véglegesen létrejön a következő szerkezeti elem; minden egyéb esetben elakad az elemzés
- nem képes egy kérdéses elemzési ponton annak környezetéből nyert információk alapján dönteni, hogy adott ponton választható több szerkezeti elem melyikével folytassa a jelölőnyelvi adatbázis szerkezetének felépítését
- sikeresen elemzettnek minősíthet szócikkeket úgy is, hogy az adatbázis adott egysége (egy szócikk) a megadott tipográfiai szerkezetleírás szerint többértelmű, és a program nem a helyes szerkezetet választotta; a program ilyenkor *nem jelzi* a szerkezeti többértelműséget
- csak szöveges bemenetet tud fogadni; így speciális karakterek illetve kevert kódlapok esetében nincs biztosítva a tartalom egységes átkódolása – esetlegesen tartalomváltozást, tartalomvesztést okozva

Ezekre a hiányosságokra kellett egy létrehozandó új programnak megoldást találnia.

2.3. A MarkUpWizard, egy új szerkezetelemző és -kialakító program

Az előző fejezetben vázolt problémákra a cégen belül fejlesztett MarkUpWizard (jelölőnyelv-varázsló) program[38] jelentett megoldást. A program legfontosabb adottsága, hogy adott formális nyelvtan (előre definiált dokumentumszerkezet-leírás, valamint tipográfiai és tartalmi feltételek) szabályainak eleget tevő RTF dokumentumot képes XML dokumentummá alakítani. (Ez azt is jelenti, hogy képes minden olyan adatállományt feldolgozni, melynek formátuma tartalmi és formai információvesztés nélkül RTF-fé konvertálható.)

Az alkalmazás szükségtelenné tette tipográfiakivonatoló program használatát. A felhasználónak két feladata van: a szótár szerkezetét leíró formális nyelvtan elkészítése, és a szerkezeti elemekhez rendelt tartalmi és tipográfiai tulajdonságok leírása. Ezek elkészítésével el lehet kezdeni a feldolgozandó dokumentum átalakítását. A program UNICODE karakterkódolás használatával küszöböli ki a nem kívánt adatvesztést, és adatváltozást.

Az elemzés eredménye első próbálkozásra szinte biztos, hogy nem tökéletes. Elég, ha csak a szótárakban jelen levő formai, sorrendi stb. hibákra gondolunk. A munka elején azonban a felhasználó legtöbbször még nem méri fel és írja le tökéletesen a szótár szerkezetét. Így az elemzés eredményétől függően a nyelvtan és/vagy az RTF dokumentum módosítása után a második vagy mindkét lépést meg kell ismételni, egészen a kívánt eredmény eléréséig. (Tehát az elemzési eredmények hatással vannak a szótári tartalomra.) További fontos tulajdonsága a programnak, hogy egy szócikk elemzése közben az éppen elemzésre következő tipografikus adatból kialakított szerkezeti elemről nem dönti el azonnal, vajon az megengedhető-e a szótár szerkezetében. Amikor az elemző a feldolgozandó bemenetben olyan döntési ponthoz jut, melynél az elemzés folytatását többféleképpen[39], vagyis egymástól eltérő szerkezeti egységek kialakításával is folytathatja, kiválasztja az egyik lehetőséget, és a dokumentumban sorrendben később következő elemek tipográfiai információi igazolják vagy cáfolják a választás helyességét. Amennyiben a több lehetőség közül választott szerkezeti elem után következő adatok nem igazolják a választást és az elemzés elakad, az elemző visszalép a kérdéses pontig, és adott ponton megengedhető, addig még nem próbált szerkezeti elem létrehozásával kísérli meg érvényes szerkezet létrehozását. Amennyiben több ilyen módon megalkotható teljes szócikkszerkezet létezik, az eredményfájlban a szerkezet többértelműségére megjegyzés figyelmeztet. Ha csak egy, a kívánt formai követelményeknek eleget tevő teljes elemzés születik, az helyesnek minősül. Amennyiben az elemzés sikertelen, a program a leghosszabb, még elemezhető részfát XML-ben az eredményfájlba írja, a nem elemzett részeket pedig szabványos jelölőnyelvi megjegyzésben szerepelteti. A lezáratlan, de megkezdett elemeket a program lezárja. Az elemzési eredményeket mindvégig egy állományban tartja, ugyanabban a fájlban épülnek fel a helyes szerkezetű szócikkek, ugyanitt jelennek meg a hibaüzenetek és a többértelműségre utaló megjegyzések is. Mindezt

---

[38] A programot Pál Miklós, cégünk egyik vezető fejlesztője írta

39 Általában egy szótári szócikk szerkezetében számos olyan hely található, ahol a szerkezet vagylagos struktúrákat, elemeket, elemcsoportokat enged meg. Adott bemenő adatsorban az elemzés elérhet olyan pontig, melyen az elemzésre következő adat tipográfiai jellemzői többféle, adott elemzési helyen kialakítható szerkezeti elem tipográfiai kívánalmainak is eleget tesznek. Ezt nevezzük döntési pontnak; ilyenkor egy elemzőnek választania kell, milyen elem kialakításával építi tovább a kialakítandó szerkezetet.

úgy teszi meg, hogy a létrejövő adatok szintaktikailag mindvégig kielégítik az XML-szabványt, lehetővé téve, hogy az állománnyal XML-kezelő eszközökkel is dolgozhassunk.

3 Tapasztalatok a szótár jelölőnyelvi adatbázissá alakításáig

3.1 A jelölőnyelvvé alakítás hasznai - egységes szöveg; hiányok, hibák feltárása

Ahogy az előzőekből kiderült, a szerkezetelemző program a felmért szerkezet egyre pontosabb leírásával, illetve a kiinduló dokumentum hibás részeinek javításával közelít a végleges adatbázis kialakítása felé. Mivel a szerkezet leírása igen szigorúan igazodik a szótár valódi tartalmi, szerkezeti struktúrájához, ezért a jelölőnyelvi adatbázis kialakítása közben rengeteg hibára, hiányosságra, figyelmetlenségre fény derül. A szerkezetleíró állományban pontosan definiált kötelező és opcionális elemeket, a kötött tartalmakat, az elemek meghatározott sorrendjét, a szöveg grafikai tagolását, az adott szerkezeti elemekre vonatkozó pontosan megadott tipográfiai tulajdonságokat a program a dokumentum egészén számon kéri. Éppen ezért ez a szövegfeldolgozó rendszer nagyon jól használható a szöveg egységének, tartalmi helyességének javítására is. Az elemzés közben kapott eredmények rámutatnak a hiányokra, hibákra, hiszen az elemzés akkor akad el, amikor az ember által leírt és a programon keresztül megkövetelt formai, szerkezeti feltételeknek a dokumentum nem felel meg. Így az elemzés visszahat a dokumentumra, javítva formai, tartalmi egységét, felhívva a figyelmet az át nem gondolt szerkezetre, hibákra, tévesztésekre.

Általános tapasztalatunk a szövegek feldolgozásánál az, hogy hibátlan szótár nincs! Amelyik szótárban a munka egyötödének ráfordításával a szócikkek 75 százaléka szerkezetileg helyesen létrejön, az már egy átgondolt, nagyon jól szerkesztett szótár! A fennmaradó részben találhatók azok a szerkezeti sajátosságok, problémák, hibák, melyeket további szerkezetmódosítással, vagy a kiinduló dokumentum javításával kell megoldani.

3.2 Az átalakítás tapasztalatai – hibák, megoldási javaslatok

A szótárakban többféle típushibával találkozunk, melyek nehezítik a jelölőnyelvi dokumentum létrehozását. Előfordul például elemek sorrendjének felcserélődése, a szótárban elvileg meglévő kötött tartalmak tartalmi variálódása, kötelező tipográfiai tulajdonságok átalakulása vagy elveszése, a szövegtagolás megszokott formalizmustól való eltérése. Ezek a legtöbb szótárban előfordulnak.

Annak érdekében, hogy ezek a hibák ne forduljanak elő a szövegben, előre definiálni kell a szótár szerkezetét. Érdemes próbaszócikkekkel kialakítani egy olyan, szigorú szerkezetet, melyet aztán minden szótárszerkesztő leírva kézhez kap, a kötelező elemtartalmak (szófajok, földrajzi használati körök, szakjelzetek stb.) pontos halmazelem-leírásával, az elemek kötelezőségének vagy elhagyhatóságának gondos feltüntetésével. Egy XML-szerkesztővel írt szótárban elvileg a program képes figyelni számos ilyen információra. Mégis, tapasztalatunk az, hogy egyrészt a mai szótáríró gene-

ráció még nem szívesen használ ilyen programokat, vagy, ha használ is, sokszor nem veszi figyelembe a szerkezeti elemek valós szótárbeli szerepét. Megelégszik azzal, ha talál egy olyan szerkezeti elemet, melyre nézve a tipográfiai megszorítások megegyeznek azzal a formai megjelenéssel, melyet az adott leírandó információnak elképzelt. Mivel ezek az elemek legtöbbször az elhagyható – és általában nem kötött tartalmú – elemek kategóriájából kerülnek ki, nagy esélye van különböző szerepű, de azonos tipográfiájú elemtartalmak keveredésének. Ez a keveredés azonban elkerülhető, ha a szótárszerkesztők a különböző szótári tartalmakat más-más tipográfiával jelenítik meg. Két-három fonttípus helyett érdemes négy-öt, egymáshoz illeszkedő, mégis markánsan más fontot használni.

Könnyebb a szótári szöveg egységének megőrzése, ha jól látható grafikus szimbólumok, jelek választják el egymástól a nagyobb szerkezeti egységeket. Ráadásul ennek kettős előnye van: segíti a szótár áttekinthetőségét, segít eligazodni a szótárban, és a szöveg feldolgozhatóságát is gyorsabbá, egyszerűbbé teszi. Természetesen lehet egy szótár áttekinthető a különböző grafikai elemek, sematikus ábrák, szövegegységeket tagoló szimbólumok alkalmazása nélkül is. Sokat segíthet a betűtípusok variálása, gondos megválasztása, a szöveg megfelelő tagolása központozással, új sorban kezdett nagyobb szerkezeti egységekkel, az egymástól elkülönítendő sorok szövegbehúzási tulajdonságainak megváltoztatásával. Véleményünk szerint mégis hasznos, ha a szöveg áttekinthetőségét grafikai elemek bevezetésével segítjük.

Annak igazolására, hogy legalább tartalmi és szerkezeti ellenőrzésre érdemes igénybe venni szerkezetelemzőnket, álljon itt néhány feldolgozott szótárból vett példa. Ezekkel illusztráljuk, hányféle hiba érhető tetten még azokban a szótárakban is, melyeket jól szerkesztetteknek mondunk, sőt, azokban is, melyek elvileg már XML-szerkezetben vannak.

Volt olyan szótárunk, melyet két külön formátumból kellett egységes alakra hoznunk. Ennek egyik részét egy szoftvercég kódolta jelölőnyelvi adatbázissá. Az anyag valóban megfelelt a jelölőnyelvi adatbázisokkal szemben felállított formai követelményeknek. Azonban az általánosabb szerkezeti egységek kialakításán túl, a bonyolultabb, és szerkezeti következetlenségeket is tartalmazó anyag feldolgozásánál már nem vették a fáradságot arra, hogy definiálják az elemek sorrendjét, és azt, hogy a szerkezetben kötelezőek-e vagy sem. Így fordulhatott elő, hogy több száz (!) helyen hiányzott a forrásnyelvi kifejezések magyar fordítása, amit a szerzővel, több heti munkával utólag kellett pótolnunk.

Íme még egy fontos tapasztalat: nem mind jelölőnyelvi adatbázis az, ami annak látszik! Egy adatbázisnak nem csak formai követelményeknek kell eleget tennie, pontosan tükröznie kell az eredeti dokumentum tartalmi és szerkezeti sajátosságait is! Mivel feldolgozási módszerünk egyik alapkövetelménye, hogy a szótárban fellelhető szerkezetet olyan szigorú szerkezetben írjuk le, amennyire az magának a szerzőnek koncepciója volt, így nálunk nem fordulhat elő, hogy ezek a hibák rejtve maradjanak.

Találkoztunk még olyan hibákkal, mint például a kötelező elemek kitöltetlenül hagyása – ez legtöbbször szófajok esetében fordult elő. Ha egy szerkezetellenőrző programban jól definiáltak a kötelező tartalmak, minden ilyen jellegű hibára fény derül. Általában véletlen hibák ezek, elfelejtődnek a kitöltendő szerkezetek.

Sokszor fordul elő egymás melletti – főleg minősítési és körülírásra szolgáló – elemek sorrendi következetlensége. Jellemző, hogy amikor fel lehetne állítani sorrendi precedenciákat, akkor sem teszik a szerkesztők ezt meg. Főleg a különböző minősítések, használati rétegek – pl. földrajzi használat, szakjelzetek, stílusrétegek –

keverednek. Nagyon sok szótár nem is tesz tipográfiai különbséget az ilyen adatok között. A feldolgozott szótárak szolgálnak megoldással is: megkülönböztethetjük az adatokat, ha például a szakjelzetek nagybetűvel kezdődnek. A minősítések jelölésére szolgálhat az adatok zárójelben (), < >, való szerepeltetése. Ezek csak ötletek. A lényeg: egymástól jól megkülönböztethető, lehetőleg minél kevesebb célra használt jelek, tipográfiai tulajdonságok alkalmazása.

Elgondolkodtató a szótárakban elvileg kötött tartalmak variálódása is. Volt olyan szótárunk, ahol a tárgyas ige megadására a következő elemek szolgáltak: *i ts, ts i, tsi, ts*. Ugyanebben a szótárban találunk példát a tartalmi ellenőrizetlenség súlyosabb esetére is. A szótárban szerepelő **lótenyésztő** szó szófajaként a *ló* volt megjelölve. Szintén a tartalmi ellenőrizetlenség példája a következő: Ha a hozzánk érkezett elektronikus anyag változatlan formában jelent volna meg papíron, a **szovhoz** szó szófaja 'szocialista realista főnév' (*fnszoc.r.*) maradt volna. Jó példa ez azonos tipográfiai tulajdonságú, eltérő szerepű elemek 'összeolvadására'. A papírváltozatban természetesen már a helyes, *... fn; szoc.r. ...* szerkezetű adat szerepelt. Ez utóbbi két példát egy – elvileg – XML-szerkezetű szótárból gyűjtöttük. Látható, hogy még az ilyen, szerkesztettnek mondott szótáraknál is összekeveredhetnek tartalmak, pusztán tipográfiájuk hasonlósága vagy azonossága miatt. Ez is bizonyítja, hogy mennyire fontos a kötött tartalmak valódi ellenőrzése. Természetesen előfordul olyan eset is, hogy a tartalom helyes, de nem a neki megfelelő tipográfiai megjelenítést kapja. Mivel elemzőnk egyszerre értékel formai információkat, és a szövegegységek kialakítandó szerkezetben elfoglalt helyét is, az ilyen hibák sem maradnak rejtve.

A szerkezet kialakítása közben is érdekes dolgokat tapasztalhatunk. A következő sematikus szócikkrészlet példa a felesleges információismétlésre, megmutatva, sokszor mennyire nem átgondolt egy szótári struktúra:

<címszó> *mn*, **<címszóvariáns>** mn ...

Valóban nincs értelme *egy* szócikkben szerepeltetett több címszóvariáns mindegyikéről elmondani, hogy melléknév. Nem tartjuk szükségesnek egyedi és típushibák további felsorolását. Célunk az volt, hogy a fenti néhány példa egyértelművé tegye: a szótárak szerkezeti és tartalmi egységességének kialakításához szükség van mind szerkezeti, mind tartalmi ellenőrzésre.

Végezetül álljon itt még egy általános tapasztalat. A szerkezetkialakító munka vége felé sokszor találkoztunk a következő problémával: a néhány, még nem elemzett szócikk szerkezete megkövetelheti újabb elemek szerkezetbe vételét. Néha valóban szükség van a szótárstruktúra leírásának bővítésére, de legtöbbször az ilyen 'új' szerkezetelemek bevezetése felesleges. Nagyon valószínű, hogy inkább a szótár átgondolatlanságáról van szó, és a közölni kívánt dolgokat le lehetne írni más, a struktúrában már meglévő elemmel. Előfordul persze ellenpélda, de ha egy elem csak öt-tíz, vagy kevesebb esetben fordul elő, érdemes gyanakodni arra, hogy a szerző másként, egységesebben, egyszerűbben is leírhatta volna ugyanazt.

4 A nyelvi egyértelműsítés, előkészítés az intelligens kereshetőségre

4.1 A nyelvi egyértelműsítés szerepe, fontossága

A kívánt szerkezetű XML létrejötte után kezdődik az a munkaszakasz, amelyik már a kerestetni kívánt szövegtartalmak intelligens megtalálásáért felelős. Tudjuk, hogy a papírszótárak, főleg helytakarékosság és a fölösleges ismétlések elkerülése érdekében élnek a szövegbehelyettesítés (tilde), szövegrövidítés (perjelek, per-kötőjelek), jelentésrész-összevonás (perjelek, elhagyható zárójelek) módszerével. Míg azonban az emberi olvasásra szánt ilyen jellegű szótárakban a felhasználó – általában – intuitív módon be tudja helyettesíteni az adott, hiányzó szövegrész(eke)t, értelmezni tudja a sűrített, rövidített tartalmakat, addig a számítógép önmaga nem képes értelmezni a szótárakban ilyen megoldással előforduló tartalmak összetartozó egységeit. Így, előzetes értelmezés, segítség nélkül nem is lenne képes arra, hogy valóban kereshessen az ilyen tartalmakban. Ezt a jelentésegyértelműsítést, jelentésfeloldást, jelentésszöveg-dekódolást végzi el az adatbázison a nyelvi normalizálás munkafázisa.

4.2 A nyelvi egyértelműsítés munkafázisai, megoldandó feladatai

A nyelvi egyértelműsítés első szakaszában, a normalizálásban ezeket a szöveget rövidítő jeleket oldjuk fel, és értelmezzük. Az összetartozó szövegelemeket egymáshoz illesztjük, és a már egységbe fogott, egyértelműsített szövegegységeket a dokumentumban nekik megfelelő, rövidített szövegegységek mellé illesztjük. A szövegkeresések a továbbiakban az egyértelműsített szövegekhez fordulnak, míg a szövegek megjelenítésére továbbra is az eredeti szövegek szolgálnak.

Az intelligens keresést segítő következő munkafázis már a keresendő szótári tartalom nyelvi háttéradatokkal, szótövekkel való kiegészítése. Ezt a munkaszakaszt azonban, mivel felmerülő nehézségei, problémái nem szótárszerkesztési problémák, a következő, tapasztalatokat, javaslatokat tárgyaló részben már nem érintjük.

4.3 Hibák, tapasztalatok – hogyan tud segíteni a szótáríró?

Minden szövegrövidítési típusra adhatók olyan általános szabályok, melyeket általában a rövidített szövegrészek többségénél alkalmazni lehet az egyértelműsítésre. A tilde karakter feloldásakor például általában a címszó a rövidített, behelyettesítendő tartalom. Volt azonban olyan szótár, melyben egy szócikken belül sem volt egységes az alkalmazása: hol a teljes szócikket kellett behelyettesíteni, hol a ragozott alakokhoz tartozó megrövidült tőallomorfot. A szövegegyértelműsítést a jelölési szokások szócikkenkénti különbözősége nehezítette. Tehát jól át kell gondolnunk a szövegrövidítés alkalmazott módjait. Ha nem egységesen használjuk ezeket, automatikusan nem lehet a teljes tartalmat reprodukálni. Márpedig egy szótárfeldolgozó rendszerben ez lenne a cél.

Az alapelvek a szövegrövidítés bármely formájával kapcsolatban ugyanazok: úgy használjuk őket, hogy a szótár bármely pontján ugyanazt a szövegegység-

visszaállítási algoritmust alkalmazva minden azonos típusú feloldandó tartalom egyformán, helyesen egyértelműsítve hozza létre a teljes egybetartozó szótári tartalmat. A továbbiakban bemutatjuk azt a szövegrövidítési eljárást, melynél az egyértelműsítést segítő egységes szövegkódolásra megoldást is tudunk ajánlani.

4.4. A perjelek használata – javaslat összetartozó szövegek kódolására

A perjel a szövegben felcserélhető tartalmak jelölésére szolgál. Általában két, egymással a jelentésben felcserélhető szót választ el egymástól. Sokszor azonban a perjelek két oldalán szereplő alternatívák nem csak egy-egy szót, hanem egész szövegrészeket fednek le. A számítógép maga nem tudja ezeket a szokásostól eltérő szerkezeteket feloldani. Álljon itt egy példa: *nem szavazó/szavazásra nem jogosító részvény.* Vajon értené-e egy a magyar nyelvvel csak ismerkedő szótárolvasó, mit jelent a *nem szavazó nem jogosító részvény,* vagy a *nem szavazásra nem jogosító részvény*? A problémára több megoldás is létezik. El is kerülhetjük az ilyen szerkezetek leírását, de segíthetjük is a szótárolvasót, amint az a következő, nyomtatásban megjelent szótári szócikkrészletben is látszik:

{arcára fagy/ráfagy az arcára/lefagy az arcáról} a mosoly arcának (...) érzelmi megnyilvánulása hirtelen megszűnik ...

Itt az összetartozó tartalmak egyértelműen egymáshoz rendelődnek. Nemcsak a szöveg feldolgozását segíti az ilyen megoldás, de az olvasó is könnyebben eligazodik a szövegben. Természetesen fenti példa csak javaslat, több, általunk is látott megoldás lehetséges az összetartozó szövegek egymáshoz rendelésére.

A nyelvi egyértelműsítés a megfelelő szabályok megalkotásával és betartásával egyszerű, sablonok mentén végezhető munka. Minden szövegrövidítési eljáráshoz megtalálható egységes szövegösszerendelő formalizmus, egyértelműsítési filozófia. Ezek közül néhányat tartalmilag is megjeleníthetünk a feldolgozandó szövegben, segítve a gépnek a szótárszöveg feldolgozásában. Azokat, melyek a szótárfelhasználó számára is egyértelműbbé, átláthatóbbá, érthetőbbé teszik a szótárszöveget, érdemes a szótár felhasználóknak szánt formájában is szerepeltetni.

5. Összegzés

Cikkünk összegzéseként elmondhatjuk, hogy a szótáradatbázisok minőségén, tartalmi egységességén sokat javít szerkezetelemzési módszerünk, legyen az tipografizált formában írt, vagy XML-szerkezetben lévő szótár. A feldolgozás eredményeként a szótár helyes szerkezetben, egységesen, áttekinthetően kerül a felhasználók kezébe. A kialakított jelölőnyelvi adatbázissal pedig kihasználhatók mindazok az előnyök, melyeket cikkünk elején már felsoroltunk.

„tök jó, de nincsenek benne csúnya mondatok”[40]

egy WAP-alapú szótári rendszer üzemeltetésének tapasztalatai

Földes András

MorphoLogic Kft; Orbánhegyi út 5. 1126 Budapest
lafoldes@morphologic.hu

Kivonat: A MorphoLogic 2003 nyara óta működteti a MoBiWAP szótári rendszert. A MoBiWAP nem más, mint a jól ismert MoBiDic szótári rendszernek a WAP lehetőségei szerint átalakított változata. A program a kezdetektől fogva naplózza a kéréseket. A cikk következtetései ennek a naplófájlnak a vizsgálatán alapulnak. A keresett szavak meglepően nagy hányada obszcén. Az adatok elemzése a jelenség okainak kutatásához próbál támpontokat nyújtani. A szokványos statisztikákon túl, az egyszerre keresett szavak közti kapcsolatokat egy skálafüggetlen gráfban is ábrázolom. Az így kapott koincidenciagráf a szavak szemantikai kapcsolatainak a feltárásában is segíthet. A kapott adatok hozzájárulnak a felhasználói elégedettség növeléséhez.

Figyelmeztetés

Az alábbi cikk természetéből fakadóan nagy számban tartalmaz obszcén, pornográf és más offenzív kifejezéseket. A trágár szavak esetenként olyan sűrűn és nagy mennyiségben fordulnak elő, hogy mindennemű körülírás, helyettesítés vagy „kipontozás” a szöveg érhetőségének a rovására ment volna.
A cikk jórészt egyedi azonosításra alkalmatlan, összesített statisztikai adatokat mutat be. Abban a néhány esetben, ahol a mondandóm illusztrálásra egyedi szövegeket használok, azokból minden, esetlegesen a személyes azonosításra alkalmas adatot eltávolítottam.

A MoBiWAP rendszer

A MorphoLogic 2003 nyara óta működteti a MoBiWAP szótári rendszert. A MoBiWAP nem más, mint a jól ismert MoBiDic[41] szótárunknak a WAP lehetőségei szerint átalakított változata.

[40] Egy a sok ezer beérkezett vélemény közül

A felhasználók a lekérdezés nyelvi irányának megadása után (az alapértelmezés a bármi-bármi „nyelvpár”, ekkor a keresés az összes lehetséges nyelven történik) beírják a keresendő szót, vagy többszavas kifejezést és elindítják a tényleges keresést. A keresés eredménye vagy a keresett szóhoz tartozó szócikk, vagy „kifejezésben keresés” esetén minden, a szót tartalmazó kifejezés-szócikk. A többi MoBiDic szótárhoz hasonlóan, a keresés előtti szótövesítésnek köszönhetően ragozott alakban is megadható a keresendő szó.
Túl sok találat esetén az eredmények között lapozni lehet, ha pedig nincs találat, akkor a szótár felkínálja a keresett szó ábécésorrend szerinti környezetét.
A **Vélemény** menüpontban a felhasználók értékelhetik a szótár működését, és javaslatokat tehetnek a továbbfejlesztésre.

1. ábra: A MoBiWAP használatának fontosabb lépései

A rendszerben jelenleg egy kéziszótár méretű, mindkét irányból kereshető angol–magyar és német–magyar középszótár, illetve magyar szinonimaszótár működik. A szócikkek méretét a WAP igényeinek megfelelően csökkentettük.
A MoBiWAP része a Pannon GSM WAP-portáljának[42], ennek köszönhetően eddig közel három és fél millió kérést szolgált ki; véleményből eddig több mint hatezer érkezett.

A keresések naplózása

Az összes szótári keresés adata egy naplófájlba kerül. A fájl (jelen cikk szempontjából érdekes) mezői a következők:

1. táblázat: A naplófájl rekordleírása

Mező	**Lehetséges értékek**
Időpont	éééé-hh-nn óó:pp:mp
Telefontípus + WAP verzió	szöveg
Lekérdezési mód	0 (keresés); 1 (lapozás); 2 (környezet)
Keresendő szó, kifejezés	szöveg
Keresési mód	0 (szó); 1 (kifejezés részeként)
Forrásnyelv	0 (bármi); 1038 (magyar); 2057 (angol), 1031 (német)
Célnyelv	0 (bármi); 1038 (magyar); 2057 (angol), 1031 (német)
A találatok száma	szám

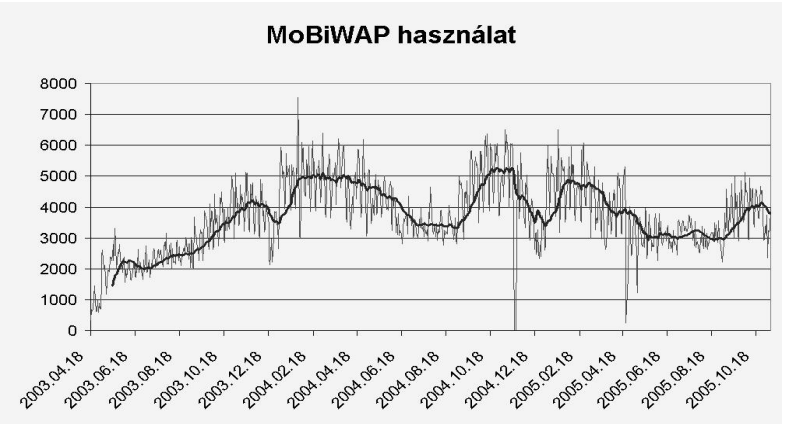
A fenti szerkezetű naplófájlt egy MySQL adatbázisba töltöttem be. A cikk ennek az adatbázisnak az elemzése során nyert adatokra épül.

[41] A MoBiDic rendszer leírása megtalálható a www.morphologic.hu oldalon.
[42] A MoBiWAP a **Hasznos/Szótár** menüpontban érhető el. Más rendszerekből a www.mobidic.hu/scripts/mobiwap.exe. címen lehet hozzáférni.

Üzemeltetési adatok

Időbeli eloszlás

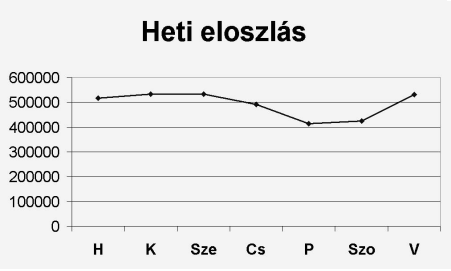
A vizsgált időszakban (2003.04.18. – 2005.11.07.) közel három és félmillió rekord került naplófájlba. Ilyen mennyiségű adat egyedülálló lehetőséget nyújt a szótárhasználat alapos statisztikai elemzéséhez.



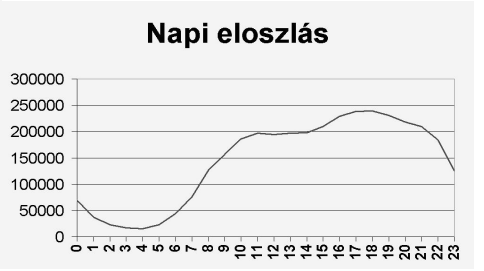
2. ábra

Az ábrán a napi hozzáférések száma látható a teljes időszakban. A kezdeti felfutási időt leszámítva átlagosan napi 3-5 ezer kérés érkezik. A szótárhasználat mérsékeltebb a nyári szünetben, karácsony és újév között**.**

A keresések számának 2005-ös kismértékű általános visszaesése feltehetőleg a Pannon portál átrendezésével magyarázható.
A szótárhasználat heti periodicitása is jól követhető (3. ábra). A forgalom pénteken és szombaton esik vissza, ekkor a potenciális fiatal felhasználók nagy része nyilván nem idegen szavak után kutat a neten.



3. ábra

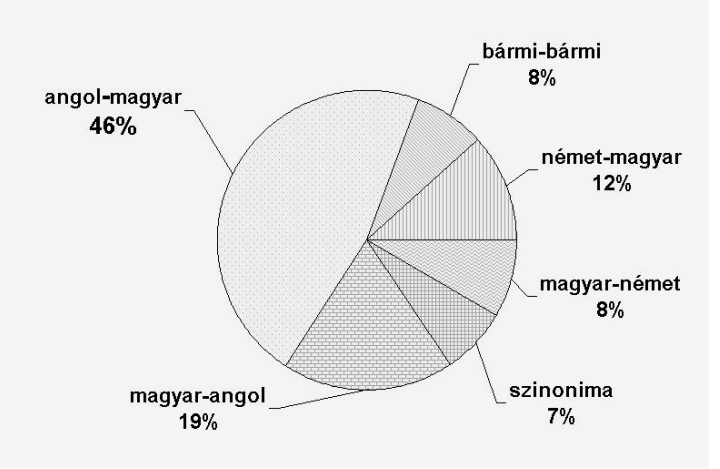


4. ábra

A napi forgalommegoszlás (4. ábra) érdekes módon a felhasználás esti emelkedését mutatja.

Nyelvek szerinti eloszlás

A forgalom nyelvenkénti megoszlása (5. ábra) megfelel az idegen nyelvek iránti magyarországi érdeklődésnek.[43] (További nyelveket még kevesebben igényelnének.) A szinonimaszótár 7%-os adata egy hosszú üzemszünet miatt alacsonyabb a valóságos igénynél.

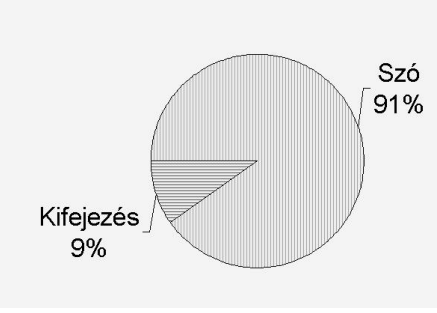


5. ábra: a keresések nyelvi irány szerinti megoszlása

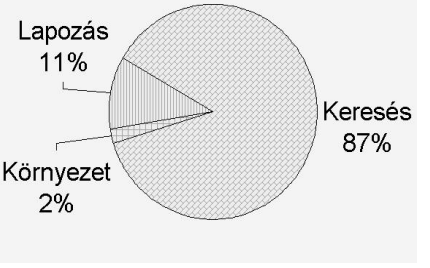
Meglepő, hogy viszonylag kevesen élnek az alapértelmezett nyelvbeállítással (bármi-bármi) ehelyett inkább - egy plusz lépésben - kiválasztják a keresett nyelvi irányt is.

A felhasználói felület kezelésének adatai

A felhasználói felület funkcióinak kihasználatlanságát illusztrálja az alábbi két ábra is. A felhasználók túlnyomó része nem él a kifejezésben keresés, illetve a szótár tartalmában való navigálás lehetőségével.



6. ábra



7. ábra

[43] A MorphoLogic és más kiadók szótáreladásai is hasonló eloszlást mutatnak.

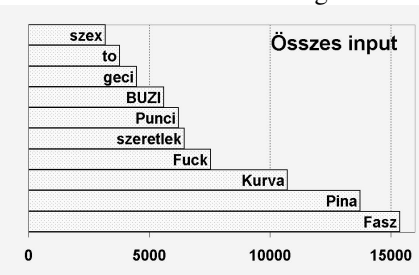
Szolgáltatásunkkal sok olyan embert is elértünk, akinek nem igazán volt még szótár a kezében. Az ő esetükben összemosódik a szótár, a fordítógép, az idegen szavak szótára, a nagylexikon, és a mindent meghallgató beszélgetőtárs fogalma. Mindezt jól illusztrálja a következő néhány „szótári lekérdezés”:

```
Ez a terület,ahol semmi változás nem tapasztalható az elodhöz képest.
Finally out comes a Crow, Comimg quickly to a stop.
Menjen egyenesen elore és a harmadik keresztezodésnél balra.
Bazdmeg de egy köcsög buzi vagy.Te szemét szutyok paraszt csicska.
```

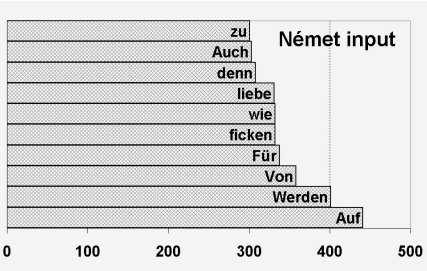
Gyakorisági adatok

Gyakoriság általában és nyelvenként

A statisztikai elemzés első kézenfekvő lépése: megkeresni a gyakran kért szavakat. A 8. ábra a teljes adatmennyiség 10 leggyakrabban keresett kifejezését tartalmazza. Megdöbbentő az obszcén szavak elsöprő többsége.
A helyzet a további pozíciókban sem igazán javul, egészen a két-háromszázadik helyig kell elmennünk, hogy „valódi” szótári kereséseket találjunk.
A hihetetlenül sok trágár szó jelenléte döntő részben nyilván nem az idegen nyelvi jelentések iránti érdeklődéssel indokolható.[44] A jelenség mindenképpen pszicholingvisztikai vagy pszichológiai magyarázatra szorul. Ehhez a későbbiekben a további érdekes adatokkal szolgálok.



8. ábra



9. ábra

Ha a szógyakoriságokat forrásnyelv szerinti megbontásban vizsgáljuk, megállapítható hogy az angol nyelv esetében némileg, a német nyelv esetében jelentősen csökken az obszcenitás túlsúlya. Úgy látszik (legalábbis a hazai WAP-felhasználók körében) a magyar és az angol trágárság is része az „általános műveltségnek”, a német nem. (9. ábra)
Elegendő mélységben vizsgálva a szógyakorisági listát a szavak három többé-kevésbé jól elkülöníthető csoportba oszthatók:

[44] Vajon ki tud elképzelni olyan szituációt, amikor valakinek égető szüksége van a “*szutyok paraszt csicska*” angol jelentésére, ezért gyorsan utánanéz WAPon?

Disznó és más szexuális vonatkozású szavak: *fasz, pina, kurva, fuck, szeretkezés, szar, segg, bitch, ficken, muschi, szeretlek, csók, hiányzol* és így tovább, oldalakon át.
Tesztszavak: Ezek gyakori vagy ritkább szavak, főleg főnevek. A felhasználó feltehetőleg nem igazán szó jelentésére kíváncsi, hanem a szótárat teszteli.[45]: *szép*, *autó, asztal, have, jó, ablak, alma, kutya.*
Ténylegesen keresett szavak: ezek között több az ige és a kifejezés, a (magyar) jelentés gyakran nehezen adható meg egy szóval, magyarul a szó több jelentésű. Ide tartozik a kifejezések szavankénti fordításából származó sok prepozíció, segédige is: *issue, cool, imaginative, distress, serendipity* illetve *for, with, have, get, take, could.*

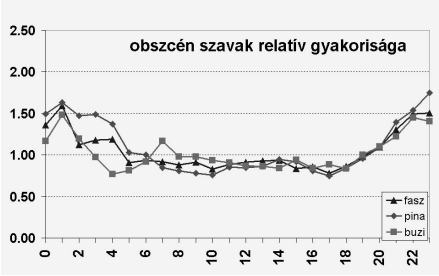
Gyakoriság időben és egyes szócsoportokra

További érdekes megállapításokat tehetünk, ha a szógyakorisági táblázatokat a teljes anyag különböző szempontok szerint kiválogatott részhalmazaira készítjük el:

Időbeli eloszlás:
A táblázat a szógyakoriság-toplistát mutatja különböző időpontokban. Látható, hogy az „élmezőny” gyakorlatilag változatlan:

idő/helyezés	0-1	6-7	12-13	18-19
1	*fasz*	*fasz*	*fasz*	*fasz*
2	*pina*	*pina*	*pina*	*pina*
3	*kurva*	*szeretlek*	*kurva*	*kurva*
4	*szeretlek*	*kurva*	*fuck*	*fuck*
5	*fuck*	*fuck*	*punci*	*szeretlek*

Az alábbi ábrán egyes obszcén szócsoportok használatának relatív gyakorisága látható az idő függvényében (10. ábra).
Látható, hogy az obszcén szavakat napközben az átlagosnál kevésbé, este és késő este viszont jóval gyakrabban keresik.



10. ábra

11. ábra

Néhány beküldött véleményből kiderül[46], hogy sokan szódolgozatok írásakor puskázásra is használják a MoBiWAPot. A jelenség ellenőrzésére megvizsgáltam a közép-

[45] Az én gyakori tesztszavaim: *alma, almafa, mosómedve, apple, raccoon, Tag*

[46] Ilyen vélemények többek közt: `Nagyon zsir ez a xotàr! Tök jòl lehet wele puskàzni!; Kirra a szotar, ezzel puskaztuk a dogankat.`

iskolai tananyagban szereplő néhány szó[47] relatív gyakoriságát (11. ábra). Az ábrán egyértelműen látszik, hogy ezeknek a szavaknak iskolaidőben átlag feletti a gyakorisága. Bár a vizsgált részhalmaz elég kicsi, kis „beleérző képességgel" a görbén talán az ebédszünet is látszik. Ugyanezen az ábrán látható az étlapokon szereplő[48] szavak csoportjának relatív gyakorisága is. A görbe mintha kötődne az étkezési időpontokhoz.

A találati arány javítása

A sok érdekes és meghökkentő megállapítás mellett a naplófájl elemzésének a legfontosabb haszna a találati hatékonyság és általában a felhasználói elégedettség növelése.

A szótár működtetésének első hónapja után elemeztem a „nincs találat" válaszok lehetséges okait.

Nyolc kategóriát különítettem el:

1. **A nagyszótárakban megvan.** pl.: *középpályás*
2. **Más nyelven a nagyszótárakban megvan.** A felhasználó létező szót írt be, de hibásan adta meg a nyelvi irányt. Pl.: *Spielen*, magyar–német irányban
3. **Gyengén ékezetesítve megvan**. Gyenge ékezetesítésen az í, ó, ő, ú, ű magánhangzók rövid párjukkal való helyettesítését értem. Pl.: *vizköpö*. Nem tartoznak ide a vegyes (néhány ékezet hibás, néhány nem), a fordított (rövid helyett hosszú magánhangzó) és a ragozott (a morfológia nem működik hibás ékezetekkel) alakok.
4. **Erősen ékezetesítve megvan.** Az erős ékezetesítés esetében az összes ékezetes magánhangzót az ékezet nélküli megfelelőjével (a, e, i, o, u) helyettesítettem. Pl.: *gyulolet*
5. **Szavanként megvan**. A beírt többszavas kifejezés kifejezésként nincs meg a szótárban, de az őt alkotó szavak egyenként igen. Pl.: *minél gazdagabb leszel*
6. **Gyengén ékezetesítve szavanként megvan**.
7. **Erősen ékezetesítve szavanként megvan**. A megfelelő esetek kombinációi.
8. **Egyéb**. Minden más eset. Ide tartoznak a fent említett bonyolultabb ékezetesítések (*A testvérem gyereket szűlt*), a helyesírási hibák *(Alles klahr, himveszo*), a kisbetűvel írt német főnevek (*zuschauer*), a nagyon hosszú beírások (*Menj a jó büdös kurva anyádba te kétszínü durva francos nagyseggü majom* és még sokkal hosszabbak is), a nem szöveges input (*1m1m1m1m1*), a tulajdonnevek (*AUCHAN*), más nyelvű kérések *(bune ziua, le roi est mort, vive le roi!*), stb.

[47] *factual, occasion, demand, fiction, suppose, attend, attic, agreement, intention, possible, fame, conclusion, deal, beggar*

[48] *soup, sirloin, tenderloin, beef, pork, mousse*

12. ábra

A hibakategóriák eloszlásának (12. ábra) ismeretében a következő javítások voltak elképzelhetők:

Technikai jellegű javítások

Megtörtént az indexek kiegészítése az ékezet nélküli alakokkal, és a kisbetűs német főnevekkel. Tovább javítható a találati arány a helyesírási hibák automatikus javításával. A többszavas inputok esetében szóba jöhet a MetaMorpho fordítószoftver alkalmazása is.

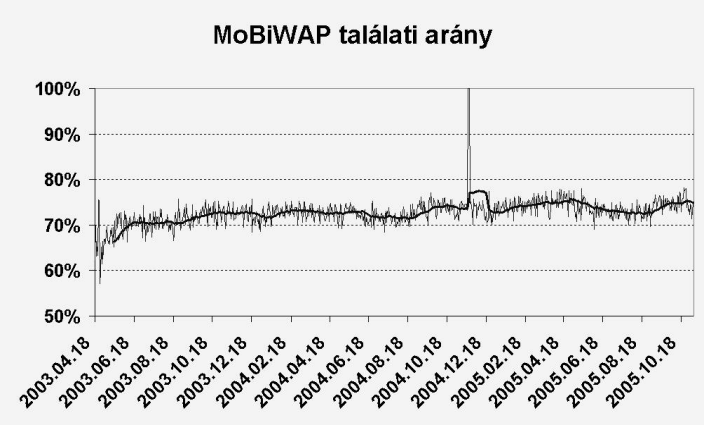
A lexikont érintő javítások

Kézenfekvő, hogy ha nagyobb a szótári adatbázis, akkor kevesebb az ismeretlen szó. Ezért megvizsgáltuk mi történne, ha a saját fejlesztésű középszótárunkat (negyvenezer szótári tétel) felcserélnénk az Akadémiai Nagyszótárral. Ez további öt százalékkal növelte sikeres lekérdezések arányát. A naplófájl alaposabb elemzése után azonban itt sem maradt el a meglepetés: nagyjából ugyanekkora javulás volt elérhető a húsz leggyakoribb, a szótárból hiányzó szó felvételével. (Hogy melyik ez a húsz szó? Kérem, lapozzanak vissza a gyakorisági statisztikákhoz...)

Új adatbázis-modulok hozzáadása

A „nem talált” listák elemzésével az is megállapítható, hogy nagy szükség lenne, egy, a WAP lehetőségeit kihasználó, magyar nyelvű, „idegenszavakszótára-lexikon-enciklopédia”-szerű adatbázismodulra is. Tudomásom szerint ilyesmi még nincs a piacon, és az adatbázis összeállítása is érdekes feladatnak ígérkezik.
A mérsékelt igény ellenére továbbra is tervezzük újabb nyelvek bevezetését.

A végrehajtott változtatásoknak köszönhetően a találati arány a kezdeti 64 – 66%-ról 73 - 75%-ra nőtt (13. ábra)



13. ábra

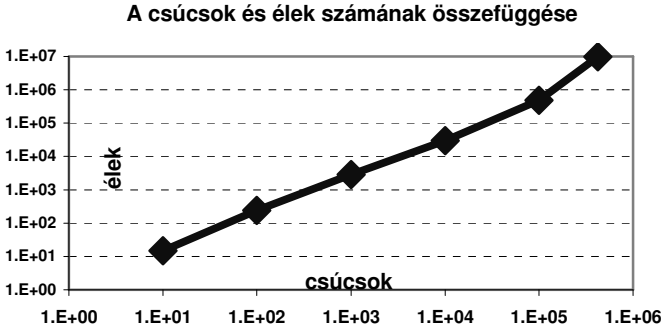
Az egy „menetben” egyszerre keresett szavak vizsgálata

Az eddigi „magától értetődő” statisztikai feldolgozás mellett az egyes tételek egymáshoz való viszonyát is vizsgálhatjuk. Azaz: akik az x szót keresik, milyen szavakat keresnek még?
A naplófájl tartalmazza a telefon típusát és a WAP-böngésző verziószámát is. Mindez elég specifikus adat ahhoz, hogy azt állíthassuk: az időben szorosan egymás után, azonos telefonról és WAP-browserről érkező kérések nagy valószínűséggel ugyanarról a telefonról, „egy menetben” érkeznek.

A „koincidenciagráf” előállítása

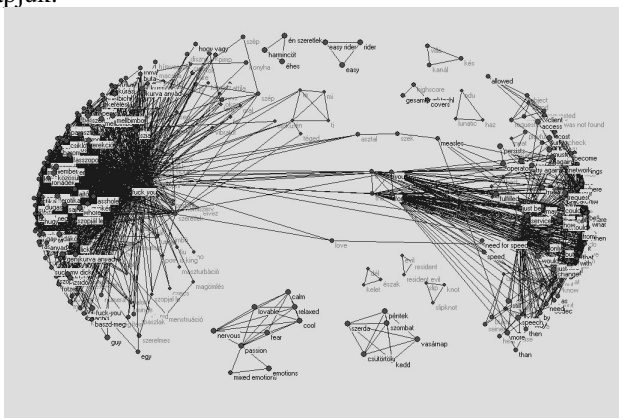
Az adatokat egy gráfban ábrázoltam, melynek csúcsai a keresett kifejezések. Két csúcsot akkor köt össze él, ha a csúcsoknak megfelelő szavak együtt szerepeltek egy lekérdezési menetben. Az élhez hozzárendeltem az együtt szereplés gyakoriságát.

Az így kapott teljes gráf 421 233 csúcsot és 9 776 746 élt tartalmaz. Ekkora adatmennyiséget sajnos egyetlen általam ismert elemző szoftver sem tud kezelni.
Az adatmennyiség csökkenthető, ha csak a nagy értékű éleket és a hozzájuk tartozó csúcsokat tartjuk meg. (Azaz csak azokat a szavakat, amelyek nagyon gyakran szerepelnek együtt egy „menetben”.)



14. ábra

A csökkentett csúcs- és élszámokat loglog skálán ábrázolva (14. ábra), megállapítható, hogy a kapott gráf nem más, mint egy – a napjainkban a legkülönbözőbb tudományterületeken „felfedezett" –skálafüggetlen hálózat[49].
A továbbiakban csak az ezer csúcsot tartalmazó részgráfot vizsgáltam. Ebben a Pajek[50] hálózatkezelő és –megjelenítő szoftver volt a segítségemre.
A gráf csúcsait a Fruchtermann–Reingold[51] algoritmussal átrendezve a következő elrendezést kapjuk:



15. ábra

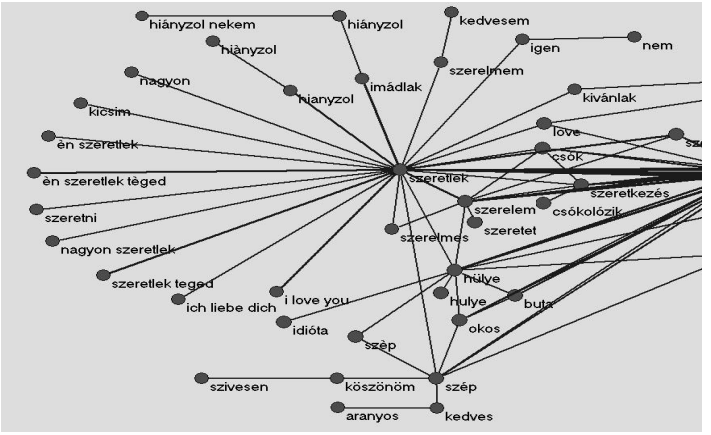
Jól látható, hogy a szavak nagy része két nagy clusterben csoportosul: A baloldali erősen centralizált cluster (a „Pornográf Birodalom") tartalmazza az obszcén szavakat, középpontban a legtöbb kapcsolattal rendelkező *pina* és *fasz* szavakkal. Jobboldalon látható, a sokkal több egyenrangú csúcsot tartalmazó részgráf, alap- és középfokú angol szavakkal (az „Angol Köztársaság"). A két tartomány között csak laza

[49] A téma jó összefoglalása olvasható Barabási Albert-László, vagy Csermely Péter könyvében
[50] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[51] http://www.boost.org/libs/graph/doc/fruchterman_reingold.html

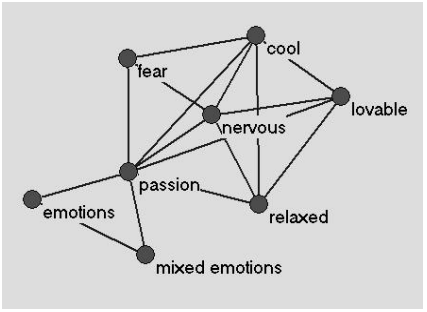
kapcsolat van tesztszavakon (*asztal, szék)* és nagyon gyakori angol szavakon *(a, the)* keresztül.

„Felfedezések" a gráfban

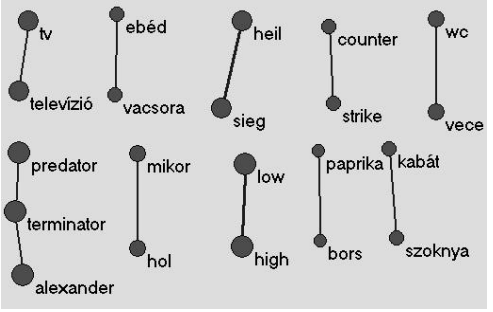
A két nagy tartomány mellett sok kicsi izolált vagy részgráf is feltűnik. Lássunk néhány érdekesebb szemantikai tartalmút:



16. ábra:
A „Pornográf Birodalom" *szerelem* szóhoz tartozó részgráfja, könnyű kitalálni hová is kapcsolódhatnak a jobboldali élek



17. ábra: „Emóciók" egy izolált gráfban



18. ábra Kételemű izolált részgráfok

Következtetések

A MoBiWAP rendszer üzemeltetése szempontjából a koincidenciagráf vizsgálatának egyértelmű a tanulsága: A sok kapcsolattal rendelkező szavakhoz tartozó szótári tartalmat különös gonddal kell összeállítani: ha a felhasználók elégedettek ezen szavak szócikkeivel, jó esélyünk van arra, hogy többször és más szavak esetében is ezt a szótárat fogják választani.

A koincidenciagráf beható vizsgálata meghaladja jelen cikk lehetőségeit, de (különösen, ha a négyszázszor nagyobb teljes gráfot tekintjük) minden bizonnyal nagyon sok hasznos információt tartalmaz a szemantikai csoportok vizsgálatához, valamint más nyelvszociológiai és nyelvészeti kutatásokhoz.

De erről majd egy másik alkalommal...

Köszönetnyilvánítás

Köszönettel tartozom Prószéky Gábornak és Kis Balázsnak, akik unszolása és biztatása nélkül soha sem írtam volna meg ezt a cikket; Vöröss Ferencnek, aki gondozta a szótári adatbázis tartalmát.
Köszönettel tartozunk a Pannon GSM-nek a MoBiWAP szolgáltatás megrendeléséért, máskülönben nem jött volna létre a hatalmas vizsgálható szöveganyag.

Bibliográfia

1. Barabási Albert László: Behálózva – a hálózatok új tudománya (Magyar Könyvklub, 2003)
2. Csermely Péter: A rejtett hálózatok ereje (Vince Kiadó, 2005)

IV. Morfológia és kivonatolás

morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis

Trón Viktor[1], Halácsy Péter[2], Rebrus Péter[3], Rung András[3],
Simon Eszter[4], és Vajda Péter[3]

[1] International Graduate College Language Technology and Cognitive Systems
University of Edinburgh és Saarland University
v.tron@ed.ac.uk
[2] BME – Média Oktató és Kutató Központ
hp@mokk.bme.hu
[3] MTA Nyelvtudományi Intézet – MTA-ELTE Elméleti nyelvészet program
{rebrus,runga,vajda}@nytud.hu
[4] BME – Kognitív Tudományi Tanszék
esimon@cogsci.bme.hu

Kivonat: Cikkünkben a morphdb.hu adatbázist mutatjuk be, amely a magyar nyelv egy minden eddiginél teljesebb és elméleti alapokon álló morfológiai leírása. A leírás a hunlex keretrendszerben van megfogalmazva, így a hunmorph szóelemző eszközkészlet segítségével az adatbázis helyesírásellenőrzéshez, tövezéshez, morfológiai elemzéshez és számos egyéb annotációs feladathoz használható elsődleges nyelvi erőforrásként.

Bevezetés

A szabály alapú szóelemzők működése – akár a végesállapotú, akár affixumlevágásos architektúrára gondolunk – a nyelv szókészletét és morfológiáját leíró erőforrásokat feltételez. Cikkünkben a morphdb.hu adatbázist mutatjuk be, amely a magyar nyelv minden eddiginél teljesebb és elméleti alapokon álló morfológiai leírása.

A morphdb.hu szóanyaga a helyesírásellenőrzésre használt Magyar Ispell szótár [6], az Elekfi László jegyezte Magyar Ragozási Szótár [2], valamint az ún. FKP-szótár [3] szóanyagának kritikus összefésülésével készült.

A szótár és a hozzá tartozó morfológiai nyelvtan leírását a hunlex keretrendszerben [8] végeztük el. A hunlex a morfológiai leírásból olyan kimeneti állományokat állít elő automatikusan, amelyeket a hunmorph szóelemző-algoritmusai (és hasonló ispell típusú erőforrást használó affixumlevágásos szóelemzők [9]) igényelnek. A hunlex és a hunmorph segítségével a morphdb.hu adatbázis így helyesírásellenőrzéshez, tövezéshez, morfológiai elemzéshez és számos egyéb annotációs feladathoz használható elsődleges nyelvi erőforrásként.

A hunlex rendszer morfológiai jelenségek formalizálására kidolgozott leíró nyelve alkalmasnak bizonyult a magyar nyelv komplex morfológiai jelenségeinek leírására. Az alábbiakban először a morfológiai nyelvtan néhány vonását mutatjuk be (§1), majd rátérünk a szótári anyag ismertetésére (§2). Összegzésképpen megmutatjuk, hogy miként mérhető az adatbázis lefedettsége és pontossága, valamint kitérünk jövőbeni terveinkre (§3).

1. A morfológiai leírás alapelvei és szerkezete

Tőtár és morfológiai folyamatok A hunlex rendszer leíró nyelve Item-and-Process [4] típusú morfológiai leírások formális keretéül szolgál. Egy hunlex morfológiai leírás lényegében egy tőtárból (lexicon állomány) és egy morfológiai operációkat formalizáló nyelvtani leírásból (grammar) áll. Az operációk konceptuálisan két részre oszthatók: (i) egyik részük konkrét morfoszintaktikailag additív szabályok, vagyis olyanok, amelyeket morf hozzáadásként is értelmezhetünk; (ii) másik részük pedig absztrakt morfofonológiai folyamatokat ír le, vagy diakritikumok (idioszinkratikus jegyek) és fonológiai-ortográfiai mintázatok közötti megfeleléseket fogalmaz meg. Az utóbbi, (ii)-es típusú szabályokat *szűrőszabályoknak* is nevezhetjük. Ilyen például a hangkivetés vagy a rövidülés mint absztrakt morfofonológiai folyamatok, vagy az ikességnek és a múlt idejű ragozásnak az összefüggését kimondó szabályok. Bizonyos szabályok (például a rövidülést leíró szabályok) egyazon absztrakt folyamat (rövidülés) megvalósulási formáinak tekinthetők, amelyeket közösen *szűrőnek* hívtunk.

Toldalékszabályok és szűrők A nyelvtan implementálásakor konkrét döntéseket kellett hoznunk arról, hogy a nyelvészeti elemzések által feltételezett morfofonológiai folyamatok közül melyeket általánosítjuk és melyeket nem. A leíráskor szem előtt tartottuk azt, hogy a szabályok leírása a legközvetlenebbül tükrözze a hagyományosan allomorfiának nevezett váltakozásokat, így a nem-konkatenatív morfofonológiai folyamatokat (pl. a magánhangzó-harmóniát kezelő fonológiai szabályt) nem absztrakt ((ii)-es típusú) szabályként írtuk le, hanem az allomorfikus szabályokban adtuk meg. Ezt illusztrálja az 1. ábra. (A hunlex leíró nyelvének technikai részleteire itt nem áll módunkban kitérni.)

```
CAS_INE
        IF: analytic lengthened cas_ine
        TAG: <CAS<INE>>

, +ban IF: back
, +ben IF: front
;
```

1. ábra. A vesszővel elválasztott szabályok a toldalékmorféma allomorfjainak (*-ban*, ill. *-ben* morfok) feleltethetők meg, a teljes szabály (CAS_INE) pedig így magának az affixummorfémának (az *inessivus* esetragnak).

A nyelvtan morfoszintaktikailag additív ((i)-es típusú) szabályait aszerint csoportosítjuk, hogy milyen morfoszintaktikai tulajdonságok (az 1. ábrán például az

inessivus eset) kifejezéséért felelnek. És megfordítva, egy ilyen szabálycsoport hagyományosan egyetlen morfémának tekintett toldalékot ír le.[52] Így ezeket a szabályegyütteseket (az 1. ábrán `CAS_INE`) némiképp lazán toldalékmorfémának nevezzük, az egyes szabályokat pedig toldalékallomorfoknak. A fenti elvek eredményeképpen lényegesen egyszerűsödik és átláthatóbb lesz a nyelvtan, és ezáltal új toldalékokkal való bővítése az Item-and-Arrangement szemléletet ismerő nyelvész vagy a fonológiai folyamatokhoz nem értő laikus számára is könnyebb.

Toldalékváltozatok kondicionálása A folyamatoknak a bemenetre való alkalmazása számos feltételhez lehet kötve. A `hunlex` leírásban a feltételek között szerepelhet mind a mintaillesztés, mind a jegyekre való hivatkozás (az 1. ábrán például a magánhangzó-harmónia jegyeire hivatkoznak a szabályok). A toldalékallomorf kiválasztása gyakran idioszinkratikus tulajdonsága a tőnek (pl. *hal<u>at</u>*, de *dal<u>t</u>*), ezeket tehát absztrakt jegyek ellenőrzésével kell kezelnünk. Ráadásul a fonológiailag megjósolhatóan kondicionált váltakozásokat a helyesírási konvenciók miatt egyes esetekben szintén önkényesen kellett kezelnünk (pl. *Voltaire-rel*). A legtöbb allomorfikus szabály alkalmazási feltételeit tehát jegyek segítségével fejeztük ki.

Lexéma-alapú tőtár és alulspecifikáció A lexikon bővítése akkor a legegyszerűbb, ha a (i) szótári bejegyzések lexémáknak[53] felelnek meg, valamint (ii) a szótári bejegyzéshez a lehető legminimálisabb morfológiai információt kell megadni. A szótárak a lexikográfiai hagyomány szerint lexéma-alapúak és a bejegyzések az ún. 'szótári alak' mellett csak a kivételességek (és megjósolhatatlan információ) feltüntetésével kerülnek az állományba. Ezt szem előtt tartva a nyelvtan részeként olyan folyamatokat (szűrőszabályokat) is feltételeznünk kellett, amelyek (i) a szótári alakokból előállítják a tőváltozatokat, valamint amelyek (ii) a toldalékváltakozások feltételeként használt megjósolható, de a lexikonban alulspecifikált jegyeket a tőváltozatokhoz asszociálják. A nyelvtan érdekes része tehát az a szűrőlánc, amely a morfofonológiai jegyeket kiosztja és a tőalternánsokat előállítja (például hangbetoldási, rövidülési és nyúlási folyamatok alkalmazásával).

Unáris licenciálás és opcionalitás A morfológiai adatbázisban túlnyomó részben unáris (egyértékű) jegyeket használtunk. Unáris jegynek nevezünk egy jegyet, ha a szabályok alkalmazási feltételei csak a jegy meglétére hivatkoznak. Például azoknál a toldalékolási folyamatoknál, ahol a kötőhangzó középső, illetve alsó nyelvállású (nyílt) is lehet (pl. többesszám *-ak*, *-ok*) a bemeneti relatív tőtől függően, ott a nyílt, illetve középső nyelvállást egy-egy jegy engedélyezi. A nyitótöveket (pl. *hal*, ill. nyitó relatív töveket, pl. *fák-*) a `low`, a nemnyitókat (pl. *dal*) pedig a `non_low` jegygyel kell ellátni ahhoz, hogy a megfelelő toldalékokat megkaphassák. Azok a tövek, amelyek opcionálisan nyitók (pl. *öröm*) egyszerűen mindkét jeggyel rendelkeznek. Ezzel a megközelítéssel elérhető, hogy (i) a szabályokat átlátható módon pozitív feltételhez kössük, ugyanakkor (ii) már a jelölés szintjén megmutatkozzon az opcionalitás jelöltsége, hiszen egy opcionális tőnél az adott jegydimenzió (nyitás)

[52] Ettől csak néhány praktikus esetben tértünk el, amikor két morféma szétválasztása bonyolítaná a nyelvtant, például a múlt idő jelét és az azt követő személyjeleket nem választottuk külön.

[53] Pontosabban alaplexémáknak (azaz nem megjósolhatóan képzett lexémáknak).

mindegyik unáris jegyét specifikálni kell. Azt, hogy egy kategória (pl. névszó) az adott dimenzióban (pl. nyitás) mindenképp felvesz egy értéket, itt azt jelenti, hogy a dimenziót kifejező unáris jegyek közül a jelöletlen vagy default jegyet (a nyitás esetében `non_low`, vagyis nemnyitó) az adott kategória minden eleme meg kell, hogy kapja. Ezt a nyelvtanban egy külön szűrő fejezi ki, amely a nyitást szabályozza: ez a szűrő a 2. ábrán látható.

```
NOM_LOWERING_FILTER

    FREE: false
    FILTER: low non_low
    OUT: NOM_KEEP_ALL_FEATURES
    OUT: NOM_ACC_FILTER


,OUT: non_low
;
```

2. ábra. Példa egy szűrőre: a nyitás szűrője minden nyitásra nem specifikált névszóhoz a (`non_low`) jegyet rendeli.

A nyitáshoz hasonlóan, azoknál a morfofonológiai tulajdonságoknál, amelyeknél felmerül az opcionalitás, következetesen unáris jegyeket használtunk. Ilyen tulajdonságra példa még az előlségi harmónia, az ikesség, a tárgyasság és a legtöbb tőváltakozás. Továbbá minden olyan tulajdonságra, amelynél az idioszintkatikus jegyek közül az egyik jelöltnek tekinthető, ott a nyitóhoz hasonló szűrőszabályt vezettünk be. Ilyenre példa az igéknél a tárgyasság.

Analitikusság és tőváltakozások Hasonlóan kezeltük a tőváltozatok választását befolyásoló ún. szintetikus illetve analitikus toldalékolást is [7]. Tőalternációt mutató lexéma esetén az egyes tőváltozatok kapják a jegyek valamelyikét. Itt a tőváltakozást nem mutató, egyalakú tövek opcionális töveknek felelnek meg, hiszen az egyalakú tő mind analitikus mind szintetikus toldalékolási folyamatokban részt vehet. Az analitikusság jegyei a nyelvtan belső jegyei, közvetlenül a lexikonban csak a tőváltakozás típusát (rövidülő, hangkivető, *v*-vel bővülő, *sz-d* tő, stb.) kódoló jegy jelenik meg. A szintetikus és analitikus tőváltozatok előállítása során a megfelelő jegyek a változatokhoz rendelődnek, az egyalakú szótári tövek pedig automatikusan mindkét jegyet megkapják.[54]

Magánhangzónyúlás Hasonlóan kezeltük a névszói magánhangzónyúlást is. Azok a toldalékok, amelyek kiváltják a magánhangzó-nyúlást (többesszám, birtokjel, a legtöbb esetrag, stb.) az affixumszabály feltételeként ún. „hosszú magánhangzós tő" jegyet (lengthened) követelnek meg, míg a nem nyújtó toldalékok (pl. *-ként*, *-kor*, *-ság/ség*) a „nem hosszú magánhangzós tő" jegyet (`non_lengthened`). A magánhangzónyúlás, mint folyamat így egy virtuális morf segítségével kezelhető, amely a

[54] Bizonyos tövek egyalakúak, de mégsem vehetnek fel minden szintetikus és analitikus toldalékokat: ezek az ún. defektív tövek (pl. *rejlik* (vö. **rej(e)ljen*), amelyek már a lexikonban szintetikusként szerepelnek.

megnyújtott véghangzós tövet (*fá-*) előállítja és ellátja a hosszú jeggyel, a szótári tövet (*fa*) pedig a nem-hosszú jeggyel látja el. Azok a tövek, amelyek végződésük szerint soha nem nyúlnak (pl. *mozi*, *bot*) megkapják mind a hosszú mind a nem-hosszú jegyet, így természetes módon az egyalakú tövekhez hasonlóan viselkednek.[55]

Idegen helyesírású szavak kiejtés szerinti toldalékolása A fonológiailag és ortográfiailag kondicionált allomorfia jegyekkel történő kezelése egyéb pontokon is hasznosnak bizonyult. Egyrészt az affixumszabályok leírásakor az absztrakt feltételek megadása tömörebbé válik, és ezek koordinálhatók lesznek (több konjunktív feltétel megadható egyszerre, pl. hosszú és mély hangrendű), ami az elemző algoritmus korlátai miatt nem lehetséges közvetlen illesztési kifejezés esetén. A másik nagyon fontos előny pontosan ahhoz kapcsolódik, hogy a nyelv hangtanilag kondicionált szabályszerűségeit az ortográfia speciális szabályai szerint írjuk le. Az idegen helyesírású, de kiejtés szerint ragozott szavak (pl. *Voltaire*) kezeléséhez a toldalékválasztást adó jegyek elengedhetetlenek. Ez a fajta ortográfiai önkényesség egyszerű módon kezelhető, ha bizonyos idegen szavak a kiejtésükkel együtt vannak felvéve a lexikonba (pl. *Voltaire/volter*). A `hunlex` keretrendszer lehetővé teszi, hogy az ilyen szavaknál, az egyes toldalékok az írásképhez járuljanak, de úgy, hogy a toldalékallomorfok kiválasztása mégis a kiejtésük szerint történjen. Ehhez csupán annyi kell, hogy a kivételes kiejtésű szótári tételeknél fel legyen tüntetve a filterekben található minták szempontjából releváns kiejtés. Ekkor a tő a jegyeket a kiejtés szerint kapja meg, majd a szűrőlánc végén egy szabály egyszerűen törli a kiejtésre vonatkozó részt a tőből. A `hunlex` azt is lehetővé teszi, hogy bizonyos kiejtési szabályokat is a nyelvtanban adjunk meg. Ezt egyelőre csak a szavak egy speciális csoportjánál, a betűszavaknál (pl. *http*) vezettük be, ahol a kiejtés egyértelműen rekonstruálható (*hátétépé*).

Morfoszintaktikai jegyek és hiányos paradigmák Mivel az igék és a névszók releváns morfofonológiai jegyei nagyrészt diszjunktak, ez a két nagy kategória külön szűrőláncot kívánt. Hasonlóan a névszókon belül számos (főként képzési) folyamat érzékeny arra, hogy az illető névszó főnév, melléknév vagy számnév. Az ilyen alkategória-érzékeny toldalékolás miatt az egyes morfémák (*–s* képző, *-szoroz* képző, stb.) és morfémacsoportok (esetrag, birtokos személyjel, stb.) kapcsolódását szintén jegyek meglétéhez kötöttük. Az ilyen (unáris) morfoszintaktikai jegyek alkotják a `morphdb.hu` jegyeinek másik részét. Hasonlóképpen ilyen jegyekkel értük el, hogy az igéknél a tárgyasságra, vagy a névszóknál a köznév–tulajdonnév különbségére érzékeny szabályok ne generáljanak túl. Egyes morfoszintaktikai jegyek a lexéma alaptulajdonságai (főnév, számnév, stb.), így a tőtárban minden bejegyzésnél szerepelnek. A lexikonállományban más jegyek csak jelölt esetben (tulajdonnév, tárgyas ige) szerepelnek, ilyenkor az adott dimenzión belüli default tulajdonságokat egyszerű

[55] A nyúlás jegyei szintén a nyelvtan belső jegyei, azonban a kiosztásukat szabályozó filter nem lexikai jegyekre, hanem fonológiai mintázatokra hivatkozik. A nyúlást leíró folyamat default szűrőként való implementálását az motiválja, hogy egyes alakok (*la*, *Che*, *Mandrake*) a releváns mássalhangzó ellenére nem váltakoznak, vagyis ezeket a lexikonban az opcionális alakokhoz hasonlóan a dimenzió minden jegyével el kell látni.

szűrőszabályok segítségével rendeltük az alakokhoz.[56] Ezek a morfoszintaktikai jegyek szabályozzák a hiányos paradigmájú alakokat is (pl. *léptek*, *két*, *ismerszik*).

Összefoglalásképpen a `morphdb.hu` igei szűrőláncának összetevőit soroljuk fel.

- az ikesség felismerése a szótári alak alapján automatikus, tő előállítása
- a default tőtípus hozzárendelése a nemkivételes egyalakú lexémákhoz
- az elölségi harmóniát szabályozó unáris jegyek hozzárendelése mintaillesztéssel
- tőváltozatok előállítása: hangkivetés és -betoldás, *sz-d*(*-v*)-tövek, *v*-tövek, rövidülés
- kerekségi harmónia unáris jegyeinek hozzárendelése mintaillesztéssel
- szótagszámot kódoló jegyek hozzárendelése (a szótagszámérzékeny szabályok miatt) mintaillesztéssel
- alanyi ill. tárgyas ragozást engedélyező jegyek hozzárendelése: a default csak alanyi ragozás
- kvázianalitikus toldalékolást szabályozó jegyek hozzárendelése mintaillesztéssel (*hoznak* vs. *vonzanak*)
- a múlt idő változatait szabályozó jegyek hozzárendelése mintaillesztéssel (*hoztak* vs. *vonzottak*)
- összetételi határokat törlő szabályok
- kategóriaérzékeny képzést engedélyező morfoszintaktikai jegyek kiosztása

Alulspecifikáció és kivételesség A mintaillesztéssel történő allomorfkiválasztás (jegy-hozzárendelés) csak akkor lehetne lehetséges, ha a szóban forgó jelenség tökéletesen megjósolható lenne. Bár számos ilyen jelenség van (pl. az ikesség a szótári alak alapján, vagy a kerekségi harmónia a kiejtés alapján tökéletesen megjósolható), sok esetben a toldalékolás önkényes, azaz megköveteli, hogy egyes jegyeket a lexikonban adjunk meg. Ilyen esetekben azt az elvet követtük, hogy minél tágabb körű általánosítást fogalmazunk meg a nyelvtan szűrőiben úgy, hogy a kivételesnek tekintett (tehát a lexikonban jeggyel ellátott) alakok lehetőség szerint véges zárt osztályt alkossanak. Egy adott dimenzióra nézve ennek a zárt osztálynak a tagjai tekinthetők kivételesnek. Ezzel a szótárbővítési munkát remélhetőleg minimálisra csökkentjük, hiszen ez a módszer csak olyan jegyek specifikációját kényeszeríti a szótárfejlesztőkre, amelyek egy nyílt osztályra is megjósolhatatlanok. A morfofonológiai jegyek közül csak a névszói birtokos toldalékolás típusát (*-a/e* vagy *-ja/-je*) szabályozó jegy ilyen.

[56] Mivel vannak csak tárgyas részparadigmát megengedő igék (*megemberel(i magát)*), ezért a tárgyasság dimenziója is két unáris jeggyel ábrázolódik: `verb_indef` (engedélyezi az alanyi részparadigma toldalékait), és `verb_def` (engedélyezi az tárgyas részparadigma toldalékait). Hasonlóan a hangrendileg ingadozó opcionalitáshoz, a tárgyas igéket (amelyek mindkét részparadigmát engedélyezik) a két jegy együttes jelenlétével adjuk meg.

2. A magyar morfológiai adatbázis szóanyaga

A szótári anyag A `morphdb.hu` szótári anyagának elkészítéséhez három önmagában is nagy lefedettségű elektronikus szótárat használtunk fel. A Magyar Ispell szótár [6] a szabad forráskódú szoftverek világában domináns `ispell` alapú helyesírásellenőrző magyar nyelvű erőforrása. A Németh László vezetése alatt közös munkaként elkészült Magyar Ispell szótár a mai magyar nyelv egyik legteljesebb és legnaprakészebb szóanyagát tartalmazó anyag. A `magyarispell` átalakításakor a `hunmorph` morfológiai elemzőhöz készült szótár nyers változatából indultunk ki. A több mint 100 ezer szavas szóállomány témakörök, stílusminősítések alapján van csoportosítva, az ún. alapszókincs mintegy 37 ezer szóból áll. A tematikus szótárak anyagát is átvettük, megtartva a témakört, mint a szótári bejegyzéshez adott használati információt.

Másik forrásunk a Magyar Ragozási Szótárnak [2] a Nyelvtudományi Intézet Korpusznyelvészeti osztálya által digitalizált változata. Az Elekfi-szótár az Értelmező Kéziszótár közel 70 ezer lemmáját tartalmazza paradigmaosztályokba sorolva. A szótár jelöli a komplex szavak belső szerkezetét is, és számos összetettségi típust elkülönít. Ezek közül némelyeket egybevontunk, de az összetett szavakat valamint igekötős igéket felbontásukkal együtt átvettük.

Harmadik forrásunk, az FKP-szótár, Papp Ferenc Szóvégmutató Szótárából és a Füredi-Kelemen-féle Gyakorisági szótárból állt elő [3] és kb. 70 ezer tételt tartalmaz.

A források átalakítása A `morphdb.hu` szótár előállításának első részeként ezen létező szótárak anyagát kellett a `hunlex` lexikon formátumára hozni, úgy, hogy a morfológiai információikat az első részben vázolt elvek figyelembevételével a `hunlex` nyelvtanban használt jegyekké átalakítsuk. A morfofonológiai jegyeket egy-egy szónál csak akkor kívántuk felvenni, ha a nyelvtanban meghatározott szabálytól eltérően viselkednek, azaz a szó ebből a szempontból kivételes. A források viszont minden információt tartalmaznak, így a nyelvtanunk szempontjából szabályosnak tekinthető folyamatokat szabályozó default jegyeket is. Az átalakításnál ennek a redundanciának a kiküszöbölése volt az egyik probléma.

Második lépésként az átalakított és redundanciamentesített forrásszótárakat kellett összefésülni. Ez a több forrásban is előforduló bejegyzések esetében annak eldöntését is megkívánta, hogy melyik „autoritás" által adott morfológiai leírást fogadjuk el a legpontosabbnak.

Célunk nem csupán egy minden eddiginél nagyobb, de elméletileg is jobban megalapozott morfológiai leírást követő szótár előállítása volt. A morfológiai információ átvételében csak a nyílt szóosztályok (névszók és igék) esetén követtük az eredeti források morfológiai leírását. A határozószavak, névutók, névmások, kötőszók és egyéb szófajok feldolgozásánál csupán a szóanyagot vettük át, de megkíséreltünk egy adekvátabb leírást, csoportosítást kialakítani.

A jegyek hozzárendelése: nyílt szóosztályok A magyarispell szóállomány már bizonyos morfofonológiai tulajdonságok alapján (hangrend, birtokos *-j* megléte, tőallomorfia, tárgyasság, stb.) csoportosítva tartalmazza a szavakat és a kivételeket, ezért ezek átalakítása a nyelvtanunkban használt jegyekké nem ütközött nehézségekbe.

A Magyar Ragozási Szótár a bejegyzéseket lehetséges toldalékaik alapján diszjunkt csoportokba, paradigmákba sorolja. Egy-egy paradigmába csak a teljesen

egyformán viselkedő (ugyanazon affixumokat felvevő) szavak kerültek, és a paradigmaosztályok száma összesen 1700. Az osztályok számozása nyelvészeti szempontból nem rendszerezett, ami azt jelenti, hogy a kódok közti különbség a legtöbb esetben (kivéve az előlségi és a kerekségi harmóniát, illetve az ikes és nem-ikes igéket) nem tükrözi az adott kódú csoportokba sorolt szavak közti ragozásbeli különbséget. Ezért az átalakításához egy olyan táblázatot kellett elkészíteni, mely minden egyes paradigmakódhoz megadja, hogy a `morphdb.hu` szótárban az adott csoport milyen jegyeket kap. Bár a paradigmák nem veszik figyelembe a tárgyas és nem tárgyas igék különbségét, maga a szótár tartalmazza ezt az információt, amelyet ily módon egyszerűen át tudtunk venni.

Az FKP-szótár bejegyzései a lemmákhoz tartozó morfológiai információt nem paradigmabesorolással, hanem a `morphdb.hu` szelleméhez közelebb álló módon tartalmazza. Például egyes mezők egy toldalék vagy toldalékcsoport allomorfjai közül való választást adják meg (pl. tárgyeset, *-at*, vagy *t*), mások közvetetten absztrakt tulajdonságokat (pl. tőtípus) specifikálnak. Az FKP szótár átalakítását úgy végeztük, hogy egy táblázatban minden mezőértékhez megadtunk egy `morphdb.hu`-beli jeghalmazt, amelyet a mezőérték jelenléte esetén a szótári bejegyzéshez asszociáltunk.

Mivel a lemmákhoz csak a minimális, nem megjósolható információt kívántuk tárolni, az átalakítás részeként a nyelvtan által a szavakhoz rendelt, és így a szótárban redundáns jegyeket töröltük. Ezt a `morphdb.hu` szűrőinek alkalmazásával tettük meg: ha egy szűrő egy bizonyos dimenzió jegyeit helyesen rendeli a szóhoz, akkor a szótárból annál a szónál töröljük az illető jegyeket.

A szótárak összefésülése A három forrásszótár szóanyagának átalakítása és redundanciamentesítése után a közös szókincs kiszűrését kellett megoldanunk. A mindhárom szótárban előforduló szavak száma 28 ezer, és tízezres nagyságrendű a páronként közös, illetve a csak az egyik szótárban előforduló szavak száma. A három szótárat összevonva jelenleg a lemmák száma 150 ezer körüli. Megfigyelhető, hogy az egyes szótárak egyedi hozadéka szintén 10 ezres nagyságrendű (az `Ispell` 70 ezer egyedi tétele a hatalmas tulajdonnévtárnak köszönhető, amit a többi szótár nem tartalmaz), vagyis bármelyik nélkül lényegesen szegényebb lenne a szóállományunk. Az egyes forrásállományok bejegyzéseinek, valamint az átfedő tételek számát mutatja az 1. táblázat.

1. táblázat. A forrásszótárak számokban

SZÓTÁR	BEJEGYZÉSEK SZÁMA
Ispell	105580
Elekfi	67047
FKP	68316
Ispell ∩Elekfi	32898
Ispell ∩ FKP	30754
Elekfi ∩ FKP	54607
Ispell \ (Elekfi ∪ FKP)	70591
Elekfi \ (Ispell ∪ FKP)	8155
FKP \ (Ispell ∪ Elekfi)	11568
Ispell ∩ Elekfi ∩ FKP	28663

Az összefésülésnél ellenőrizni is tudtuk az egyes átalakításokat, mivel a több szótárban szereplő szavak esetében ugyanolyan bejegyzéseknek kell kijönniük. Ez azonban az esetek kis részében fordult csak elő. A többi esetben a különbségek egyik oka az volt, hogy az egyes erőforrások másként kezelnek egy-egy szót, más-más alakjait tartják helyesnek. Ez különösen azokban az esetekben fordult elő, ha az egyik szótár egy szót (bármilyen szempontból) ingadozónak tüntet fel.

A különbségeket automatikusan csoportosítottuk – egy-egy csoportba azok a szavak kerültek, amelyekhez azonos módon rendelődött többféle jegyhalmaz. Például azok a szavak, amelyhez az egyik forrás csak a *preverb* jegyet egy másik pedig a *preverb*, *trans* jegyeket rendelte egy csoportba kerültek. Ezeket a csoportokat kézi feldolgozással átnézve alakult ki a nyílt tokenosztályok szótárának végső állapota.

Egyéb kategóriák A névmások és névutók nem kerültek közvetlen módon átvételre, ezeket főként a nyelvtanon belül külön jegyek és toldalékolási folyamatok segítségével kezeljük. A forrásszótárak összes egyéb szófajú bejegyzését kézzel átnéztük és újra szófajokba soroltuk őket. Ezek a szófajok nem feltétlenül követik a források besorolását, és gyakran a forrásokban is ellentmondásos besorolással szerepelnek. A 2. táblázat a `morphdb.hu` főkategóriáit és a `hunlex` által kompilált kimeneti szótárban a hozzájuk tartozó bejegyzések számát adja meg.

Morfológiai annotáció A `morphdb.hu` természetesen tartalmazza azt az információt, amellyel a kimenetet használó morfológiai elemző a szavakat címkézi. A szótári adatbázist úgy készítettük el, hogy az egyes morfémákhoz tartozó morfoszintaktikai tagek változtathatóak legyenek, vagyis az elemző kimeneti annotációja rugalmas. Jelenleg a `morphdb.hu` az ún. KR-kódolást [5] támogatja. A KR jelölés a morfoszintaktikai tulajdonságoknak a jelöletlenséget jól kifejező, hierarchikus gráfreprezentációja.

2. táblázat. A morphdb.hu főkategóriái

FŐKATEGÓRIA-CÍMKE	FELOLDÁS	BEJEGYZÉSEK SZÁMA
NOUN	főnév	88026
ADJ	melléknév	17514
VERB	ige	12549
ADV	határozószó	1932
UTT-INT	mondatszó/interjekció	498
CONJ	kötőszó	258
NUM	számnév	209
DET	determináns	164
POSTP	névutó	146
PREV	igekötő/igevivő	132
ONO	hangutánzó szó	96
PUNCT	központozás	28
PREP	prepozíció	14
ART	névelő	2

A szótár egyéb információi Mindhárom forrás a toldalékolásra vonatkozó gazdag morfológiai leírás mellett számos egyéb hasznos információt tartalmaz, pl. témakör (Magyar Ispell), belső szerkezet (Elekfi), stiláris és használati információ (FKP), amelyet a `morphdb.hu` adatbázisában is megőriztünk.

Ezen kívül a névmások és a határozószók külön alkategóriákba lettek sorolva. Az alkategóriára vonatkozó információk a `hunlex` beállításával a kimeneti állományokba kerülhetnek, s ezáltal az elemző tetszőleges annotációs eszközként is használható.

3. Konklúzió

A `morphdb.hu` szóanyagának és nyelvtanának pontosságát és fedettségét a kézzel taggelt Szeged Korpusszal [1] való összehasonlítással ellenőriztük. A korpuszban magadott MSD-kódokat a `morphdb.hu` kimeneti annotációjára [5] alakítottuk át egy konverziós táblázat segítségével, majd a korpusz szavait leelemeztük a `morphdb.hu` erőforrást használva is. Így a kézzel és az elemzővel taggelt változat összehasonlíthatóvá vált. A két elemzés közötti eltérések egy része a készítők eltérő nyelvészeti felfogásából ered. Az összehasonlítás érdekében ezeket a konkrét elemzéseket átsoroltuk, azaz az MSD-kódokat nem a neki megfelelő `morphdb.hu` kódra (ha volt ilyen[57]), hanem az általunk helyesnek tartott elemzésre cseréltük. Például az *is* szó tagjének (`Cccp`) „fordítása” `CONJ` (kötőszó), a `morphdb.hu` azonban `ADV`-ként tartja számon. A névutók, névmások kapcsán sok ilyen különbséget szisztematikusan kezeltünk. E folyamat közben a Szeged Korpusz számos hibásan taggelt szavát is sikerült azonosítanunk.

Az átalakítások után az elemző jellemző fedettsége 90%-os, a látszólag magas 10% elemzetlen alak közül a legtöbb tulajdonnév, rövidítés, illetve olyan névszók, amely az Elekfi és FKP szótárokban egyszerűen a koruknál fogva nem szerepelnek (pl. *rendszergazda*, *internetező, limit, fájl*). Már az első teszteléstől fogva ezekkel a szavakkal bővítjük a `morphdb.hu` anyagát, végső célunk a közel 100%-os fedettség elérése.

A `morphdb.hu` pontosságának mérése nehezebb feladat, hiszen még a korpuszban előforduló szóalakoknak is számos a korpuszban nem szereplő alternatív elemzése lehet. Emiatt a pontosságot a korpusz segítségével csak azzal tudjuk tesztelni, hogy a `morphdb.hu` erőforrással kiadott elemzések között szerepel-e a korpusz szerint helyes elemzés.[58] Az ilyen hiányzó elemzés jelenleg a szóalakok kevesebb, mint 1%-ánál fordul elő, és ezek is ritka szavak, hiszen tokenszázalékban ez a korpusznak hozzávetőleg 0.1%-a. Célunk az ilyen hibák lehetőleg teljes kiküszöbölése. Ez ezután is szakértői munkát igényel, hiszen szem előtt kell tartani a `morphdb.hu` kategorizálásának a Szeged Korpusztól való szándékolt eltéréseit.

A jövőbeni terveink között szerepel a `morphdb.hu` további kézzel egyértelműsített szövegekkel (pl. Nemzeti Szövegtár) történő tesztelése és ezzel párhuzamosan a szótár bővítése. Számos tulajdonnévlista (magyar vezetéknevek, cégne-

[57] Pl. a *-hAt* morfémát az MSD-kód nem jelöli, a `morphdb.hu`-ban pedig inflexióként szerepel.

[58] Ezzel persze a túlelemzések nem kezelhetők, azok szűrésére más módszert kell találnunk.

vek, földrajzi nevek) is rendelkezésünkre áll, amelyeknek `hunlex` formátumra történő alakítása nem ütközhet nagy nehézségbe. Fontos tervünk a szóanyag normatív ellenőrzése is, vagyis a szubsztandard (helyesírású) alakok pontosabb megjelölése, hogy az adatbázisból egy a Magyar Ispell szótár minőségét elérő helyesírás-ellenőrző erőforrást lehessen generálni. Tervünk a szóösszetételi modul finomítása is, amely jelenlegi formájában igen megengedő.

Köszönetnyílvánítás

A `morphdb.hu` létrejöttében sokan segítségünkre voltak. Külön köszönet illeti Kornai András, Németh László és Varga Dánielt. Köszönet a Magyar Telecomnak a projekt anyagi és infrastrukturális támogatásáért.

Bibliográfia

1. Csendes Dóra, Hatvani Csaba, Alexin Zoltán, Csirik János, Gyimóthy Tibor, Prószéky Gábor, Váradi Tamás. Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In II. Magyar Számítógépes Nyelvészeti Konferencia, 238–245. Szegedi Tudományegyetem, 2003.
2. László Elekfi. Magyar ragozási szótár. MTA Nyelvtudományi Intézet, Budapest, 1994.
3. Mihály Füredi, András Kornai, and Gábor Prószéky. A szo1ta1r adatbázis. Kézirat, 2004.
4. Charles F. Hockett. Two models of grammatical description. Word, 10:210–234, 1954.
5. András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, and Viktor Trón. általános célú morfológiai elemző kimeneti formalizmusa. In II. Magyar Számítógépes Nyelvészeti Konferencia, 172–176. Szegedi Tudományegyetem, 2004.
6. László Németh. Magyar Ispell – Válasz a Helyes-e?-re. In IV. GNU/Linux szakmai konferencia, pages 99–107. Linux-felhasználók Magyarországi Egyesülete, 2002.
7. Péter Rebrus. Morfofonológiai jelenségek [morphophonological phenomena]. In Ferenc Kiefer, editor, Strukturális magyar nyelvtan. 3. Morfológia. [Hungarian structural grammar. 3. Morphology.], 763–948. Akadémiai Kiadó, Budapest, 2000.
8. Viktor Trón. Hunlex - morfológiai szótárkezelő rendszer. In II Magyar Számítógépes Nyelvészeti Konferencia, 177–182. Szegedi Tudományegyetem, 2004.
9. Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In Proceedings of the ACL05 Software Workshop. Ann Arbour, 2005.

Morfológiai egyértelműsítés maximum entrópia módszerrel

Halácsy Péter[1], Kornai András[1], Varga Dániel[1]

[1] Budapesti Műszaki Egyetem -- Média Oktató és Kutató Központ,
1111, Budapest, Stoczek u. 2.
{hp, kornai, daniel}@mokk.bme.hu

Kivonat: Cikkünkben olyan magyar nyelvű statisztikai morfológiai egyértelműsítő modelleket hasonlítunk össze, amelyekbe a korpusztól független morfológiai elemzőt is beleépítettünk. Ismeretes, hogy magyar nyelvre a morfológiai elemző alkalmazása megnöveli a pontosságot a tisztán statisztikus módszerekhez képest. Modelljeink ugyanakkor a maximum entrópia módszer segítségével hatékony becslést adnak a morfológiai elemző által fel nem ismert szavakra is, tehát robusztusan viselkednek olyan tesztkorpuszokon is, amelyekhez a morfológiai elemző nem lett adaptálva.

1. Bevezetés

A morfológiai analízis (MA) a magyar, és általában az összetettebb morfológiájú nyelvek számítógépes kezelésének egyik központi feladata: a helyesírás-ellenőrzéstől a gépi fordításig szinte nincs is olyan gyakorlati alkalmazás, amelyhez valamilyen formában ne lenne szükséges MA. De még ha tökéletes (minden szót ismerő, és hibát soha nem vétő) MA algoritmus állna is rendelkezésünkre, akkor is szembe kell néznünk azzal a ténnyel, hogy a magyarban számos szóalak többértelmű, és hogy melyik elemzés a helyes, azt csak a szövegkörnyezet alapján lehet eldönteni.

Cikkünkben a morfológiai egyértelműsítés problémáját a statisztikai módszerek szemszögéből tárgyaljuk: ennek fő előnye, hogy a kontextus vizsgálatát egyértelműen korpusznyelvészeti alapokra helyezi. A címkézési feladatra a legjobb eredményt nyelvünkre tudomásunk szerint eddig Oravecz és Dienes [10] érte el 98.11% pontossággal. Ők a *TnT* rejtett Markov modell (HMM) alapú rendszert [2] módosították: a legnehezebb feladathoz, a tanítókorpuszban nem látott szavak helyes címkézéséhez a Humor morfológiai elemzőt hívták segítségül.

Cikkünk első részében bevezetjük a valószínűségi MA (WMA, weighted MA) fogalmát, és ennek segítségével a morfológiai egyértelműsítési probléma nehézségére adunk előzetes becslést. A második részben egy a magyar nyelvre eddig még nem alkalmazott, a maximum entrópia elvén alapuló szófaji címkéző módszert ismertetünk. Ehhez morfológiai elemző komponensként a hunmorph rendszert [12] alkal-

maztuk a morphdb.hu nyelvi erőforrással [14]. Az eredményeket a harmadik részben ismertetjük és értékeljük.

Magyar nyelvre a korábbi vizsgálatok elsősorban egy idealizált (a tesztanyag minden szavát garantáltan ismerő) morfológiai elemzőre támaszkodtak, ezért általános felhasználási értékük némileg megkérdőjelezhető, különösen akkor, amikor olyan kicsi és stilisztikailag homogén korpuszon alapulnak, mint a MULTEXT-East *1984* anyaga [3]. Munkacsoportunk az itt bemutatott algoritmus tanításához és teszteléséhez a Szeged Korpusz 2. változatát [4] használta, ennek az *1984* csupán 8%-a, és az Oravecz és Dienes [10] által használt korpusszal (280 ezer szövegszó) stílusában leginkább összemérhető `wholenews` szekció (ezt a sajtó és az üzleti rövidhír részkorpuszok összevonásával hoztuk létre) is némileg nagyobb a Szeged Korpuszban (350 ezer szövegszó).

Bár az `1984` anyagon elért 97.91%, a `wholenews` anyagon elért 98.38%, és Szeged Korpusz egészén elért 98.17% numerikusan nem jelentenek hatalmas javulást, úgy véljük, hogy rendszerünk a gyakorlatban jobban használható lesz. Nem csak azért, mert kritikus komponensei, beleértve a `WMA`-t, nyílt forráskódúak és szabadon módosíthatóak, hanem mert az általunk javasolt algoritmus robusztusan ellenáll a korpuszhoz nem igazított MA algoritmusok gyakorlatban nem ritka lefedettségi hiányosságainak, és mint ilyen, lehetővé teszi az eddiginél nagyobb változatosságú, pl. a dinamikusan növekvő magyar web kiaknázásával épült korpuszok [6] morfológiai elemzését is.

2. A címkézési feladat

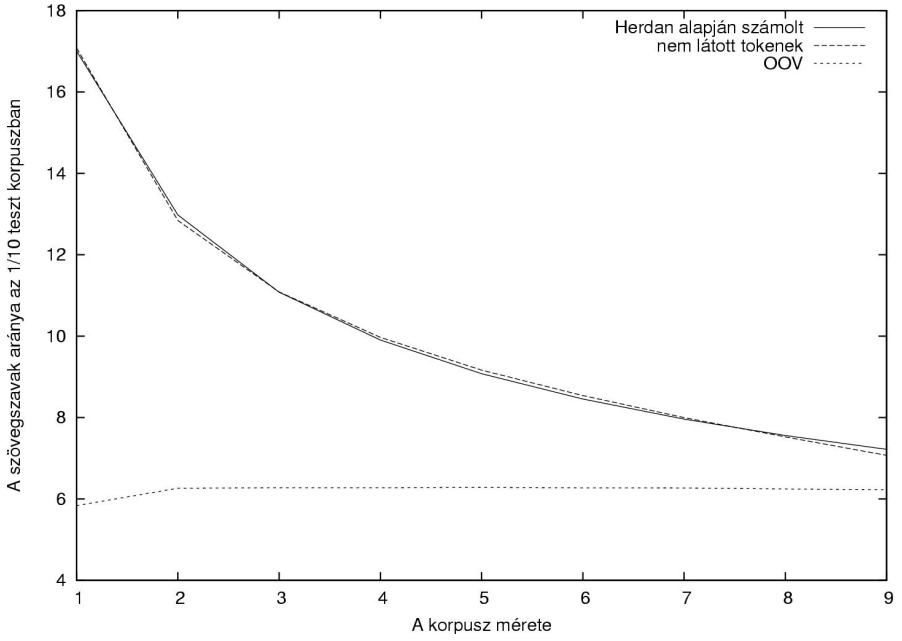
A morfológiai egyértelműsítés központi feladata a több elemzéssel rendelkező szavak esetében a helyes elemzés kiválasztása: ennek a feladatnak a nehézségét szokás a többelemzésű szövegszavak arányával [4], illetve az egy szövegszóra jutó elemzések átlagos számával [13] mérni. Ezeket a számokat azonban erősen torzítják a gyakori, de nem minden elemzést egyforma valószínűséggel nyerő szövegszavak (pl. *az* tipikusan névelő de lehet mutató névmás is, *én* tipikusan névmás, de pszichológiai szakszövegben gyakran főnév), hiszen a legegyszerűbb maximum likelihood címkézési stratégia számára ezek nem igazán problémásak.

A feladat nehézségének helyes mérőszáma tehát az egy szó egyértelműsítéséhez átlagban szükséges információmennyiség. Ha a *w* szó a T_i címkét $P(T_i \mid w)$ valószínűséggel kapja (címkézett korpuszból ezt a $C(T_i, w) / C(w)$ hányadossal becsülhetjük empirikusan, ahol *C* az előfordulások száma) akkor e szó címke-entrópiája $H(w) = -\sum_i P(T_i \mid w) \log P(T_i \mid w)$, és a címkézési feladat egészének nehézségét ezen entrópiáknak a *w* szavak gyakorisága szerint súlyozott átlaga adja, vagyis: $\sum_w P(w) H(w)$. Ez a Szeged Korpuszon durván 0.1 bit/szó (a pontos érték a választott címkerendszertől függ), tehát messze nem olyan nagy, mint azt a többelemzésű szavak aranyából gondolhatnánk: ha a lehetőségek mindig éppen egyformán valószínűek és a korpusz fele kétértelmű [4], akkor az entrópia akár 0.5 bit/szó.

A gyakorlatban természetesen a morfológiai elemző nem tökéletes, az egyes szavak gyakoriságát és címke-entrópiáját pedig csak becsülni tudjuk. Különösen érdeke-

sek számunkra azok a módszerek, amelyek e becsléseket a morfológiai elemző kiküszöbölésével, egyenesen a korpuszból végzik, hiszen ezek a morfológiai analízis (MA) nélkül működő, csak a korpuszból tanuló címkéző algoritmusoknak felelnek meg. A címkézési feladatot már ilyen algoritmusokkal is meglehetősen sikeresen meg lehet oldani: ha például minden adott szövegszóhoz a tanítókorpuszban látott szövegszavak esetén a típus leggyakrabban előforduló címkéjét, a nem látott típusok esetén pedig a nyílt kategóriák közül a leggyakoribb (egyes szám alanyesetű főnév) címkét rendeljük, akkor a Szeged Korpuszon (90% tanítás, 10% teszt, 10-szeres keresztvalidáció) 92% pontosságot érünk el. Ugyanezt az algoritmust tekinti alapszintnek (baseline) [10], de ott csak 81.2% pontosságot mérnek. A különbségnek az az oka, hogy a mi tanító- és tesztkorpuszaink egy nagyságrenddel nagyobbak, és így esetünkben csupán 10.7% a nem látott szövegszavak aránya, szemben az általuk tapasztalt 17.13%-kal.

Általában, ha a tanítókorpusz mérete *N*, a tesztkorpuszé ennek konstans hányada (pl. *N/10*), akkor Herdan törvénye szerint a tesztben az új szavak aránya cN^{q-1} ahol q a Zipf konstans reciproka. Az 1. ábrából látható, hogy a korpusz méretének növekedésével a fix arányú tanító- és tesztkorpusz esetén a nem látott szavak száma folyamatosan csökken: a mért és a Herdan-törvény segítségével számolt értékek megdöbbentően közel állnak egymáshoz (*q* és *c* paramétereket a korpusz alapján becsültük).



1. ábra. A tesztkorpuszban nem látott szavak arányának csökkenése eredeti korpuszon és a kevert változaton.

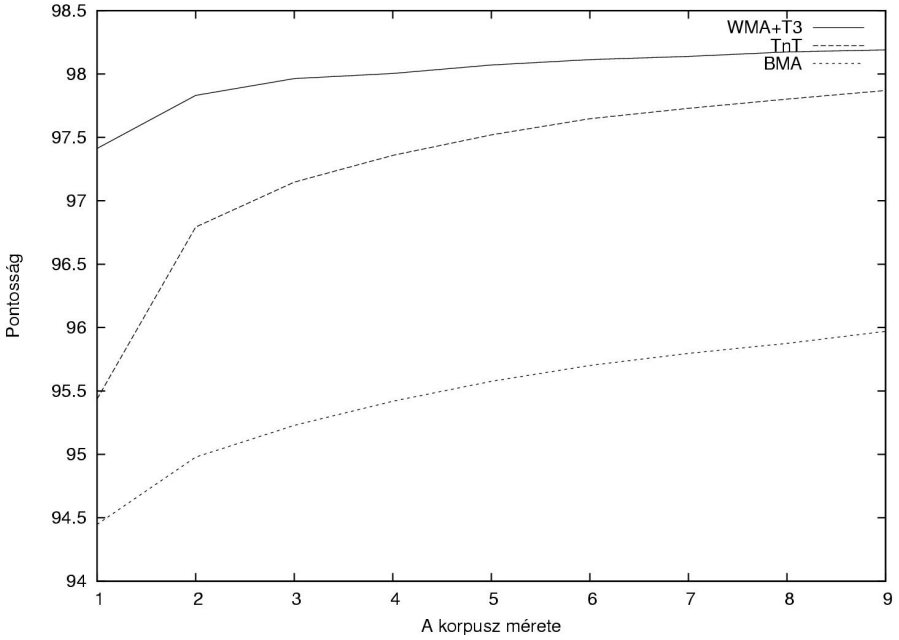
A Szeged Korpusz több, egymástól műfajában és nehézségben teljesen különböző szekcióból áll. Hogy az 1. ábra és 2. ábra görbéit elég nagy korpuszra is fel tudjuk rajzolni, a korpuszt még tanító- és tesztkorpuszra bontás előtt összekevertük. Az ezen

a korpuszon mért pontosság (2. ábrán) nem vethető össze a hagyományos 10-es keresztellenőrzéssel nyert eredményeinkkel, mert a keverés hatására a nem látott szavak aránya nagyon lecsökken a tesztkorpuszban (akár 30%-kal is).

Már [10] is kiemeli, hogy a produktív magyar morfológia miatt a magyar nyelvű korpuszokon nagyobb a nem látott szavak aránya, mint egy ugyanakkora méretű angol korpuszon. (270,830 szövegszó esetén mértek magyarra 17.13%, angolra 4.5%-ot.) Miután a nem látott szavak aránya igen erősen befolyásolja az alapszintűnél összetettebb módszerek hatékonyságát is, alapjában három utat követhetünk:

(A) növeljük a tanítókorpusz méretét, hogy az ilyen szavak arányát csökkentsük,
(B) a nem látott szavakat a már látott szavakkal rokonítjuk, vagy
(C) a nem látott szavakra vonatkozó heurisztikát javítjuk, pl. MA igénybevételével.

Közhelyszámba megy, hogy a gyakorlatban a leghatékonyabb az (A) módszer „there is no data like more data”, és ezt mutatják a mi vizsgálataink is.



2. ábra. Különböző algoritmusok tanulási görbéje kevert korpuszon.

Jó példa a (B) módszerre az alapszintű algoritmus alábbi módosítása (ehhez hasonlót javasol [7] is), amire a későbbiekben `BMA`-ként (baseline MA) hivatkozunk:

1. Ha *w* a tanítókorpuszban szerepel, akkor a $T = \arg\max(T_i \mid w)$ címkét kapja, egyébként
2. ha az MA ismeri és egy címkét rendel a szóhoz, akkor ezt kapja,
3. ha az MA ismeri, de nem egyértelmű a szó, akkor az MA által kiadott $T_{w,i}$ címkék közül a tanítókorpuszban leggyakoribb címkét adjuk, minden egyéb esetben
4. a címkét `NOUN`-nak vesszük.

Ez a módszer a Szeged Korpuszon 95.40%, az *1984*-en pedig 95.84% pontosságot ér el, ami összemérhető a transzformáció-alapú tanuló-rendszerek eredményeivel ([7], [1], [9]), de messze marad a Markov modellel elérhető 98.11%-tól [10]. Mivel a módszer a látott szavakra igen magas pontosságot ad, és a nem látott szavak aránya monoton csökken a korpusz méretének növelésével, a teljes pontosság monoton növelhető a korpusz méretével, ahogy a 2. ábra mutatja.

Ugyanezen az ábrán látható a morfológiai elemző hatása is. Az MA nélkül működő rejtett Markov modellen alapuló `TnT` [2] a `BMA` modell felett teljesít, mert figyelembe tudja venni a szó környezetét is. Ugyanakkor, ha a rejtett Markov modellezést kiegészítjük úgy, hogy a nem látott szavaknál az MA kimeneti címkéire támaszkodjon, hasonlóan [10]-hez, akkor jelentősen megnő a pontosság. Ezt a módszert mi `WMA+T3`-ként jelöltük, mert tekinthető egy súlyozott MA (weighted morphological analyzer) és a három szó méretű kontextust figyelembe vevő Markov-lánc együttesének. Ezt a modellt a következő fejezetben részletesebben mutatjuk be.

A 2. ábrából az is kiolvasható, hogy az MA jótékony hatása a korpusz növekedésével, és így a nem látott szavak arányának csökkenésével egyre kisebb lesz. Ahogy növeljük a korpusz méretét, a `TnT` és a `WMA+T3` hibaszázalékai közötti különbség egyre csökken. Közöttük a fő különbség csupán az, hogy a nem látott szavakra a `WMA+T3` az MA kimeneti címkéi közül tud választani.

A morfológiai egyértelműsítők hibája értelemszerűen a tesztkorpusz olyan szövegszavainál a legnagyobb, amelyek sem a tanítókorpuszban nem szerepeltek (mint láttuk ezek aránya a korpusz növekedésével csökken), sem az MA nem ismeri őket (out of vocabulary, OOV). Ezek aránya a korpusz méretétől független: az ilyenek teszik ki a tesztkorpusz 2%-át. Egy adott korpuszon az OOV tetszőlegesen csökkenthető, sőt akár ki is küszöbölhető az MA tőtárának növelésével (különösen hasznos lehet ez az eljárás az 1984 újbeszédének lefedéséhez). De hosszú távon, dinamikusan növő korpuszon (amilyen például a magyar web) 2% alatti OOV nemigen várható, hiszen a köznyelv állandóan bővül új szavakkal, különösen tulajdonnevekkel. A magyar szófaji címkéző szakirodalomban eddig egységesen követett eljárás, hogy az MA építést előre, a tanító- és a tesztkorpusz különválasztása előtt, a teljes korpusz alapján elvégzik. Ez azonban csupán az OOV problémát a mérésből kiküszöbölő egyszerűsítésnek tekinthető, és ezért az eddigi eredményeknek egy új korpuszon való reprodukálhatósága megkérdőjelezhető.

3. A maxent modell

A maximum entrópia (maxent) modellt szófaji címkézésre először Ratnaparkhi [11] javasolta. Ebben a keretben minden osztályozandó objektumhoz (esetünkben szövegszóhoz) úgynevezett jegyek (predikátumok, angolul features) halmazát rendeljük, és a rendszer ezek alapján tanulja meg a kimeneti címkéket (melyeket szintén jegyként kezel). A jegyek meghatározásakor nemcsak az éppen aktuális szót, hanem annak környezetét (rendszerünkben a közvetlen szomszédait) is figyelembe vehetjük. A maximum entrópia modellezéshez az OpenNLP maxent programkönyvtárat (`http://maxent.sourceforge.net/`) alkalmaztuk.

Míg az előző szakaszban tárgyalt (B) eljárás a morfológiai elemzést csak a tesztszót a már látott tanítószavakkal való rokonítására használja, az alábbiakban javasolt

architektúra inkább a (C) úthoz áll közelebb, amennyiben túllép az MA által adott ambiguitási osztályokon, és a címke-valószínűségekre explicit becslést tesz.

A következőkben a mondatokat szavak $w_1,\ldots w_n$ sorozatának tekintjük, amelyhez tanításkor ismert a $t_1,\ldots t_n$ címke-sorozat. A maximum entrópia modell egy együttes eloszlást határoz meg a lehetséges t_i címkék és az aktuális c_i kontextus között,

$$p(t_i, w_i) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(t_i, c_i)}$$

ahol π egy konstans normalizációs faktor, $\{\alpha_1,\ldots,\alpha_k\}$ a modell paraméterei és a $\{f_1,\ldots f_k\}$ a modellben használt bináris jegyek, amik minden címkére és kontextusra {0,1} értéket vehetnek fel (az 1 érték jelenti az adott predikátum teljesülését). Gyakorlatban a bináris jegyek helyett egyértékű predikátumokat is meg tudunk adni, amik bináris jegyekké alakíthatóak át. Jelenleg a következő jegyeket használjuk:

1. a szóalak kisbetűsítve[59]
2. nem mondatkezdő szó esetén a megelőző szó kisbetűs alakja
3. nem mondatzáró szó esetén a következő szó kisbetűs alakja
4. az MA elemzéseiből alkotott ambiguitási osztály
5. tartalmaz-e a szóalak számot, nemalfabetikus karaktert
6. csupa nagybetűs-e, nagy kezdőbetűs-e
7. ha 5 karakternél hosszabb a szó, akkor az utolsó 2, 3, és 4 karaktere külön-külön

Nem nyilvánvaló, hogy az MA elemzéseit hogyan kell jegyekké alakítani. A legjobb eredményt úgy értük el, ha az MA elemzéseinek halmazát (az ún. ambiguitási osztályt) egyetlen jegyként vettük fel. A szó utolsó néhány karakterére és a felszíni alakra vonatkozó jegyek alapjában az OOV probléma megoldását szolgálják: amikor a szót sem az MA nem ismeri sem a tanítókorpuszban nem szerepelt, akkor a modell csak a környező szavak és végződés adta jegyeket használja.

A tesztkorpusz címkézésénél a maxent modell által meghatározott együttes eloszlás alapján kiszámoljuk, hogy mi a kontextusra jellemző címke-eloszlás, azaz a mondat *i*. szavára, minden egyes lehetséges címkére kiszámoljuk a

$$P(t_i = T_k \mid c_i) = \frac{P(t_i = T_k \mid c_i)}{\sum_{t \in T} P(t_i = T_k, c_i)}$$

[59] A szó, előző szó, következő szó, a szuffixumok, az ambiguitási oszály, stb. mind predikátumok, amelyekből annyi különböző jegy lesz, ahány különböző szótípus, megelőző szótítpus, stb. található a korpuszban; a továbbiakban ezt a megkülönböztetést nem jelöljük.

valószínűséget. A maxent modell tehát nem hoz döntést, csupán minden egyes lehetséges címkére megadja annak valószínűségét. A maxent modell – bár jegyként megkapja az MA által adott címkéket – a tanítókorpuszban látott minden címke-típushoz pozitív valószínűséget rendel.
Első modellünk, a továbbiakban MA+ME, egyszerűen a fenti maxent modell alapján egy szóhoz a következő címkét rendeli:

1. Ha az MA ismeri a szót, akkor ezek közül választjuk a maxent modell által legvalószínűbbnek tartott címkét. (Speciálisan, ha az MA csak egyetlen elemzést ismer, akkor azt választjuk.)
2. OOV szóalak esetében a maxent modell választ.

Ez a modell csak lokális információkra hagyatkozik: egy adott szó címkézésénél nem veszi figyelembe a szó kontextusában lévő szavak címkéjét, ellentétben például a HMM alapú TnT-vel. Ezért két további modellt javasolunk.

A WMA+T3-nak nevezett modell a maxent modell és egy trigram-simítás kombinálása. A maxent modell és az MA kombinálásával súlyozott MA-t (Weighted Morphological Analyzer, WMA) építhetünk, amely a szóhoz hozzárendeli címkék egy valószínűségeloszlását, az alábbi módon:

1. Ha a szó szerepelt a tanítókorpuszban, akkor a szó címkéinek valószínűségét maximum likelihood módszerrel becsüljük, mint az alapszintű módszereknél.
2. Ha az MA ismeri a szót, akkor pontosan az általa kiadott címkéket engedjük meg, és a maxent által ezekre adott valószínűségeket egyre normalizáljuk. Speciálisan, ha az MA csak egyetlen elemzést ismer, akkor annak egy valószínűséget adunk.
   Előfordulhat, hogy az MA olyan címkét ad ki, amit a maxent modell a tanítókorpuszban nem látott. Ennek most mi egy konstans valószínűséget adunk normalizálás előtt.
3. OOV szóalak esetében a maxent modell által legvalószínűbbnek ítélt három elemzést engedjük meg, és ezeket normalizáljuk.

A WMA tehát minden egyes szóra megadja lehetséges címkéit súlyokkal. A címkék közül ki kell választani azokat, amik megadják a mondathoz rendelhető legvalószínűbb címke-szekvenciát. Formálisan:

$$\arg\max P(t_1,...,t_n \mid w_1,...,w_n) = \arg\max P(w_1,...,w_n \mid t_1,...,t_n)P(t_1,...,t_n),$$

ahol az első a szorzat első tagját a WMA kimenete, a másodikat a tanítókorpuszban látott címke-szekvenciák alapján épített másodrendű Markov modell szolgáltatja. A Markov modell építéséhez, és a legvalószínűbb szekvencia megkereséséhez (Viterbi algoritmussal), a SRILM[60] programcsomagot használtuk. Ennél a modellnél a maxent modellből ki kell hagyni a megelőző és következő szó jegyeit (tehát a WMA

[60] http://www.speech.sri.com/projects/srilm/

kontextusfüggetlen), hogy a kombinált modellben a két komponens független legyen. A `WMA+T3` modell gyakorlatilag analóg Oravecz és Dienes [10] modelljével.

Az utolsó modellünk, a `TNT+MA+ME`, szintén érzékeny a címke-szekvenciára. Az előbbiekben bemutatott `MA+ME` modell jegyei közé felvesszük még a szó, a megelőző, és a következő szó címkéit. Tanítási fázisban ezek adottak, címkézéskor pedig ezeket a jegyeket a tanítási korpuszon betanított `TnT` modell jósolja meg.

4. Értékelés

Ahhoz, hogy a Szeged Korpuszt, mint tanító- és tesztkorpuszt alkalmazni tudjuk, konverzióra volt szükség az MSD címkék és hunmorph által használt KR címkék [8] között. A konverzió nem teljesen triviális feladat, mert a két rendszer még az inflexiós kódok tekintetében sem vág teljesen egybe (pl. a marginális esetragok és a familiáris többes kezelésében).

A reziduális főkategóriájú (`X`, `Z`, `O`) MSD-címkéket tartalmazó mondatokat elhagytuk a korpuszból. A hunmorph ugyan számos `X` elemet (ismeretlen szó) felismer, és a vele közös tőtárú hunspell számos `Z` (sajtóhiba) elemet ki tud javítani, de célunk nem az előfeldolgozás, hanem a morfológiai egyértelműsítés vizsgálata, és ezekhez az elemekhez a Szeged Korpusz nem adja meg azt a javított kódot (ground truth), amivel rendszerünk eredményeit össze lehetne hasonlítani. Az `O` főkategóriájú nyílt címkeosztály esetében pedig úgy tapasztaltuk, hogy a Szeged Korpusz szerkesztési elvei még nem teljesen kiforrottak ezekre nézve, ezek az elemek még manuálisan sem különíthetőek el megfelelő pontossággal egymástól és más kategóriáktól.

Az eredeti Szeged Korpusz 82,098 mondatából így végül 70,084 mondatot tartottunk meg. A korpuszból elhagyott mondatokat későbbi robusztussági tesztjeinkhez alkalmaztuk, `hard` részkorpusz néven. Bár szemünkben a tulajdonnévi csoportok kijelölése (named entity recognition) is külön feladat lenne, megtartottuk a szóközt tartalmazó tokeneket, amelyek a korpusz 1.37%-át teszik ki. Mivel az általunk használt MA ezeket nem ismeri, ezek méréseinkben garantáltan az OOV szavak számát növelik.

Összességében 1001 MSD címkét 744 KR címkére konvertáltunk, ami látszólag egyszerűsíti a címkézési feladatot, valójában azonban nem, mert a KR címke és a tő ismeretében az MSD címke gyakorlatilag 100%-ban visszaállítható, azaz nincs két címke összevonásából adódó információvesztés. Másképpen fogalmazva: egy adott százalékban korrekt KR címkézés mechanikusan, egy statikus táblázat segítségével ugyanilyen, vagy még nagyobb százalékban korrekt MSD címkézéssé alakítható.

1. táblázat. A modellek pontossága a Szeged Korpusz szekcióin.

szekció	méret	oov	alapszint	BMA	TnT	MA+ME	WMA +T3	TNT+MA +ME
irodalom	209785	5.79	86.20	95.46	96.02	97.37	97.63	**97.83**
iskola	290167	1.62	90.17	96.34	96.97	97.73	97.80	**98.01**
Sajtó	355311	9.98	82.68	94.36	97.32	97.93	98.14	**98.38**
számtech	157969	8.43	86.06	94.44	97.02	97.53	97.91	**98.11**
Jog	147766	4.97	91.41	96.89	98.44	98.76	98.96	**99.04**
teljes	1161016	5.64	89.70	95.40	97.42	97.72	97.93	**98.17**

Az egyes részkorpuszokat jellemző méret és OOV adatok után a két alapszintű modell (MA nélküli és MA-val működő) és négy statisztikai modell eredményeit közöljük: `TnT` a Brants-féle trigram modell, `MA+ME` a tisztán maxenten alapuló, a `WMA+T3` egy MA-t használó saját trigram modell, `TNT+MA+ME` pedig a `MA+ME` modell, amely a `TnT` kimenetét is megkapja bemeneti jegyként. A rendszerek hatékonysági sorrendje a szekció kiválasztásától teljesen függetlennek bizonyult.

A táblázatban látható, hogy a morfológiai egyértelműsítésnél fontos a címkeszekvencia mint információforrás. A `MA+ME` modell csak lokális információk alapján dönt, a környező szavak címkéjét nem veszi figyelembe. Ezzel szemben a `WMA+T3` és a `TNT+MA+ME` modellek nem szavanként hoznak egymástól független döntéseket, hanem az egész mondatra határozzák meg a legjobb címke-szekvenciát.

A tisztán statisztikai `TNT+MA+ME` pontossága felülmúlja az összes általunk ismert szabálytanuló rendszerét: [9] 96.52% pontosságot ér el a teljes Szeged Korpuszra és 98.26%-t a hírekre. [7] 98.03%-os pontosságot ér el az *1984* feladaton, ahol mi jelenlegi módszertanunk mellett csupán 97.91%-ot mérünk. Ehhez a korpuszból idealizált (azaz a tesztanyag minden szavát garantáltan ismerő) MA-t épít az egyértelműsítés fázisa előtt. Ha a rendszerünkben használt független MA-t kicseréljük egy korpuszból épített morfológiai szótárra, akkor [7]-tel immáron azonos feltételek mellett 98.50%-os pontosságot érünk el.

A robusztusságukat ellenőrizendő a rendszereink pontosságát megmértük a teljes `hard` részkorpuszon tesztelve, a standard korpusz megfelelő méretű véletlenszerűen választott részén tanítva, a pontosságba nem mérve bele a kezelhetetlen címkéket. Azt tapasztaltuk, hogy a `TNT+MA+ME` pontossága ebben a felállásban 97.80%, ami csupán fél százalékpontnyi csökkenés az ugyanekkora, véletlenszerűen választott tanító- és tesztkorpusszal mért 98.31%-os eredményhez képest. A kontextust kevésbé figyelembe vevő `MA+ME` esetében a csökkenés nagyobb, itt 97.87%-ról 96.93%.-ra változik a pontosság.

Az eredményekből látható, hogy a tisztán statisztikai elven működő modellek eredményesen kombinálhatóak az erőforrás alapú morfológiai elemzővel. Magyar nyelvre ezt először [10] demonstrálta. Modelljeink előnye az általunk alkalmazotthoz képest abban áll, hogy az OOV szavakat is képesek robusztusan kezelni. Eredményeink nem teljesen hasonlíthatóak össze, mert méréseinket más (bár hasonló méretű és jellegű) korpuszokon végeztük. A legjobb rendszerünk teljes Szeged Korpuszon mért 98.17% pontossága az OOV szavak kezelésén túl azért is kiemelkedő, mert műfajában nagyon különböző összetevőkből álló heterogén korpuszon keresztértékeléssel értük el ezt az eredményt. Így módszerünk remélhetőleg lehetővé teszi az eddiginél nagyobb változatosságú, például a dinamikusan növekvő magyar web kiaknázásával épült korpuszok morfológiai elemzését is.

5 Köszönet

Szarvas Györgynek és Vajda Péternek a korpuszért és annak átalakításáért, Trón Viktornak a morfológiai elemző beépítésében nyújtott segítségéért és Oravecz Csabának értékes tanácsaiért.

Irodalomjegyzék

[1] Kuba András, Bakota Tibor, Hócza András, and Csaba Oravecz. A magyar nyelv néhány szófaji elemzőjének összevetése. I. Magyar Számítógépes Nyelvészeti Konferencia, pages 16–22. 2003.

[2] T. Brants. TnT – a statistical part-of-speech tagger, 2000.

[3] Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, pages 315–319, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] Csendes Dóra, Hatvani Csaba, Alexin Zoltán, Csirik János, Tibor Gyimóthy, Prószéky Gábor, and Tamás Váradi. Kézzel annotált magyar nyelvi korpusz: a szeged korpusz. In II. Magyar Számítógépes Nyelvészeti Konferencia, pages 238–245. Szegedi Tudományegyetem, 2003.

[5] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Szógyakoriság és helyesírás-ellenőrzés. In Proceedings of the 1st Hungarian Computational Linguistics Conference, pages 211–217. Szegedi Tudományegyetem, 2003.

[6] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In Proceedings of Language Resources and Evaluation Conference (LREC04). European Language Resources Association, 2004.

[7] Tamás Horváth, Zoltán Alexin, Tibor Gyimóthy, and Stefan Wrobel. Application of different learning methods to Hungarian part-of-speech tagging. In ILP, pages 128–139, 1999.

[8] András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, and Viktor Trón. Általános célú morfológiai elemző kimeneti formalizmusa. II. Magyar Számítógépes Nyelvészeti Konferencia, pages 172–176. Szegedi Tudományegyetem, 2004.

[9] András Kuba, László Felföldi, and András Kocsor. Pos tagger combinations on hungarian text. In 2nd International Joint Conference on Natural Language Processing, IJCNLP, 2005.

[10] Csaba Oravecz and Péter Dienes. Efficient stochastic part-of-speech tagging for Hungarian. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002), pages 710–717, 2002.

[11] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[12] Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*, 2005.

[13] D. Tufis, P. Dienes, C. Oravecz, and T. Váradi. Principled hidden tagset design for tiered tagging of Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, 2000.

[14] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, and Vajda Péter. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, 2005. megjelenés alatt.

Ismert névelemek felismerése és morfológiai annotálása szabad szövegben

Tikk Domonkos[1], Szidarovszky Ferenc P.[2], Kardkovács Zsolt Tivadar[1], Magyar Gábor[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, H-1117 Budapest, Magyar Tudósok krt. 2.
{tikk, kardkovacs, magyar}@tmit.bme.hu
[2] Szidarovszky Kft. H-1392 Budapest, Pf. 283.
ferenc.szidarovszky@szidarovszky.com

Kivonat: „A szavak hálójában” projekt[61] keretében készülő internetes keresőszolgáltatásnak egyik célja az, hogy lehetőséget nyújtson természetes nyelvű magyar kérdésekkel internetes adatbázisok tartalmában az ún. mélyhálóban való keresésre. Az adatbázisokból ki lehet nyerni azokat az *egyedi azonosítókat*, amelyek együttese lehetővé teszi, hogy a felhasználói keresések információigénye és a mélyhálós tartalmak között kapcsolatot teremtsünk. Az egyedi azonosítókat *névelemnek* nevezzük. A természetes nyelvű kérdések feldolgozásának kiemelt fontosságú része a bennük szereplő ismert névelemek felismerése, valamint a kérdésben betöltött szerepük meghatározásához a felismert névelemek morfológiai jegyeinek meghatározása. Cikkünkben bemutatjuk a probléma megoldására javasolt és megvalósított algoritmusunkat, amely számítási igényt tekintve is hatékonyan oldja meg a felvázolt feladatokat.

1 Bevezetés

„A szavak hálójában” projekt[2] keretében készülő internetes keresőszolgáltatást végző alkalmazásnak egyik célja az, hogy lehetőséget nyújtson természetes nyelvű magyar kérdésekkel internetes adatbázisok tartalmában az ún. mélyhálóban való keresésre. Az adatbázisokból az adatgazdákkal történő együttműködés eredményeként ki lehet nyerni azokat az *egyedi azonosítókat* pl. könyvtári adatbázis esetén a szerzők, kiadók, műcímek stb. nevének listáját , amelyek együttese lehetővé teszi, hogy a felhasználói keresések információigénye, és a mélyhálós tartalmak között kapcsolatot teremtsünk, és ezzel a kérdés megválaszolását megkönnyítsük. A mélyháló jellegzetességei és keresésének jelentősége [1, 6], valamint a projekt keretében kidolgozott mélyhálós internettartalmak keresését végző rendszerünk [4, 5] felől érdeklődő Olvasók számára – terjedelmi okok miatt – a megadott irodalmi forrásokat ajánljuk.

Cikkünk felépítése a következő. Először a 2. szakaszban meghatározzuk az általunk feldolgozott névelemek körét, és ismertetjük, hogy milyen problémákat kell

[61] NKFP-0019/2002 projekt

megoldania a névelem felismerő algoritmusnak. A 3. szakaszban részletesen ismertetjük az általunk javasolt algoritmust, majd a 4. szakaszban működését példákon keresztül is bemutatjuk. Végül az 5. szakaszban röviden összegezzük a cikk lényeges eredményeit.

2 Névelemek és felismerésük problematikája

Az egyedi azonosítókat *szótári*, vagy *ismert névelemnek* nevezzük, amelyeket a *névelemtárban* tárolunk. A szótári jelzőt a *minták alapján felismert névelemektől* (pl. dátumok, postai és internetes címek, stb.) való megkülönböztetésre használjuk, hangsúlyozandó azt, hogy a névelemtárban szereplő névelem bejegyzéseket szótári (kanonikus) alaknak tekintjük. A szótári névelemek nagy részét a fenti meghatározás miatt a tulajdonnevek teszik ki, azonban alkalmazásunkban a fogalomba beleértjük az olyan rögzített alakú közneveket is, amelyeknek kiemelt szerepe van bizonyos minták alapján felismert névelemtípusok (mennyiségek, címek, stb.) és egyéb, az elemzett kérdés további feldolgozása szempontjából fontos fogalmak azonosítása során. Eszerint névelemnek tekintjük pl. az alábbi csoportokba tartozó közneveket: a pénznemek jelölései (forint, euró, stb.), nemzetiségnevek (magyar, szlovák, angol, stb.), közterülettípus (út, utca, tér, stb.), stb.

A névelemtárnak az adatbázisból történő feltöltése során szemantikai információkat rendelünk az egyes elemekhez, amelyeket az adat adatbázisbeli séma- és attribútum-információiból nyerünk ki. A névelemtárban lehetőség van a kanonikus alak lehetséges szinonimáinak[62] megadására is (pl. *Petőfi Sándor* bejegyzéshez *Petőfi* szinonima, vagy a *forint* bejegyzéshez a *HUF* szinonima).

A névelemtár elemei meghatározzák azt az információs teret, amelyben a felhasználó kérdésére választ tudunk adni. Ez azt jelenti, hogy csak azokat a kérdéseket tudjuk megválaszolni a mélyhálós tartalmak segítségével, amelyekben ezen tartalmakból kinyert névelemek szerepelnek. Összességében az alábbi megszorításokat tesszük a felhasználó kérdéseire vonatkozóan, a listában szerepelnek a tartalmi vonatkozású megkötések is:

- csak egyszerű, azaz nem összetett mondatokat fogadunk el;
- csak helyesen írt, és nyelvtanilag helyes mondatokat fogadunk el;
- csak kérdőszóval kezdődő, nem eldöntendő kérdést fogadunk el; a lehetséges kérdőszavakat is korlátozzuk;
- Szubjektív (*Hány éves a kapitány?*), ok-okozati viszonyra irányuló (*Miért tört ki a II. világháború?*), vagy egyéb nem tényszerű, illetve nem a fenti információs térben található mondatok helyes megválaszolását nem garantáljuk.

A természetes nyelvű kérdések feldolgozásának tehát kiemelt fontosságú része a bennük szereplő ismert névelemek felismerése, valamint a kérdésben betöltött szerepük meghatározásához a felismert névelemek morfológiai jegyeinek meghatározása .Ez a toldalékoló magyar nyelv esetén korántsem egyszerű feladat, mivel a névelemek nem feltétetlen rögzített alakjukban (beleértve a szinonimákat) fordulnak elő, hanem többnyire toldalékolt alakban. A toldalék megváltoztathatja a névelem szótövét, illetve ha a szótári alak már eleve toldalékolt, akkor ezt is módosíthatja[63]. Továb-

[62] Nem toldalékolt alakok, csak lehetséges különböző előfordulásai a kanonikus alaknak

[63] Ld. *Vissza a jövőbe* és *Hol adják a Vissza a jövőbét?*

bi gondot jelenthet az egymásba ágyazott névelemeknél a névelem határainak meghatározása[64] [3]. Ha ez utóbbi esetben több értelmezés lehetséges, akkor alternatívákat állítunk elő. A morfológiai jegyek meghatározásánál a nem alanyesetű kanonikus alakok és a nem magyar (azaz morfológiai elemző által fel nem ismert) névelemek *speciális* esetei kívánnak külön megfontolást.

Cikkünkben bemutatjuk a probléma megoldására javasolt és megvalósított algoritmusunkat, amely azon kívül, hogy a fenti feladatokat megoldja, mindezt a számítási igényt tekintve hatékonyan valósítja meg. Az ismertetett módszer a HunMorph [2] szabad forráskódú statisztikai alapú morfológiai elemzőt használja, ennek megfelelően a példáknál található morfológiai elemző eredmények is a HunMorph kódolása szerint vannak megadva.

Fontosnak tartjuk kiemelni, hogy a módszer *nem felügyelt tanuláson alapul*, mivel célja nem ismeretlen névelemek felismerése, hanem az ismertek pontos azonosítása.

3 Szótári névelemek felismerése

A szótári névelem (ezentúl itt csak *névelem*) felismerőnek két fő célja van:

– *keresés*: a mondatban szereplő névelemek megtalálása;

– *annotálás* (vagy *címkézés*): a névelemek morfológiai jegyeinek meghatározása.

A keresés és annotálás folyamata általában összekapcsolódik, így önmagukban nem hajthatók végre.

Mivel egy névelem több szóból is állhat, a kérdőmondat tetszőleges szegmense (szavak rögzített sorrendű sorozata) lehet névelem. Egy n szavas kérdőmondat szegmenseinek száma $n(n+1)/2$. Egy átlagos kérdőmondat 7-10 szóból áll, míg a névelemtár mérete 10^6 nagyságrendű is lehet. Így sokkal hatékonyabb a mondatszegmensekből kiindulva keresni, mint a névelemtárból kiindulva. Egy kifejezés keresése a névelemtárban gyorsítható a névelemtár elemeinek hash-elésével. A mondatszegmensek összevetése a névelemtárral a szegmensek hossza szerint csökkenő sorrendben történik.

A névelem felismerés egy másik problémája az, hogy egy névelem tartalmazhat egy másikat (pl. a *The New York Times* egy napilap). Míg a Blitz NL feldolgozó [3] a felismert névelemek közül csak egyet választ ki konfidencia értékek alapján, mi fel kívánjuk ismerni az összes névelemet, különböző mondat alternatívákat létrehozva. Ebből kifolyólag az összevetés a keresés eredményétől függetlenül tovább folytatódik a rövidebb szegmensekkel.

A szegmensek összevetése az alábbi sorrendben történik:

1. A teljes mondattal kezdjük: [1,...,n], és vesszük az első szóval kezdődő egyre rövidebb szegmenseket: [1,...,j], ahol $j=n$-1,...,1.
2. Vesszük a második szóval kezdődő egyre rövidebb szegmenseket: [2,...,j], ahol $j=n$,...,2.
3. Általánosan, az összes szegmenst megvizsgáljuk a kezdőszó mondatbeli pozíciója szerint növekvő, majd azon belül a szegmens hossza szerint csökkenő sorrendben: [i,...,j], ahol i=3,...,n, $j=n$,...,i.

[64] *New York Times sport rovata* tartalmazza a New York, York, Times, és New York Times-t.

1. megjegyzés: Nyilván nem mindegyik mondatszegmens lehet valóban névelem. Ha figyelembe vesszük, hogy a mondat első szavának a megszorítások miatt feltétlenül kérdőszónak kell lennie, akkor kezdhetünk a 2. lépéssel ([2,...,*n*] szegmenstől), a vizsgálandó részletek számát $n(n-1)/2$-re csökkentve.

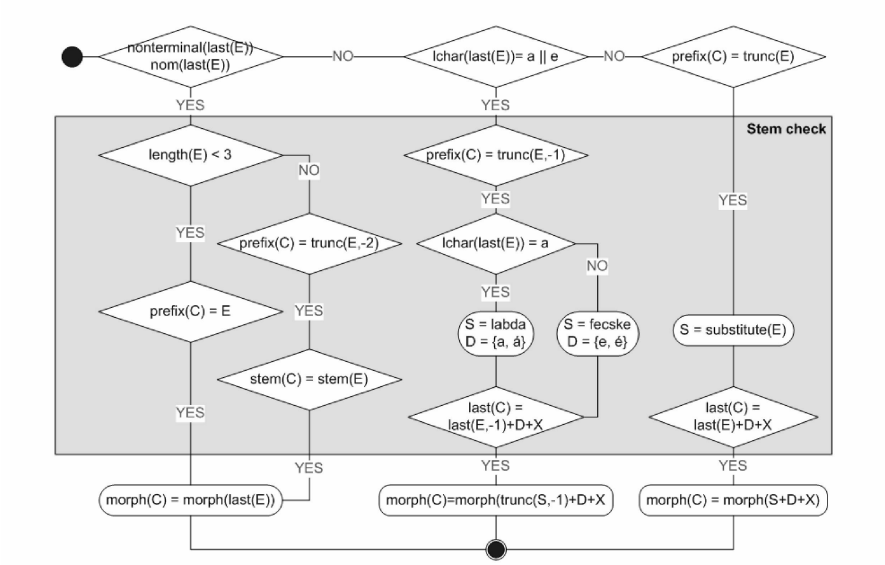
A továbbiakban a névelem felismerést egy konkrét mondatszegmens (ezentúl *jelölt*) kapcsolatban ismertetjük. A magyar nyelvben a szavak töve változhat toldalékolásnál. Az esetek nagy részében a szótőnek csak az utolsó két betűje változhat (*tűz* □ *tüzet*; *álom* □ *álmot*). Hasonlóan, egy toldalék megváltozhat egy következő toldaléktól (ez csak akkor fordul elő, ha a névelem magában is toldalékolt, és azt a mondatban tovább toldalékoljuk, ld. 3. lábjegyzet). Ebben az esetben csak az utolsó betű változhat. Mindezeket a névelem felismerés keresés fázisában figyelembe kell vennünk.

A névelemek jelentős része nem magyar nyelvű, így a morfológiai elemző nem képes azokat elemezni. Ennek ellenére a névelem felismerő ezen névelemeket is el kell lássa morfológiai jegyekkel. Erre a feladatra ún. *helyettesítő szavakat* használunk. A helyettesítő szónak a névelemek toldalékainak meghatározásánál van szerepe. Feltételezzük, hogy minden névelemhez rendelkezünk egy helyettesítő szóval, mely morfológiailag elemezhető és pontosan ugyanúgy ragozódik (kiejtés szerint azonos hangrendű, főnév), mint a névelem utolsó szava. Ez gyakran a névelem utolsó szava (ha az egy alanyesetű magyar főnév), vagy algoritmikusan előállítható mikor a névelem bekerül a névelemtárba. A helyettesítő szónak mindig főnévnek kell lennie, mivel az ismert névelemek előfordulásai egyedi entitásokat jelölnek, tehát a mondatban főnévi szerepben állnak és eszerint ragozódnak. Kivételt képeznek a 2. szakaszban ismertetett egyéb névelemtípusokat egyes esetei, de ezek a morfológiai elemző által ismert magyar szavak, ahol tehát a morfológiai jegyek megállapítására nincs szükség helyettesítő szóra.

Az alábbi jelöléseket használjuk:

– last(x) jelöli az x kifejezés utolsó szavát.
– length(x) jelöli az x szó betűinek számát.
– trunc(x,i) jelöli az x szót az utolsó i betűje nélkül.
– lchar(x) jelöli az x szó utolsó betűjét.

Továbbá jelölje *C* a jelöltet, *S* a helyettesítő szót és *E* a névelemet. Az algoritmus folyamatábráját az 1. ábra szemlélteti:



1. ábra Az algoritmus folyamatábrája

1. Ha last(*E*) toldalékolható, alanyesetű, magyar szó (azaz a morfológiai elemző felismeri)
   1.1 keresés
      1.1.a ha length(last(*E*)) ≥ 3, ellenőrizzük, hogy *C* trunc(*E*,2)-vel kezdődik-e.
      1.1.b ha length(last(*E*)) < 3, ellenőrizzük, hogy *C E*-vel kezdődik-e.
   1.2 szótő ellenőrzés: Ha 1.1.a igaz, azaz *C* trunc(*E*,2)-vel kezdődik, akkor meg kell határozni, hogy last(*C*) és last(*E*) szótöve megegyezik-e. Erre azért van szükség, mert a betűelhagyás miatt a csonkolt szó több értelmes szónak is a prefixe lehet. Ez a lépés kihagyható, ha 1.1.b igaz.
   1.3 annotáció: Ha 1.2-ben a szótövek megegyeznek, akkor *C* az *E* névelem, melynek morfológiai jegyei a last(*C*) jegyei. Ha *E* és *C* egyaránt rendelkezik nem záró morfémával, azt kihagyjuk az annotációból (lásd 4. példa).
2. Ha last(*E*) nem felel meg az 1. feltételeinek, azaz a morfológiai elemző nem ismeri fel, vagy nem toldalékolható, vagy nem alanyesetű.
   2.1 keresés
      2.1.a Ha lchar(last(*E*)) = *a* vagy = *e*, ellenőrizzük, hogy *C* trunc(*E*,1)-vel kezdődik-e.
      2.1.b Ha lchar(last(*E*)) ≠ *a* és ≠ *e*, ellenőrizzük, hogy *C E*-vel kezdődik-e.
   2.2 helyettesítő szó megállapítása
      2.2.a Ha 2.1.a igaz és lchar(last(*E*)) = *a*, akkor S = *labda*, ha lchar(last(*E*)) = *e*, akkor S = *fecske*.
      2.2.b Ha 2.1.b igaz, akkor vesszük a névelemtárban *E*-hez megadott *S*-t.
   2.3 annotáció

2.3.a *C* utolsó szavának alakja a következő: [trunc(last(*E*),1){*a*,*e*}*marad*], ahol *marad* a (*C*) végén lévő maradék betűkből áll (ha vannak). A következő szövegeket elemeztetjük a morfológiai elemzővel: [trunc(last(*S*),1){*á*}marad], ill. [trunc(last(*S*),1){*é*}marad] ha lchar(*E*) = *a*, ill. lchar(*E*) = *e*, azaz a szóvégi magánhangzót hosszúra cseréljük . Csak az egyik szöveg lesz helyes szó, és ismeri fel a morfológiai elemző. A *C* morfológiai jegyei a helyes szó jegyei lesznek.

2.3.b *C* utolsó szavának alakja a következő: [last(*E*) *marad*]. A következő szöveget elemeztetjük a morfológiai elemzővel: [*S marad*]. A *C* morfológiai jegyei az [*S marad*] szó jegyei lesznek.

1. megjegyzés: Látható, hogy az első esetben a keresés bonyolultabb, mert a toldalékolható szavak esetén a helyes szótő azonosítása nehezebb. A második esetben viszont az annotálás a bonyolultabb, mert a toldalékok meghatározása csak egy megfelelő helyettesítő szóval lehetséges.

2. megjegyzés: A névelemek keresett alakja a névelemtár feltöltésekor számítható és tárolható, így jelentős időt nyerünk a keresésnél.

3. megjegyzés: A 2.3-nál ha length(*marad*) = 0, akkor kihagyható a morfológiai elemző használata, mert ez azt jelenti, hogy a névelemen nincsenek toldalékok és az egy alanyesetű főnévnek tekinthető.

4. megjegyzés: A 2.2.b-ben használt, a névelemhez rendelt helyettesítő szó meghatározásánál egy fél-heurisztikus algoritmust használunk. A helyettesítő szavakat már a névelemtár feltöltésekor offline, a névelem utolsó mássalhangzója és az utolsó szavának magánhangzói alapján határozzuk meg. Míg ez (kiejtett) magánhangzóra végződő szavak esetén triviális, mássalhangzóra végződő szavak esetén több körültekintést igényel. Ez az eljárás pl. az idegen szavak kiejtés követő toldalékolása miatt nem 100%-osan tökéletes, de az esetek túlnyomó többségében (több mint 98%-ban) jó helyettesítő szavakat eredményez.

4 Példák

A továbbiakban néhány példán keresztül bemutatjuk az algoritmus működését.

1. példa: Lásd 2. ábra

Milyen költők vannak Arany Jánostól József Attiláig?

E = *József Attila*, last(*E*)-t felismeri a morfológiai elemző mint

Attila[noun_prs]+[NOM]

így ez az 1-es eset. A keresés *József Atti* kifejezéssel végezzük,, ami alapján a *C* = *József Attiláig* szegmenst találjuk (mivel ezekben a példákban a *C* választása triviális, a következőkben külön nem térünk ki rá). A last(*C*) morfológiai elemzése

Attila[noun_prs]+[TERM]

Így az *E* névelemet felismertük *C*-ben és a morfológiai jegyei [TERM].

2. példa: Lásd 2. ábra

Ki rendezte az Anyádat ist?

E = *Anyádat is*, ez a 2 (b) eset, mert az *is* kötőszó, mely nem toldalékolható. Legyen S a *kés*, így a morfológiai elemzővel a *kést* szöveget elemeztetjük. Az eredmény

kés[noun]+[ACC]

így a felismert névelem: *Anyádat is*$_{\text{névelem}}$+[ACC].

3. példa: Lásd 2. ábra

Mennyit kell fizetnem az Interjú a vámpírralért?

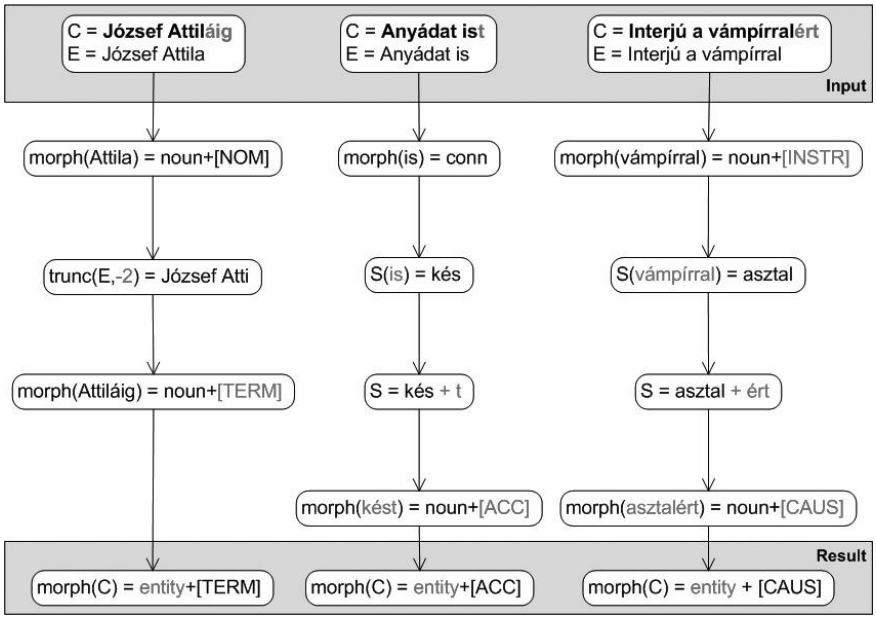
E = *Interjú a vámpírral*, ez is a 2 eset, mert last(*E*) már toldalékolt:

vámpír[noun]+[INSTR]

Legyen S az *asztal*, így a morfológiai elemzővel az *asztalért* szöveget elemeztetjük. Az eredmény

asztal[noun]+[CAUS/FIN]

így a felismert névelem: *Interjú a vámpírral*$_{\text{névelem}}$+[CAUS/FIN].



2. ábra Illusztráció az 1-3. példákhoz

4. példa: Lásd 3. ábra

Ki rendezte Az én kis mosodámat?

E = Az én kis mosodám. A névelem utolsó szava birtokos toldalékú, amit a névelem egészére mint entitásra vonatkozóan tárgyrag követ. Ebből következően a névelemet csak a tárgyraggal kell felcímkézni. Az utolsó szó morfológiai elemzése a névelem az algoritmus mindkét fő ágát aktiválja, hiszen

mosoda[noun]+[POSS_SG_1]+[ACC]
mosoda[noun]+[POSS_SG_1]+[NOM]

Az első sor a 2-es esetet aktiválja. Legyen S a *karám*, így a morfológiai elemzővel a *karámat* szöveget elemeztetjük. Mivel ezt a szót a morfológiai elemző nem ismeri fel, ez az ág nem talál névelemet.

A második sor az 1-es esetet aktiválja. A last(*E*) = *mosodám* és last(*C*) = *mosodámat* szótöve egyezik, és *C E*-vel kezdődik. Végül a morfológiai jegyeket a last(*C*) és last(*E*) morfológiai jegyeinek különbözetéből kapjuk: *Az én kis mosodám*$_{\text{névelem}}$+[ACC].

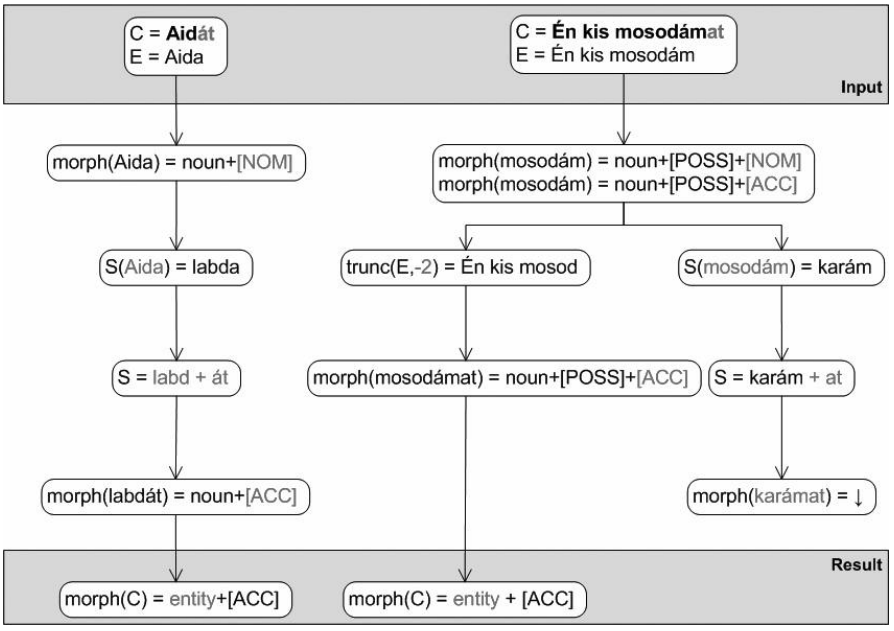
5. példa: Lásd 3. ábra

Hol játsszák az Aidát?

E = *Aida*. Ez a 2 (a) eset, mert last(*E*)-t nem ismeri fel a morfológiai elemző. Legyen S a *labda*, így a morfológiai elemzővel a *labdát* szöveget elemeztetjük, mely eredménye

labda[noun]+[ACC]

Így a névelem felismerés eredménye: $Aida_{névelem}$+[ACC].



3. ábra Illusztráció a 4-5. példákhoz

5 Összefoglalás

A fentiekben ismertettük annak a feladatnak a jelentőségét és nehézségeit, mely egy természetes magyar nyelvű kérdőmondatban a szótári névelemek összes előfordulásának megkeresése és morfológiai jegyekkel való ellátása.

Ismertettünk egy algoritmust, mely megoldás erre a feladatra, és hatékonyan végrehajtható.

6 Köszönetnyilvánítás

A cikk a Nemzeti Kutatási és Fejlesztési Pályázatok NKFP-0019/2002 jelű projektjének támogatásával készült.

Irodalomjegyzék

1. Bergman, M.K.: The deep web: surfacing hidden value. Journal of Electronic Publishing 7 (2001) http://www.press.umich.edu/jep/07-01/bergman.html.
2. Hunmorph: (2004) http://mokk.bme.hu/resources/hunmorph/
3. Katz, B., Yuret, D., Lin, J., Felshin, S., Schulman, R., Ilik, A.: Blitz: A preprocessor for detecting context-independent linguistic structures. In: Proc. of the 5th Pacific Rim Conference on Artificial Intelligence (PRICAI '98), Singapore (1998)
4. Tikk, D., Kardkovács, Zs.T., Andriska, Z., Magyar, G., Babarczy, A., Szakadát, I.: Natural language question processing for hungarian deep web searcher. In: Proc. of IEEE Int. Conf. on Computational Cybernetics (ICCC04), Wien, Austria (2004) 303–309.
5. Tikk, D. Kardkovács, Zs.T., Magyar, G.: Deep web searcher for Hungarian. International Journal of Information Technology 1(4) (2004) 191--197.
6. Winkler, H.: Suchmaschinen. metamedien im internet? In Becker, B., Paetau, M., eds.: Virtualisierung des Sozialen, Frankfurt/NY (1997) 185–202 (In German; English translation: http://www.uni-paderborn.de/~timwinkler/suchm_e.html).

Tundrai nyenyec morfológiai elemző és generátor

Novák Attila[1] és Wenszky Nóra[2]

MorphoLogic Kft. 1126 Budapest Orbánhegyi út 5.
[1]novak@morphologic.hu,
[2]nora@nytud.hu

Kivonat: Ebben a cikkben egy az Uráli nyelvcsalád északi szamojéd ágához tartozó *tundrai nyenyec* nyelvű szóalaktani elemzőprogram és szóalakgenerátor létrehozásáról számolunk be. A nyelv bemutatása és az elemző alapjául szolgáló korábbi munkák ismertetése után részletesen tárgyaljuk az elemzőprogram egyes moduljait és működését. Az elemző lexikona mintegy 19 500 tő mögöttes fonológiai reprezentációját tartalmazza, melyek 266 különböző ragozási osztályba sorolhatók. A toldaléktár 254 mögöttes toldalékalakot tartalmaz. Az elemző a nyelv inflexiós jelenségeit teljes körűen kezeli, beleértve az általában a szóképzés körében tárgyalt igenevek és gerundiumok kezelését is.

1 Bevezetés

A tundrai nyenyec szóalaktani elemzőprogram, mely erre a nyelvre az első ilyen eszköz, egy olyan projektum[65] részeként valósult meg, melynek célja korpuszok, morfológiai elemzőprogramok és egyéb elektronikus nyelvi erőforrások létrehozása volt néhány kisebb, az uráli nyelvcsaládba tartozó nyelven. A projektum keretében a tundrai nyenyec mellett a nganaszan, a komi, az udmurt, a mari és a manysi nyelvekre készült morfológiai elemzőprogram (Novák, 2004 [2]; Prószéky és Novák, 2005 [3]).

2 A tundrai nyenyec nyelv

A projektum keretében leírt másik északi szamojéd nyelvet, a nganaszant, közvetlen kihalás fenyegeti, beszélőinek száma már csak mintegy 500 fő. A tundrai nyenyecnek ezzel szemben mintegy 25 000 beszélője van. Ugyanakkor ez a beszélőközösség hatalmas területen oszlik el. A tundrai nyenyecek hagyományos lakóterületét nyugaton a Kanyin-félsziget, keleten a Jenyiszej deltája határolja. A XIX. és a XX. század-

[65] Komplex Uráli nyelvészeti adatbázis, NKFP 5/135/2001. A projektumban a Nyelvtudományi Intézet Finnugor Osztálya, különböző finnugor nyelvészeti tanszékek és a MorphoLogic Kft. vett részt.

ban Novaja Zemljára, ill. nyugaton még a Kola-félszigetre is betelepültek, keleten pedig a Tajmir-félszigeten is megjelentek. Míg keleten a tundrai nyenyec nyelv jelenleg is expanzióban van (más szamojéd nyelvek rovására), keleten jellemző a nyelvvesztés a tundrai nyenyecek körében. A nagy földrajzi távolságok ellenére a tundrai nyenyec dialektálisan nem nagyon tagolt: a beszélők mind viszonylag könnyen megértik egymást. A nyelv három nagy dialektuscsoportra (nyugati, középső és keleti) oszlik. A nyugati dialektusok jobban különböznek a középső és keleti változatoktól, mint ezek egymástól.

3 A források

Az elemző készítésénél a *Tapani Salminen* által használt latin betűs fonologikus átírást használtuk, és az ő leírására támaszkodtunk a morfofonológiai szabályok megfogalmazásánál is (Salminen, 1997 [4] és 1999 [6]). Salminen elsősorban a középső dialektuscsoport nyelvjárását írja le (bár külön kitér a nyugati változat jellegzetességeire is), így a mi elemzőnk is erre a változatra készült. Rendelkezésünkre bocsátotta morfológiai szótára anyagát (Salminen, 1998=MDTN [5]) és elküldte disszertációja szövegét is (Salminen, 1997=TNI [4]) számítógéppel kereshető formában, ami felbecsülhetetlen segítséget jelentett az elemző elkészítésében, mert a számítógéppel gyorsan meg lehetett találni a szövegben egyébként meglehetősen szétszórva előforduló információkat.

A tundrai nyenyec – és általában az északi szamojéd nyelvek – esetében a nyelvészeti leírás szempontjából talán a legnagyobb problémát a rendkívül bonyolult és produktív felszíni fonológiai–fonetikai folyamatok jelentik. Ezek nemcsak a formális számítógépes modell megalkotása és implementálása szempontjából, hanem már a nyelvi adatok puszta lejegyzése és bármilyen elfogadható grammatikai modell megalkotása szempontjából is nehézséget jelentenek.

Ami ezekben a nyelvekben egy morféma összes megjelenési formájában közös (a morféma „mögöttes reprezentációja"), az gyakran valami annyira absztrakt dolog, hogy első ránézésre szinte semmi köze nincs a morféma konkrét megjelenési formáihoz, az allomorfjaihoz. Ennek persze az az oka, hogy maguk az allomorfok sem hasonlítanak első ránézésre egymáshoz. Ezért ezeknek a nyelveknek a fonológiájáról és morfológiájáról rendkívül nehéz volt adekvát leírást készíteni, és csak a legutóbbi időben születtek meg ezek a modellek. Salminen leírására azért esett a választásunk, mert ezt a bonyolult absztrakciós folyamatot következetesen végigvitte, és konzisztensnek tűnő, ugyanakkor elég jól formalizált leírást alkotott a tundrai nyenyec morfológiáról. Salminen leírásának magas színvonala és formalizáltságának foka ritka az uráli nyelvészet körébe tartozó nyelvleírások között.

3.1 Salminen jelölésmódja

Két fonéma mind a nganaszanban, mind a tundrai nyenyecben rendkívül problematikus: a schwa (Salminen terminológiájában a tundrai nyenyec esetében „redukált magánhangzó") és a gégezárhang. Az előbbi fonetikailag rendkívül instabil, és a felszínen gyakran a magánhangzókészlet más elemeihez hasonul (illetve a tundrai nyenyecben ezen felül meglehetősen bonyolult szabályoknak engedelmeskedve tűnik el

vagy jelenik meg), és általában csak az egész nyelv fonológiai és morfológiai rendszerének figyelembevételével lehet egy-egy konkrét esetben a kilétét megállapítani. A tundrai nyenyecben ráadásul az *a* fonéma is igen változékony fonetikailag (a hossza és a minősége szempontjából) a hangsúlyviszonyok függvényében. A nyenyecben emellett legalább két különböző gégezárhang van, amelyek a fonetikai realizáció szempontjából nem különböznek egymástól, de különböző fonológiai környezetekben különbözőképpen viselkednek. A „nazalizálható gégezárhang" csak szünet előtt jelenik meg a felszínen gégezárként, egyébként obstruensek előtt homorgán nazálisként realizálódik, szonoránsok előtt pedig eltűnik. A „nem nazalizálható gégezárhang" ezzel szemben az obstruensek előtt tűnik el. Van a gégezárhangnak egy harmadik típusa is: ez nem egy fonológiailag jelen lévő szegmentum felszíni megvalósulása, hanem egy automatikus felszíni fonetikai folyamat eredményeként jelenik meg: a szünet előtti szóvégi mássalhangzó fonémák után toldódik be (hogy ez akusztikailag hogyan realizálódik, arra vonatkozólag nem állnak rendelkezésünkre adatok). A gégezárhanggal kapcsolatban még egy érdekesség megjegyzendő: intervokálisan csupán néhány szóban jelenik meg.

Salminen arra törekedett, hogy a nyelv toldalékolási paradigmáinak figyelembevételével olyan jelölésmódot hozzon létre, amely konzisztens, összhangban van a szavak fonemikus felépítésével, és tükrözi a paradigmatikus viselkedésüket is. Például a fonológiailag schwára (Salminen terminológiájában „redukált magánhangzóra") végződő szavak végén a magánhangzó fonetikailag általában nem realizálódik, de ettől a fonetikai ténytől eltekintve ezek a szavak fonológiailag és morfológiailag (a ragozási paradigmájuk szempontjából) teljes mértékben magánhangzó végű szavakként viselkednek, ezért Salminen ilyenként is ábrázolja őket. Ráadásul ezekben az esetekben az automatikus szóvégi gégezárhang-betoldás is mindig elmarad.

A cirill betűs tundrai nyenyec helyesírás a Salminen által használt fonologikus jelölésrendszerrel szemben erősen fonetikus, a hangsúlytalan a-k és schwák hasonulását az előző magánhangzókhoz általában jelöli, csakúgy, mint a betoldott fonetikai schwákat és gégezárhangokat, ugyanakkor mindössze öt magánhangzót különböztet meg. Mindezt Salminen szerint meglehetősen inkonzisztens módon. Ugyanakkor a helyesírás egyáltalán nem jelöli a mennyiségi (hosszúságbeli) különbségeket, azokat sem, amelyek tényleges fonológiai különbségeken alapulnak. Például a fonológiailag ténylegesen gégezárhangra végződő szavak esetében a gégezárhangot megelőző mássalhangzó akusztikailag jelentősen hosszabb, mint a szóvégi mássalhangzók után automatikusan betoldott gégezárhangok előtt, de ez az írásban nem tükröződik.

Mivel Salminen meglehetősen absztrakt jelölésmódjában a szavak „felszíni alakja" is csak nagyon távoli, áttételes és távolról sem egyértelmű viszonyban van ugyanezen szavak ortográfiai alakjával, ezért egyrészt az oroszországi nyelvészek kétkedéssel tekintenek Salminen jelölésmódjára, másrészt a mi elemzőnket sem lehet egyelőre közvetlenül írott nyenyec szövegek elemzésére használni. Annak természetesen nincs elvi akadálya, hogy az elemző szabályrendszerét kiegészítsük azokkal a szabályokkal, amelyek a Salminen-féle felszíni reprezentáció és a szavak ortográfiai alakjai közötti űrt áthidalják, de ez nem egészen triviális feladat, és egyelőre nem képezi részét leírásunknak.

4. A morfológiai elemző

Az elemzőt a Xerox cég reguláris relációkalkuluson alapuló morfológiai fejlesztőrendszerének, az *xfst*-nek (Xerox Finite-State Tool) felhasználásával készítettük el (Beesley–Karttunen, 2003 [1]). Ez a generatív fonológusok által megszokott kontextusfüggő újraírószabály-formalizmussal leírt szekvenciális fonológiai szabályegyüttesek megadását teszi lehetővé, és kiszámítja az egyes szabályok egymással, illetve a lexikonnal való komponálásával előálló teljes morfofonológiai leírást egyetlen kétszintű véges állapotú fordítóautomata formájában.

Miután a nganaszan elemző és szólak-generátor elkészítéséhez is ezt az eszközt használtuk (Novák, 2004 [2]), logikus döntésnek tűnt, hogy a nganaszannal közeli rokonságban álló, és hasonló bonyolultságú tundrai nyenyec esetében is ehhez a megoldáshoz folyamodjunk.

A morfológiai elemző tőadatbázisának alapjául az MDTN szótár [5] szolgált, a toldaléklexikon, a tő- és a toldaléktárat összekapcsoló általánosabb paradigmatípus-osztályozás és a fonológiai szabálykomponens pedig Salminen disszertációja (TNI=Salminen, 1997 [4]) és a *Grammatical sketch* (Salminen, 1999 [6]) alapján készült. A következőkben részletesen bemutatjuk az elemző moduljait: a tőtárat, a toldaléktárat és a szabályfájlt.

4.1 A tőtár

Az MDTN szótárt Microsoft Excel formátumban kaptuk meg. Az szótárt ISO-8859-1 kódolású szövegfájllá konvertáltuk. A konvertált szótár egy részlete alább látható:

```
   nga°                                Part          NGA ' Ø
@  (E -ibø- ~) W-C søb°bø-             Vt ¢ ø»yi     SØPØ 0«MPØ
@  (-mpø- ~ E -ibø- ~) W-C tyeb°bø-    Vt ¢ ø»yi     TYEPØ 0«MPØ
@  (E -ibø- ~) W-C tyíb°bø-            Vt ¢ ø»yi     TYÍPØ 0«MPØ
   (E -ibø- ~) W-C løbc°bø-            Vt ¢ ø»yi     LØPSØ 0«MPØ
@  (E -ibø- ~) W-C yabc°bø-            Vt ¢ ø»yi     JAPSØ ' 0«MPØ
@  (-mpø- ~ E -ibø- ~) W-C nyanc°bø-   Vt ¢ ø»yi     NYAHSØ 0«MPØ
@  (-mpø- ~) nyenc°bø-                 Vi ¢ ø»yi     NYEHSØ 0«MPØ
@  (E -ibø- ~) W-C syenc°bø-           Vt ¢ ø»yi     SYEHSØ '' 0«MPØ
   sødøb°                              N ø=          SØTØPØ
```

A tőtár létrehozásához a mögöttes alak mezőt (utolsó oszlop), a kategóriamezőt (Part, Vt ¢, N stb.) és a toldalékolási osztály (ø»yi, ø= stb.) mezőket használtuk. A felszíni alak mezőre nem volt szükségünk.

A mögöttes alak morfémákra van szegmentálva. A '-ok a homonim tövek azonosítására szolgának. A képzők illesztésénél fellépő sandhijelenségekre utaló jeleket (pl. 0«) a mögöttes alak tartalmazza a produktív hasonulási jelenségek kivételével. (A produktív hasonulásokat a külön leírt fonológiai és morfofonológiai szabálykomponens kezeli.)

A fenti adatbázisnak a Xerox rendszer által a lexikon leírására használt *lexc* formátumára való átalakításához készítettünk néhány programot. Az átalakító a fenti szótárrészletet először az alábbi formára konvertálja:

```
Root      nga^1|ø                          Part_
Root      søpø|^0«mpø                      Vt=c_ø»yi
Root      tyepø|^0«mpø                     Vt=c_ø»yi
Root      tyípø|^0«mpø                     Vt=c_ø»yi
Root      løpsø|^0«mpø                     Vt=c_ø»yi
Root      japsø^1|^0«mpø                   Vt=c_ø»yi
Root      nyahsø|^0«mpø                    Vt=c_ø»yi
Root      nyehsø|^0«mpø                    Vi=c_ø»yi
Root      syehsø^2|^0«mpø                  Vt=c_ø»yi
Root      søtøpø                           N_ø=
```

Az első oszlop azt adja meg, hogy az adott morfémasorozat melyik allexikonba kerül. Ezt követi a morfémasorozat mögöttes alakja (| jelekkel morfémákra szegmentálva). Végül az adott morfémasorozat folytatási osztálya áll (ez egy lexikonnév: minden olyan morfémasorozat, ami az adott nevű lexikonban szerepel, követheti az adott morfémasorozatot). A Root lexikonban szereplő elemek állhatnak a szó elején. A szó végét a # folytatási osztály jelzi. A következő lépésben az egyes allexikonokba tartozó elemeket egybegyűjtöttük a *lexc* formátumnak megfelelően. Az így kapott adatbázis lett az elemzőprogram tőtára.

4.2 A ragozási osztályok leírása

Salminen ragozási osztályait az egyes kategóriákon belül gyakorisági sorrendbe rendeztük és a TNI-ben felállított általánosabb toldalékolási osztályokba soroltuk. Az egyes osztályok jellegzetességeit az MDTN bevezetője elég részletesen tárgyalja. A besorolás alapja alapvetően ez a leírás volt. Ugyanakkor az MDTN bevezetője számos olyan tényt is közöl az egyes osztályoknál, amelyek jóval általánosabb fonológiai, ill. morfofonológiai folyamatok következményei. Ezeknek természetesen nem itt van a helyük, hanem a fonológiai, ill. morfofonológiai szabályok leírásánál. Itt ki kellett hagynunk ezeket a redundáns információkat. Az alábbi lista néhány tárgyasvisszaható és tárgyas igeosztály így annotált leírását mutatja, az osztály neve után annak lexikonbeli gyakorisága és szoknak az általánosabb toldalékolási osztályoknak a felsorolása következik, amelyekbe az adott toldalékolási osztály tartozik (részlet a teljes toldalékolásiosztály-listából):

```
Vt-r µ a: 162  POLYVSTEMV
Vt-r µ ye: 43  POLYVSTEMV
Vt-r µ l: 28    CSTEMV
Vt-r µ r: 15 CSTEMV
Vt-r µ ø: 13   POLYVSTEMV  ØSTEMV
...
Vt ¢ ø»yi: 2361   ALTV ALTVøi
Vt µ a: 1380   POLYVSTEMV
Vt µ ye: 579  POLYVSTEMV
```

Írtunk egy programot, amely ezt a leírást folytatási osztály alapú lexikonná alakítja. Ez a lexikonrészlet kapcsolja össze a tőlexikon folytatási osztályait az alább ismertetendő toldaléklexikon toldalékosztályaival. Az alábbi példa a leggyakoribb tárgyasvisszaható igeosztály és a leggyakoribb tárgyas igeosztály ilyen formájú leírását mutatja. A középső oszlopban []-ben szereplő címkéket az adott osztályba tartozó szavak elemzésekor az elemző kiírja. A @ jelek közötti jegyeket a program a tőhöz kapcsol-

ható toldalékosztályok szűrésére használja. Az alábbi példákban szereplő `@P.CONJ.tr@` szimbólum például a `CONJ` (igeragozás) jegy `t-r` (tárgyas-reflexív) értékre való beállítását írja elő. Ez lehetővé teszi valamennyi igei személyragosztálynak az ilyen osztályú tövekhez való kapcsolódását. Ugyanakkor pl. a `@P.CONJ.t@` (tárgyas) osztályú tövekhez a visszaható toldalékok nem kapcsolódhatnak.

```
#Class Vt-r=m_a - frq=162
Vt-r=m_a            [V][t-r][Mom]@P.CONJ.t-r@  Vt-r=m_a_
Vt-r=m_a_           POLYVSTEMV
Vt-r=m_a_           V_base
...

#Class Vt=c_ø»yi - frq=2361
Vt=c_ø»yi           [V][t][Cnt]@P.CONJ.t@      Vt=c_ø»yi_
Vt=c_ø»yi_          ALTV
Vt=c_ø»yi_          V_base
Vt=c_ø»yi_          »^yi          GFS=SFS
```

4.3 A toldaléktár

Toldaléklexikonunk és a szabályok leírásának alapjául a TNI és a *Grammatical sketch* szolgált. Ezeket a műveket többször elolvastuk, kijegyzeteltük, és több plakát méretű színes gráfot rajzoltunk, amelyek az egyes tőalakváltozatok (pl. az igéknél 'Special Finite Stem'[66] , 'General Finite Stem'[67], 'General and Special Modal Substems'[68]) és az egyes toldalékok alakváltozatait (itt nem figyelembe véve az általános fonológiai és morfofonológiai folyamatok következtében előálló alternációkat) és egymáshoz viszonyított sorrendjét ábrázolták. Mivel a nyenyec toldalékolási rendszer igen bonyolult, hasonlóan a nganaszanéhoz, ezek a gráfok tekintélyes méretűek lettek.

Mivel a forrásművek a képzőket nem tárgyalták, elemzőnk csak a ragozást kezeli produktívan. (Salminen a ragozás körébe utalta az igenevek és gerundiumok képzését, amelyeket általában képzett alakoknak tekintik, ezeket tehát elemzőnk kezeli). Ugyanakkor az MDTN szótár rengeteg képzett alakot tartalmaz (méghozzá morfémákra szegmentálva), ez tehát majd (egy következő projektum keretében) jó alapja lehet a produktív szóképzési folyamatok formális leírásának és számítógépes modellezésének.

A toldalékolási gráfokat ábrázoló plakátok alapján készítettük el a toldaléklexikont a *lexc* formátumához közeli fentebb ismertetett folytatási osztályokon alapuló formában. Ez lexikonunk harmadik része, melyet az előbb ismertetett két résszel (tőtár, toldalékolási osztályok tára) összefűzve a teljes lexikont megkaptuk. A toldaléklerás tartalmaz néhány javítást a forrásokban leírtakhoz képest, amelyeket az elemző írása, illetve tesztelése során talált hibák alapján Salminennel konzultálva végeztünk.

A toldaléktárban azoknak a jelenségeknek a kezelésére, amelyek nem szomszédos morfémák közötti megszorításokon alapulnak, jegyérték-ellenőrző kifejezéseket használtunk. Ilyenek gondoskodnak például arról, hogy a megfelelő igeiszemélyragosztályok éppen a megfelelő vonzatkeret-osztályba tartozó tövekhez járulhassanak

[66] SFS = a visszaható és a többes számú tárgyas kijelentő módú finit igealakok töve

[67] GFS = a többi finit igealak töve

[68] ugyanez a különböző egyéb igemódokra

csak (az alanyi ragozás személyragjai a reflexív igéket kivéve mindhez, a tárgyas személyragok a tárgyas és a tárgyas-reflexív igékhez, a visszaható személyragok pedig a reflexív és a tárgyas-reflexív igékhez). Hasonlóan kezeltük az enklitikus múltidő-jel megjelenésére vonatkozó megszorításokat, az opcionális palatalizáció és az *e~i°* alternáció jelenségét.

4.4 A szabályfájl

Salminen leírását a nyenyec morfofonológiáról viszonylag egyszerűen le lehetett fordítani az *xfst* szabályformalizmusára. Ugyanakkor a kézzel írott nyelvtanokban általában sok részlet homályban marad. Itt is ez volt a helyzet. Mindazokat a pontokat, ahol az eredeti leírás homályos volt (pl. a szabályok sorrendezése, alkalmazási köre, a környezet pontos leírása, egyáltalán formális megfogalmazása, a kivételek stb.) explicitté kellett tennünk. Az általunk program formájában implementált leírás a tundrai nyenyec fonológiáról és morfofonológiáról tehát az eredeti forrásokban közölt leírások jóval részletesebb és teljes formális igényű javított változatának tekinthető.

A szabályfájl definíciókkal (főleg szegmentumosztályok definícióival) kezdődik. Ezt követik a fonológiai, majd a morfofonológiai folyamatokat leíró szabályok. A fonológiai szabályok közül az igen komplex magánhangzó-redukciós folyamatot leíró szabály az első, ezt követik a különböző produktív mássalhangzó-hasonulási szabályok. A morfofonológiai szabályok az ingadozó szegmentumok viselkedését leíró szabályokkal kezdődnek, ezeket követik a különböző korlátozott morfológiai környezetekben működő (általában csak morfémahatáron, ill. csak egy-egy morféma környezetében működő) sandhiszabályok.

A szabályok sorrendezése a szabályok definíciójától függetlenül van megadva, hiszen ezt a fejlesztés–tesztelés során többször meg kellett változtatnunk.

4.5 A kész elemző

Ugyanarra a nyelvtanra alapozva több változatot is elkészítettünk az elemzőből, amelyek az elemzések „szószátyársága” szempontjából különböztek egymástól: a kevésbé bőbeszédű változat csak az abszolút tő szótári alakját és a tő és a toldalékok morfoszintaktikai jegyeit írja ki, a gazdagabb változat kiírja a toldalékok mögöttes alakját is. Tulajdonképpen az utóbbi az alapvető változat, az előbbit ebből egy megfelelő szűrő hozzákomponálásával állítjuk elő. A tömörebb elemzéseket adó változat inverzét használjuk szóalak-generátorként. Az interaktív tesztelés céljára készítettünk olyan változatokat is, amelyek a Salminen által használt karaktereknél a billentyűzeten könnyebben gépelhető (ASCII) karakterekkel dolgoznak, de egyébként ekvivalensek a Salminen jelölésrendszerét használó változattal.

Néhány példa az egyes változatok által adott elemzésekre:

ASCII tömör elemző

```
myeryuj'w@nantyih       myeryo|>^yuj'[N][Poss][Pros][Sg][3][Du]
myeryuj'w@nantyih       myeryo|>^yuj'[N][Poss][Pros][Sg][2][Du]
```

ASCII bőbeszédű elemző

```
myeryuj'w@nantyih
myeryo|>^yuj'[N][Poss]m'na[Pros][Sg]ht[3]yih[Du]
myeryuj'w@nantyih
myeryo|>^yuj'[N][Poss]m'na[Pros][Sg]ht[2]yih[Du]
```

ASCII generátor

```
myeryo|>^yuj'[N][Poss][Pros][Sg][3][Du] myeryuj'w@nantyih
```

Salminen tömör elemző

```
myeryuj°wønantyih       myeryo|»^yuj°[N][Poss][Pros][Sg][3][Du]
myeryuj°wønantyih       myeryo|»^yuj°[N][Poss][Pros][Sg][2][Du]
```

Salminen bőbeszédű elemző

```
myeryuj°wønantyih
myeryo|»^yuj°[N][Poss]m°na[Pros][Sg]ht[3]yih[Du]
myeryuj°wønantyih
myeryo|»^yuj°[N][Poss]m°na[Pros][Sg]ht[2]yih[Du]
```

5 Összefoglalás

Az elemző lexikona mintegy 19 500 tő mögöttes fonológiai reprezentációját tartalmazza. Ezeket Salminen 266 különböző ragozási osztályba sorolta. A toldaléktár 254 mögöttes toldalékalakot tartalmaz. Az elemző a nyelv inflexiós jelenségeit teljes körűen kezeli, beleértve az általában a szóképzés körében tárgyalt igenevek és gerundiumok kezelését is. A tőtárban szereplő rengeteg képzett szó morfémákra tagolását is megadja a program, annak ellenére, hogy a szóképzést produktívan nem kezeli, hiszen ez az információ az eredeti forrásban szerepelt, és a tőtár létrehozása során megtartottuk.

A létrehozott eszközök megfelelő nyelvi adatok megléte esetén a bennük implementált nyelvtan adekvátságának messzemenő tesztelését teszik lehetővé, olyan alapossággal, amely – különösen egy a tundrai nyenyechez hasonló bonyolultságú nyelv esetében – kézzel elképzelhetetlen. Az elemzőt elsőként az MDTN szótár előszavában megjelent teljes példaparadigmákon teszteltük. Az első változatban természetesen találtunk jó pár hibát, a lexikonban, a szabályok megfogalmazásában és ezek sorrendezésében is. Ezeket kijavítottuk. Másik tesztanyagként a TNI-ben szereplő példaszavakat tudtuk használni, hiszen ott is szerepel a szavak kézi elemzése is.

Sajnos nagyon kevés olyan korpusz áll rendelkezésre, amely Salminen jelölésmódjával van lejegyezve. Ezért is nagyon kívánatos lenne, hogy az elemzőt kiegészítsük azokkal a leképezési szabályokkal, amelyek lehetővé teszik, hogy közvetlenül a nyenyec ortográfiával lejegyzett szövegek elemzésére használhassuk. A másik lehetséges továbbfejlesztési lehetőség a képzők produktív kezelése, feltéve, hogy a szükséges ismeretek rendelkezésre állnak majd.

Salminennel egyébként a fejlesztés során végig szoros kapcsolatban voltunk és minden pontatlanságra és hiányra felhívtuk a figyelmét a forrásokban, amire az elem-

ző fejlesztése során fény derült. Reményeink szerint így együttműködésünk az ő számára is hasznos volt.

Bibliográfia

1. Beesley, Kenneth R. and Lauri Karttunen: *Finite State Morphology*, CSLI Publications, Ventura Hall (2003)
2. Novák Attila: Az első nganaszan szóalaktani elemző. In: *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004).* Szegedi Tudományegyetem (2004) 195–202
3. Prószéky, Gábor and Attila Novák: Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.): *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford (2005) 116–125
4. Salminen, Tapani: *Tundra Nenets inflection.* Mémoires de la Société Finno-Ougrienne 227; Helsinki (1997) = TNI
5. Salminen, Tapani: *A morphological dictionary of Tundra Nenets.* Lexica Societatis Fenno-Ugricae 26; Helsinki (1998) = MDTN
6. Salminen, Tapani: *Tundra Nenets (A grammatical sketch)*, http://www.helsinki.fi/~tasalmin/sketch.html (1999)

A magyar nyelv sajátságaihoz illeszkedő módszerek szövegek automatikus osztályozására

Németh András[1,2], Balázs László[1]

[1] Alkalmazott Logikai Laboratórium,
Hankóczy J. u. 7. 1022 Budapest,
{xandrew, bazsi}@all.hu
2 Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék,
Magyar tudósok körútja 2. 1117 Budapest,
xandrew@cs.bme.hu

Kivonat: A magyar nyelv gazdag morfológiája és agglutináló jellege megkérdőjelezi az angol nyelvre jól működő szövegklasszifikációs technikák változatlan alkalmazását. A legtöbb bevett módszerben szavak előfordulását vizsgáljuk a dokumentumokban, azonban a magyar nyelv esetében a szóalakok nagy száma miatt ez nem tűnik alkalmas megközelítésnek. Jelen cikkben két módszert javaslunk a probléma kezelésére: a már korábban is alkalmazott szótövesítést, illetve n-grammok alapján történő osztályozást. Vizsgálataink azt mutatják, hogy a kisebb apparátust igénylő n-gramm alapú technikák is a szótövesítéshez hasonlóan jó eredményt adnak, és még robosztusabbnak is bizonyulnak annál.

1 Bevezetés

Az automatikus szövegosztályozási feladatban dokumentumokat kell előre meghatározott kategóriákba sorolnunk, adott és már kategóriákba sorolt mintadokumentumok alapján.

A feladat hosszú idő óta aktív kutatási terület, és a lehetséges megoldások iránt folyamatosan nő az érdeklődés az Interneten elérhető dokumentumok számának rohamos emelkedésével. Egy jól működő osztályozó problémák rendkívül széles körében használható, például keresési eredmények strukturált megjelenítésére, beérkező levelek automatikus szortírozására, vállalati intraneten rendezetlenül megtalálható dokumentumok elérhetővé tételére témák szerint tagoltan, IR (information retrieval) rendszerek fontos részeként és még hosszan sorolhatnánk.

Az igényeknek megfelelően a nemzetközi irodalom is hatalmas, módszerek széles skáláját javasolták a különböző sztohasztikus modellektől a döntési fákon és a legközelebbi szomszéd algoritmuson keresztül a neurális hálózatokig. Az egyik legsikeresebb megközelítésnek az SVM (support vector machine, lásd pl. [6]) osztályozó használata bizonyult.

A magyar szövegek osztályozásának specifikus feladata lényegesen kisebb figyelmet kapott. Kornai és társai az origo.hu portál kulcsszó kereső és téma osztályozó

rendszerének kiépítése kapcsán foglalkoztak a feladattal, és bemutattak egy Bayes-modell alapú osztályozót és szótövesítést használó megoldást [5].

Cavnar és Trenkle már nagyon korán javasolták nyelvfelismerésre és szövegklasszifikációra az n-grammok frekvencia profile-jainak összehasonlítását [1]. Később Langdon ajánlott 3-gramm előfordulási vektorokra alkalmazott legközelebbi szomszéd osztályozót nyelvfüggetlen dokumentumklasszifikációs technikaként [7].

Jelen cikkben áttekintjük az egyik legsikeresebb módszercsaládot, és megvizsgáljuk, hogy miért kevéssé alkalmasak a klasszikus technikák magyar nyelvű szövegek esetében (2. fejezet). Megadjuk a leírt algoritmus egy egyszerű általánosítását, és ezen általánosítás speciális eseteként ajánlunk két módszert a magyar nyelvhez illesztéshez: a szótövek ill. az n-grammok alapján történő osztályozást (3. fejezet). Ismertetjük a módszerek értékelésére elvégzett kísérleteinket (4. fejezet). Végül összefoglaljuk a kísérletekből levont következtetéseinket (5. fejezet).

2 Szó alapú szövegklasszifikációs módszerek áttekintése

Ebben a fejezetben áttekintünk egy bevett, jól működő osztályozási technikát, és megvizsgáljuk a magyar nyelvre való alkalmazhatóságát.

2.1 A feladat

Adott kategória címkékkel ellátott dokumentumok egy halmaza. A feladat ezen tanítókészlet alapján valamiféle modell betanulása, és ennek segítségével címkézetlen dokumentumok osztályozása. A szövegklasszifikációs feladat speciális esete a topic felismerés. Itt egy bizonyos kategóriába tartozó (pl. adott témáról szóló, innen az elnevezés) dokumentumokat kell elkülöníteni az összes többitől. A cikkben elvégzett kísérletek erre a speciális esetre vonatkoznak.

2.2 Dokumentum reprezentáció

Az elterjedt klasszifikációs módszerek közös eleme, hogy a dokumentumokról csak azt tartják nyilván, hogy mely szavak fordulnak bennük elő. Itt a „szó” jelentése előre definiált elválasztó karakterek (pl. whitespace, központozás) közötti összefüggő karaktersorozat. A sorrend információ minden esetben eldobódik, de általában még a pontos előfordulási számmal sem foglalkozunk, így egy dokumentumot a benne előforduló szavak halmazaként reprezentálhatunk (ez az úgynevezett bag of words model).

Bár ezzel a reprezentációval nyilvánvalóan rengeteg információt elveszítünk, mégis úgy tűnik – a jó minőségű elérhető klasszifikáció miatt –, hogy az osztályozási feladat szempontjából ez még megengedhető. Ráadásul az ezen modell alapján történő osztályozásra jól működő egyszerű módszereink vannak, míg a sorrendiségben rejlő plusz segítség felhasználása nehezen képzelhető el nagyon komoly apparátus, lényegében természetes nyelvű szövegértés nélkül.

2.3 Releváns szavak kiválogatása

Ezen technikák feladata, hogy az osztályozó lépés előtt csökkentsük a feladat dimenzióját a lehető legkevesebb információ elvesztése mellett. Lényegében az összes módszer azon alapul, hogy az egyes szavak kapnak egy pontszámot, ami jellemzi a relevanciájukat az adott osztályozási feladat szempontjából, és ezen függvény szerinti legfontosabb néhány szót tartjuk meg. A lehetséges függvények széles köre és értékelésük megtalálható [3]-ban. Mi az alábbi két, széles körben alkalmazott függvényt használtuk.

Khi négyzet (χ^2 , chi-square) függvény. Ez a statisztikában különféle hipotézisvizsgálatoknál használt próbán alapul. Lényegében függetlenség vizsgálatot végzünk: független-e a szó előfordulása a dokumentumok kategóriáitól. Egy kétdimenziós diszkrét mintasorozatból a két dimenzió függetlenségének megállapítására az alábbi statisztikát használhatjuk:

$$T = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{N_{i.}N_{.j}} \qquad (\mathbf{10})$$

Itt N_{ij} azon mintaelemek száma, ahol az első dimenzióban az *i*. a másodikban pedig a j. lehetséges érték vevődik fel. $N_{i.}$ ill. $N_{.j}$ a sor ill. oszlopösszegeket jelöli, *n* pedig a minták száma. Az így kapott T statisztika a nullhipotézis teljesülése esetén eloszlásban az $(r-1)(s-1)$ rendű χ^2 eloszláshoz tart, így minél nagyobb a *T*, annál inkább el kell vetni a függetlenséget a próba során, a konkrét kritikus érték a próba terjedelme alapján számítható. Ez alapján a fenti *T* jó eszköz a két dimenzió összefüggőségének mérésére. A mi esetünkben a két dimenziós mintasorozat első dimenziója a dokumentumok kategóriája, második dimenziója pedig 0 vagy 1 attól függően, hogy a vizsgált szó szerepel-e a megfelelő dokumentumban. Tehát az (1) függvény szerinti minél nagyobb értékű szavakat kell választani.

Az információs nyereség (information gain) függvény. Ennek meghatározásához először kiszámítjuk a kategóriák eloszlásának entrópiáját a teljes dokumentum halmazon. Ezután kiszámítjuk az entrópiát a vizsgált szót tartalmazó és nem tartalmazó dokumentumokra külön-külön is. Ennek a két entrópiának a halmazok (a szót tartalmazó dokumentumok halmaza ill. a szót nem tartalmazó dokumentumok halmaza) elemszámával súlyozott átlaga kisebb egyenlő az eredeti entrópiánál. Ha kisebb, az azt jelenti, hogy azzal, hogy tudjuk, hogy a szó szerepel-e egy dokumentumban információt nyerünk annak kategóriájáról. A két fenti érték különbsége az információs nyereség, itt is a minél magasabb információs nyereségű szavakat érdemes kiválasztani.

2.4 Osztályozás

Az irodalom feltűnően egységes abban a kérdésben, hogy milyen osztályozót érdemes használni dokumentum klasszifikációhoz. A lineáris SVM mutatja a legjobb

eredményeket. Ezért itt bővebben nem foglalkozunk az osztályozó választás kérdésével, végig lineáris SVM-et használunk.

2.5 A teljes modellépítési és értékelési folyamat áttekintése

A fenti módszer egy variánsának teszteléséhez a következő lépéssorozatot kell végrehajtanunk:

1. Korpusz választás ill. előállítás: Kiválasztjuk az osztályozandó kategóriákat, és előállítunk egy (dokumentum, kategória) párokból álló halmazt, amiben egy dokumentum csak egyszer szerepelhet, és minden dokumentum a vizsgált kategorizálás szerint helyesnek tartott kategóriájával áll párban.
2. A korpuszt két diszjunkt részre, tanító és tesztkészletre bontjuk szét.
3. Kiválasztunk egy relevancia függvényt, és a tanító készlet alapján minden szónak kiszámítjuk a relevanciáját, és kiválasztjuk az első k legrelevánsabb szót.
4. A kiválogatott szavak segítségével minden dokumentumhoz egy bitsorozatot rendelünk: minden bit egyes értéke azt jelenti, hogy a hozzá tartozó szó előfordul a szövegben. Az így kapott vektorok klasszifikációjára lineáris SVM-mel építünk osztályozó modellt.
5. A teljes osztályozó modell a 3. lépésben nyert szó listából és a 4. lépésben nyert SVM modellből áll.
6. Meghatározzuk a tesztkészlet elemeire a modell által adott kategóriákat és ezt összevetjük a helyes kategorizálással, és valamilyen módon kvantitatívan értékeljük a teljesítményt

2.6 Alkalmazás a magyar nyelvre

A fenti módszercsalád sikere angol nyelvű szövegek osztályozására bizonyítja az alapvető redukciós ötlet helyességét: csak bizonyos kulcsfogalmak előfordulását vizsgálva a szövegekben - minden mást, a további szavakat és a sorrendiségi információkat eldobva - jó minőségű osztályozót készíthetünk.

Magyar nyelvű szövegek esetén azonban a fenti technikák változatlan formában történő használata lényegesen rosszabb eredményre vezet. Ez a jelenség könnyen megérthető. Ugyanis a klasszifikáció szempontjából egy bizonyos szó különböző toldalékolt alakjai lényegében ugyanazt az információt tartalmazhatják, ám mi teljesen független megjelenésekként kezeljük őket. Ezzel a helyes relevancia megállapítását is lehetetlenné tesszük, hiszen pl. könnyen előfordulhat, hogy egy releváns fogalom különbözőképpen ragozott alakjai közül egyik sem releváns (pl. egyszerűen a ritka előfordulásuk miatt). Másrészt ha egy szó több alakját is kiválasztjuk, akkor a kapott bináris vektoroknak különböző koordinátái azonos jelentésűek lesznek, feleslegesen nehezítve ezzel az SVM dolgát.

3 A magyar nyelvhez illesztés lehetőségei

Ebben a fejezetben általánosítjuk a fent leírt algoritmust, majd megadjuk a kapott általános algoritmus két, a magyar nyelv jellegéhez a klasszikus technikánál jobban illeszkedő speciális esetét.

3.1 Jellemző előfordulási modell

Vegyük észre, hogy a 2.5. szakasz lépései közül sehol sem használtuk ki, hogy amiknek az előfordulását vizsgáljuk a dokumentumokban, azok szavak. Valójában általánosítható a dokumentum ábrázolási modellünk a következő módon. Rögzítünk valahogy jellemzők egy (nem feltétlenül véges) halmazát úgy, hogy bármely szöveghez meghatározható legyen ennek egy véges részhalmaza, ami az előforduló jellemzőknek felel meg, és ezzel a részhalmazzal fogjuk reprezentálni a dokumentumainkat. (A jellemzők alaphalmazának elemei az eddigiekben az alfanumerikus karaktersorozatok voltak, és egy dokumentumhoz a benne elválasztó karakterek között előforduló sorozatok halmaza tartozott.)

A fenti reprezentációval adott tanító dokumentumkészlet segítségével kiválaszthatjuk a releváns jellemzőket, és azokból épített bináris vektorokat osztályozhatunk, azaz tulajdonképpen módosítás nélkül alkalmazhatjuk a fenti technikákat. Ennek az általánosabb reprezentációnak az előnye, hogy a jellemzők ügyes megválasztásával lényegesen javíthatunk az eredményeken.

3.2 Szótövek mint jellemző

A 2.6. részben leírt probléma kézenfekvő megoldása, hogy az előforduló szavakat a további feldolgozás előtt szótövesítjük. Tehát a vizsgált jellemzők szótő előfordulások lesznek. Erre a célra a JMorph morfológiai motort használtuk (lásd [8]).

Fontos tervezési kérdés, hogy pontosan mit is tekintsünk szótőnek, hiszen nem érdemes minden esetben az összes toldaléktól megszabadulni (jó példa erre az egészség szó). Az általunk választott megoldás az, hogy az összes jeltől és ragtól megválunk, és addig töröljük a képzőket, amíg egy már önállóan is a szótárunkban előforduló alakot nem kapunk. Ez egy jó heurisztika a megállási pont megválasztására, de persze nem adja minden esetben az optimális megoldást.

A technika hátránya a viszonylag nagy apparátus igénye. Egy morfológiai elemző önmagában bonyolultabb mint a rendszer többi része együttvéve, nem is beszélve a mögötte álló morfológiai erőforrások elkészítésének munkaigényéről (szótár, szabálykészletek). A problémát tovább súlyosbítja, hogy ha új nyelvre akarunk osztályozót készíteni, az elemző készítését lényegében nulláról kell kezdenünk.

A szótövesítés futási idő szempontjából is költséges művelet, egy komoly keresési feladat. Érdemes itt megjegyezni, hogy különösen a sikertelen elemzések tartanak sokáig, hiszen jó keresési stratégiával a legjobbnak tűnő felbontás (ha létezik) az esetek túlnyomó többségében a teljes keresési tér bejárása nélkül is megtalálható, míg negatív válasz csak az összes eshetőség végigpróbálása után adható. Tehát az ismeretlen ill. elírt szavak elemzése sokáig is tart és persze végül nem is kapunk használható eredményt. (Azon szavakat, amelyekkel az elemző nem tud mit kezdeni kísérleteinkben önálló jellemzőknek tekintjük.)

3.3 N-gramm előfordulások mint jellemzők

Másodikként egy nyelvészeti tudást nem használó, igen egyszerű módszert javaslunk. Legyenek a jellemzők egyszerűen a szövegben előforduló egymás utáni betű n-esek. Variációs lehetőség, hogy a szóhatárokon átnyúló n-grammokat eldobjuk vagy megtartjuk. A „Hideg van.” mondatban előforduló betű hármasok pl. az első esetben {'hid', 'ide', 'deg', 'van'} a másodikban pedig {'hid', 'ide', 'deg', 'eg ', 'g v', ' va', 'van'}. Mi előzetes kísérletek alapján az első változatot használtuk.

Ezzel a technikával láthatóan rengeteg teljesen értelmetlen és véletlenszerű jellemzőt kapunk, ám a válogatási lépés ezeket automatikusan eltávolítja, így az osztályozáshoz már csak a tényleg sokat mondó n-grammok maradnak meg. Nagy előny, hogy a technika teljesen nyelvfüggetlen, csak a tanítási fázist kell újra elvégezni ha más nyelvű szövegeket szeretnénk osztályozni.

Az n-gramm alapú klasszifikáció további előnye, hogy várhatóan robosztusabban viselkedik zajjal szemben (pl. elgépelések, helyesírási hibák). Egy elírt szóra a morfológiai elemző szinte biztosan hibás eredményt ad, míg egy tipikus darabja jóval nagyobb eséllyel érintetlenül megtalálható. Olyankor is alkalmazhatók az n-grammok, amikor a szóhatárok nem ismertek, pl. beszédfelismerő rendszerek kimenetének osztályozása.

4 Elvégzett kísérletek, eredmények

A kísérletek vezérfonala, hogy a klassztterező metodika elemeinek nagy részét rögzítve megvizsgáljuk, hogy a jellemzők halmazának megválasztása milyen hatással van az osztályozás minőségére.

4.1 A teljes kísérlet sorozat során rögzített választások

1. A tanító- és tesztkorpusz szétválasztása minden egyes kísérletben függetlenül és véletlenszerűen történt, 70-30 arányban (a dokumentumok 70% került a tanítókorpuszba).
2. A relevancia megállapítását az információs nyereség függvény segítségével végeztük. A kiválasztandó szavak *k* számát egy logaritmikus skála mentén változtattuk, tehát az egyes technikák minőségét egy függvénnyel jellemeztük, melynek változója a kiválogatott jellemzők száma. A rögzített k értékek: 10, 16, 27, 46, 77, 129, 216, 362, 604, 1010.
3. Az osztályozást lineáris SVM-mel végeztük, a *c* paraméter 1.0 értékével. Ehhez a libsvm (lásd [2]) nevű szabadon használható implementációt használtuk.
4. Az eredmények kvantitatív értékelése az f-measure értékelő függvénnyel történt (lásd pl. [10]).

4.2 Használt korpuszok

4.2.1 Az Index hírportálról letöltött cikkek

A kizárólag magyar nyelvű kísérletekhez az Index [4] nevű hírportál archívumából letöltött 2100 cikket használtuk. A cikkek természetes kategorizálását adja a rovat neve. Az alábbi hét rovatból szerepelnek cikkek a korpuszban, rovatonként 300 cikkel: bulvár, gazdaság, külföld, kult, sport, tech-tudomány.

4.2.2 A Hunglish korpusz

A Hunglish korpusz a Budapesti Műszaki Egyetem Média Oktató és Kutató Központja és a Magyar Tudományos Akadémia Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által készített több mint 50 millió szövegszót és 2 millió mondatot tartalmazó, mondatszinten illesztett, magyar-angol párhuzamos korpusz [9]. A magyar és angol nyelvű szövegek osztályozását összehasonlítandó ebből a korpuszból választottunk ki jogi és irodalmi szövegeket. Minden egyes osztályozandó dokumentumot a korpusz 50 egymást követő mondatból állítottuk össze, ilyen dokumentumból ezret-ezret készítettünk el a tesztjeinkhez. Az egyes dokumentumok kategóriája "law" ill. "lit" volt attól függően,. hogy jogi vagy irodalmi szövegrészletről volt-e szó.

4.3 Szó alapú osztályozás hatékonyságának összehasonlítása angol és magyar nyelvű szövegekre

Az elsőként ismertetendő kísérletünk célja az volt, hogy megvizsgáljuk, hogy a klaszszifikációs feladat szó alapú megoldásának hatékonyságáért mennyiben felelős a dokumentumok nyelve. Erre a célra tökéletesen alkalmas volt a Hunglish korpusz, hiszen segítségével pontosan ugyanazon szövegek magyar és angol nyelvű változatain végezhettünk osztályozást. A feladat a jogi és irodalmi szövegek elválasztása volt, a kiválasztott szavak számának függvényében a kapott eredményeket az alábbi grafikon foglalja össze.

A várakozásoknak megfelelően az angol nyelvű dokumentumok elkülönítése jobb eredményt adott kevés releváns szót kiválogatva. Meglepő, hogy nagy jellemzőszámok esetén a magyar nyelvű klassztterezés egy árnyalattal sikeresebb. A vizsgált feladat meglehetősen könnyű, és nem tipikus abból a szempontból, hogy itt nem csak (sőt nem elsősorban) a dokumentumok témája lehet az elkülönítés alapja, hanem azok stílusa, szóhasználata. Így a meglepő jelenség magyarázata lehet az, hogy a magyarban nagyobb különbség van a jogi és a szépirodalmi stílus között.

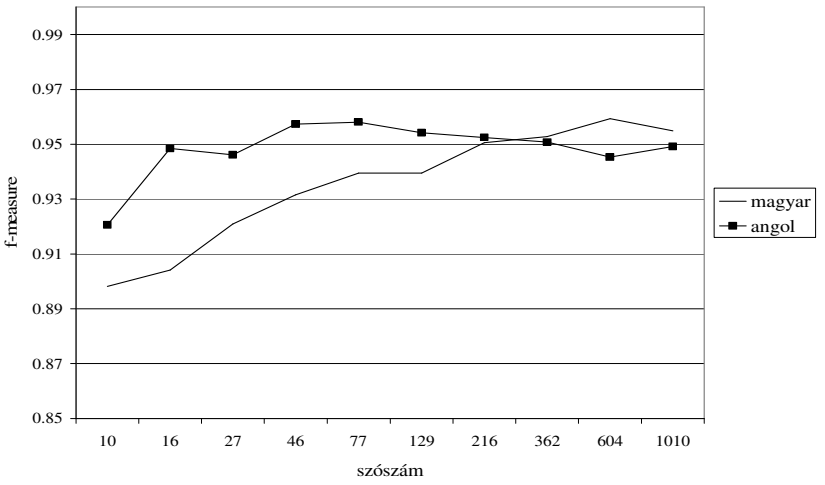


Fig. 1. A Hunglish korpusz jogi dokumentumainak elkülönítése az irodalmiaktól szavak előfordulásai alapján, magyar és angol nyelven

A várakozásoknak megfelelően az angol nyelvű dokumentumok elkülönítése jobb eredményt adott kevés releváns szót kiválogatva. Meglepő, hogy nagy jellemzőszámok esetén a magyar nyelvű klasszterezés egy árnyalattal sikeresebb. A vizsgált feladat meglehetősen könnyű, és nem tipikus abból a szempontból, hogy itt nem csak (sőt nem elsősorban) a dokumentumok témája lehet az elkülönítés alapja, hanem azok stílusa, szóhasználata. Így a meglepő jelenség magyarázata lehet az, hogy a magyarban nagyobb különbség van a jogi és a szépirodalmi stílus között.

A továbbiakban érdemes lenne egy tipikusabbnak mondható topic felismerési feladatot is kipróbálni a korpuszon, de ehhez szükség van egy jó minőségű tanító korpuszra, melynek előállítására ezen cikk megszületéséig nem volt lehetőségünk.

4.4 Az n érték választásának hatása n-grammos klasszifikáció esetén

Ebben az előkészítő jellegű mérésben arra kerestük a választ, hogy n értékének megválasztása hogyan befolyásolja az n-gramm alapú klasszifikáció eredményét. Megvizsgáltuk különböző n-ekre a jellemzőszám függvényében kapott osztályozási pontosságot, eredményeink az 2. ábrán és 3. ábrán láthatóak.

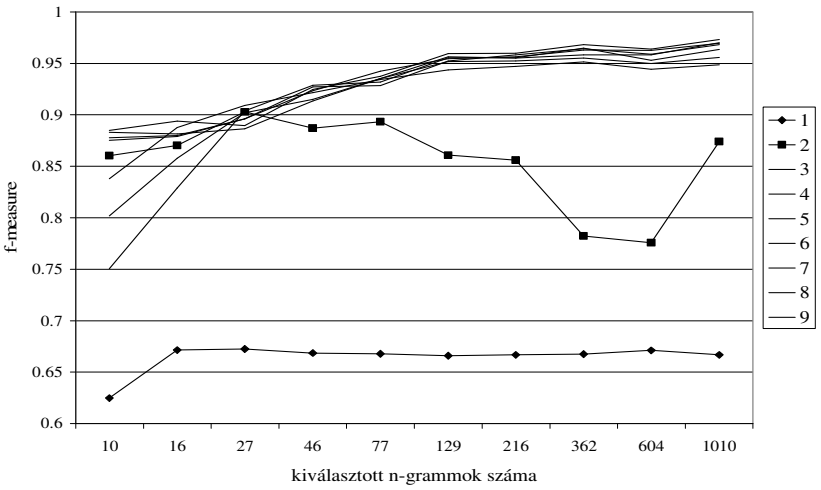


Fig. 2. Az Index Sport rovatába tartozó cikkek felismerése n-grammok (n=1, 2, 3, 4, 5, 6, 7, 8, 9) előfordulásai alapján, a használt releváns jellemzők számának függvényében

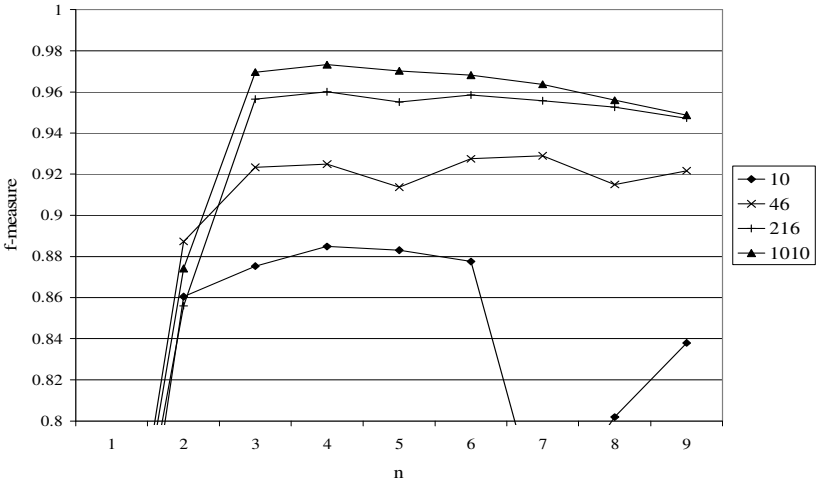


Fig. 3. Az Index Sport rovatába tartozó cikkek felismerése n-grammok előfordulásai alapján, 10, 46, 216 ill. 1010 legrelevánsabb n-gramm alapján osztályozva *n* függvényében

Az 1-gramm alapú osztályozás a várakozásoknak megfelelően használhatatlan eredményt ad, a 2-gramm már lényegesen jobb, de még mindig sokkal gyengébb a 3-gramm alapúnál, ám n=3-tól kezdve n-et tovább növelve már nem javul lényegesen az eredmény. A magasabb n-ekre a teljesítmény alakulása a 3. ábrán jobban megfigyelhető. A fenti és további hasonló kísérletek alapján végül is az n=4 választás mellett döntöttünk.

4.5 A jellemző választás hatása a magyar nyelvű szövegek osztályozására

Ebben a – cikk fő mondanivalóját alátámasztó – kísérletben az Indexes korpuszon végeztünk szövegklasszifikációt különböző jellemzőhalmaz választások mellett. A kipróbált jellemzőhalmazok a fentiekben részletesen ismertetett szavak halmaza (a klasszikus megoldás), szótövek halmaza (az előzőből szótövesítéssel kapjuk) ill. 4-grammok halmaza. A feladat a sport kategóriába tartozó cikkek elválasztása volt az összes többitől. A kapott eredményeket lásd a 4. ábrán!

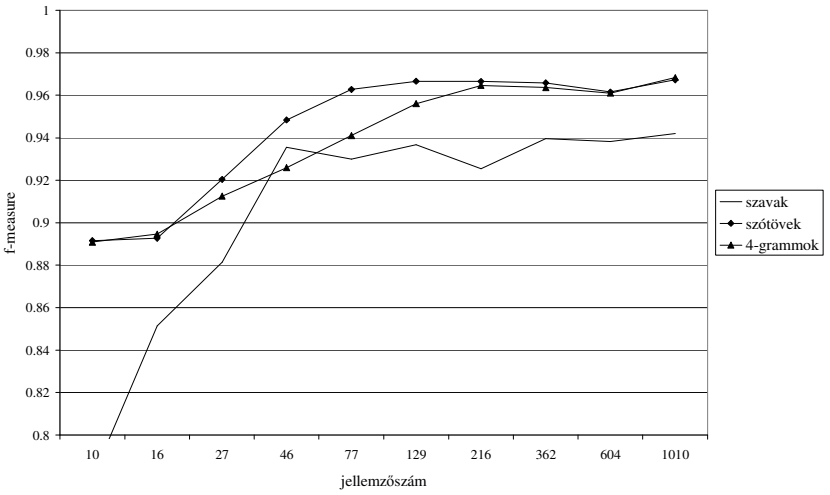


Fig. 4. Az Index Sport rovatába tartozó cikkek felismerése szó, szótő ill. n-grammok (n=4) előfordulásai alapján, a használt releváns jellemzők számának függvényében

Jól megfigyelhető, hogy a legjobb megoldást a szótövesítés adja, de a 4-gramm alapú technika teljesítménye is legfeljebb egy százalékkal marad el, közepes jellemzőszámok mellett. Nagy és kicsi jellemzőszámok mellett a két technika közel azonos teljesítményt ad.

4.6 Zajos dokumentumok vizsgálata

Ezzel a kísérlettel azt a feltevésünket próbáltuk ellenőrizni, hogy az n-gramm alapú klasszifikáció a másik két módszernél lényegesen robosztusabb hibás szövegek esetén. A szövegekben előforduló hibák szimulálására minden egyes karaktert egy adott valószínűséggel egy véletlenszerűen kiválasztott másikra cserélünk. Itt rögzítjük a jellemzőszámot, a 216 legrelevánsabb jellemzőt használjuk (ezt a választást az előző rész eredményei támasztják alá). A hibavalószínűség függvényében a három módszer által adott klasszifikáció minőségét az alábbi ábra foglalja össze.

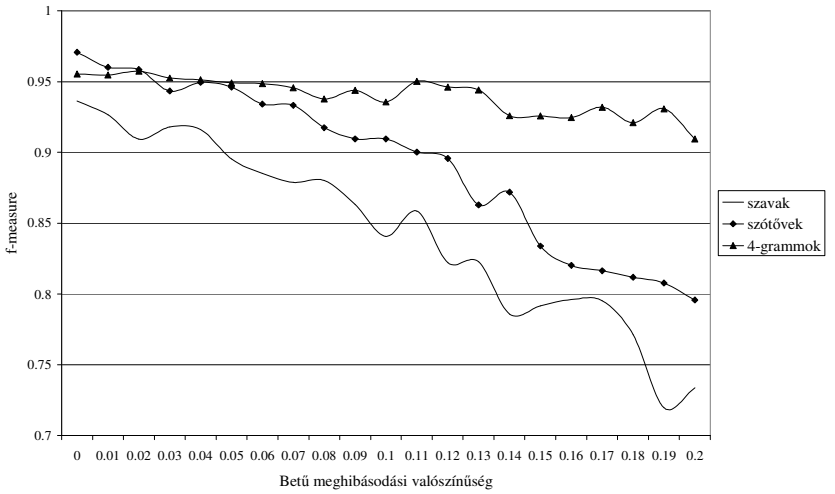


Fig. 5. Az Index Sport rovatába tartozó cikkek felismerése szó, szótő ill. n-grammok (n=4) előfordulásai alapján, a hozzáadott zaj mértékének függvényében

Az n-grammos módszer eredményei a zaj függvényében lényegesen lassabban romlanak, mint a másik két módszer esetében. Még 20%-os hiba esetén is kevesebb mint 5%-kal csökken az osztályozási hatékonyság. Ez fontos lehet olyan esetekben, amikor a szöveg valamilyen (pl. beszéd vagy írás) felismerés eredménye, vagy egyéb okból (pl. e-mail) zajos.

5 Következtetések

A magyar nyelvű szövegek klasszifikációjában komoly segítséget jelent, ha valamilyen módon megpróbálunk a különböző szóalakok közös kezelésére lehetőséget biztosítani. Erre megfelelő nyelvészeti apparátus birtokában a legjobbnak bizonyuló megoldást a szótövesítés adja. Azonban a sokkal kevesebb fejlesztési és futási időt igénylő n-gramm alapú osztályozók teljesítménye alig marad el a szótő alapú osztályozásétól.

A jelenség azzal magyarázható, hogy a betű hármasok, betű négyesek egy jelentős része már tipikusan egyetlen szó különböző szóalakjaira jellemző, így jó eszköz ezek összefogására, együtt kezelésére. Bizonyos n-grammok persze nem ilyenek, pl. a tipikus toldalékokhoz tartozó betűsorozatok, ám ezektől automatikusan megszabadulunk a relevancia szerinti válogatás során.

Ha a készítendő osztályozót várhatóan zajos dokumentumokra fogjuk alkalmazni, vagy ha nem ismerjük a szóhatárokat, akkor az n-gramm alapú osztályozás nem csak olcsóbb, de jobb megoldást is ad.

6 Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki Gyepesi Györgynek és Varga Dánielnek a JMorph morfológiai elemzővel ill. a Hunglish korpusszal kapcsolatos értékes segítségükért. A jelen cikkben ismertetett kutatást részben a Kutatás-fejlesztési Pályázati és Kutatáshasznosítási Iroda GVOP-3.1.1.-2004-05-0363/3.0 számú Orvosi szak szövegek interaktív tartalomelemzése elektronikus kórlapok kitöltésére című pályázata támogatta.

Bibliográfia

1. Cavnar, W. B., Trenkle, J. M.: N-Gram-Based Text Categorization. In Proceedings of SDAIR-94 (1994) 161–175
2. Fan, R.-E., Chen, P.-H., Lin, C.-J.: Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University (2005), http://www.csie.ntu.edu.tw/~cjlin/libsvm/
3. Forman, G.: Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Lecturre Notes in Computer Science, Vol. 2431. Springer-Verlag, London UK (2002) 150–162
4. Index.hu hírportál. http://www.index.hu
5. Kornai A., Krellenstein M., Mulligan M., Twomey D.,Veress F., Wysoker A.: Classifying the Hungarian web. In Copestake and Hajic (eds): Proceedings of EACL 2003 203–210
6. Kwok, J. T.: Automated text categorization using support vector machine. In Proceedings of ICONIP'98, 5th International Conference on Neural Information Processing, Kitakyushu, Japan (1998) 347–351
7. Langdon, W. B.: Natural Language Text Classification and Filtering with Trigrams and Evolutionary NN Classifiers. In Darrell Whitley (ed): Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference, Las Vegas, Nevada, USA (2000) 210–217
8. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga D.: Hunmorph: open source word analysis. In Proceedings of ACL Software Workshop (2005) 77–85
9. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V.: Parallel corpora for medium density languages. Proeedings of RANLP 2005, to appear
10. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Int. J. Information Retrieval Vol. 1/1-2 (1999) 69–90

Az automatikus terminológiakivonatolás módszerei és eredményei

Kis Balázs[1,2], Pohl Gábor[3]

[1] MorphoLogic Kft.
kis@morphologic.hu
[2] SZAK Kiadó Kft.
balazs.kis@szak.hu
[3] Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar
pohl@itk.ppke.hu

Kivonat: A terminológiakivonatolás létfontosságú mind a szakfordítási, mind pedig a terminográfiai/lexikográfiai munkában. Ennek elsősorban gazdasági jelentősége van: az általunk kifejlesztett módszerekkel jelenleg 3-6 óra alatt autentikus terminuslistához lehet jutni egy 600 oldalas (kb. 180 000 szövegszavas) szövegből, míg ugyanennek a listának a manuális előállítása 3-6 embernap munkát igényel.

Az előadás az előző évi, hasonló témájú előadás óta elvégzett kísérleteket és azok eredményeit mutatja be. Esettanulmányokon keresztül ismerteti az eddig kifejlesztett módszerek gyakorlati felhasználását.

A szakfordítás egyszerre néhány tíz–néhány száz oldalnyi (10 000–200 000 szövegszónyi) szöveggel foglalkozik. Ez erősen korlátozza a statisztikai módszerek alkalmazását, hiszen gyakorisági devianciákat, illetve asszociációs mértékeket csak jóval nagyobb korpuszokon lehet eredményesen számítani. Emiatt a terminológiakivonatolásra elsősorban szótáralapú, mintaillesztéses, illetve a legújabban környezetvizsgáló eljárásokat alkalmazunk.

1 A terminológiakivonatolás rendeltetése

A terminológia a szakmai nyelvhasználat elsődleges eszköze. A szakmai kommunikáció jelentős része fordításokon keresztül zajlik, ahol a fordítási folyamat sikerét elsősorban a megfelelő és konzisztens terminológiahasználat biztosítja. [12]

A fordítási folyamat résztvevői azonban nem férnek hozzá egykönnyen a terminológiához, amelynek használata a forrásnyelvi szövegben csak implicit módon jelenik meg. Emiatt, ha a nagyobb szövegek lefordítására rövid határidővel van szükség, a fordítás párhuzamosítása előtt a terminológiát elő kell készíteni [7]. A terminológia előkészítése hosszadalmas művelet, mert maximális minőségi követelmények esetén megkívánja a teljes forrásnyelvi szöveg végigolvasását. A gépi terminológiakivonatolás ezt a lépést rövidíti le jelentősen.

A gépi terminológiakivonatolás fontos szerepet tölthet be kiadványok tárgymutatóinak előkészítésében és a lexikográfiai munkában is: korpuszalapú szótárak készíté-

sekor a különböző kivonatolási eljárások a címszavak kiválasztását könnyíthetik meg. [9]

A fentiek mellett a gépi fordítás is hasznot húzhat az terminológiakivonatolásból: a fordítórendszerek által használt témaspecifikus gépi szótárak (machine-readable dictionaries; MRD) összeállításához használható, különösen akkor, ha kétnyelvű eljárások alkalmazásával a forrásnyelvi kifejezéseket és azok célnyelvi megfelelőit egyaránt előállítja. [3]

2 A kivonatolási eljárások

A terminológiakivonatoláshoz sokféle módszer használható, azonban nem alkalmazhatók változatlan formában azok a módszerek, amelyeket általánosan használunk többszavas kifejezések (lexémák) kivonására korpuszokból. Ennek az az oka, hogy a feldolgozandó forrásszövegek potenciálisan nem elég nagy terjedelműek ahhoz, hogy a korpusznyelvészet statisztikai eljárásai alkalmazhatóak legyenek rájuk. [6][10][11] E ponton már félreérthetetlenül látszik, hogy a fejlesztésünk tisztán gyakorlati jellegű, vagyis arra összpontosítunk, hogy eszközünk valódi fordítási feladatokban, valódi forrásnyelvi szövegeken jelentős hatékonyságnövekedést eredményezzen.

A terminológia modellezése – amelyre a terminológiakivonatolási eljárások épülnek – nem könnyű feladat. A fejlesztés során a következő alapfeltevésekből indultunk ki, amelyek a terminológiakivonatolási feladat három megközelítését nyújtják:

(1) A terminológia a szakmai nyelvhasználat alapvető attribútuma. Kis (2003) szerint a szakmai nyelvhasználat a terminológiai magatartás eredménye [7], ahol a szakmai szöveg a terminológia elemei – a terminus technicusok – mint váz köré épül. A terminológiakivonatolás feladata a váz elemeinek kielemelése.

(2) A terminusok olyan egy- vagy többszavas lexémák, amelyek a szövegben terminológiai helyzetben szerepelnek [7]. A terminológiakivonatolás feladata a szövegben szereplő szavak és kollokációk terminológiai helyzetének megállapítása vagy cáfolata.

(3) Ugyanaz a szó vagy kollokáció lehet része a szakmai szövegbeli terminológiának, más előfordulásaiazonban a diskurzust, illetve a szöveg koherenciáját biztosító nyelvi elemek közé tartozhatnak. Emellett, bár a szabványos terminológiával szemben elvárás a monoszémia, egyes terminus technicusok meghatározott szövegekben többértelműek is lehetnek (ennek oka például az lehet, hogy a szöveg több szakmai területhez is tartozik). A terminológiakivonatolás feladata, hogy megállapítsa a szöveg egyes lexémáinak (terminológiai) szerepét.

Az itt következő eljárások elsősorban a terminológiai helyzet felismerésére irányulnak, vagyis különböző kritériumokat állítanak fel arra, hogy egy adott szó vagy kollokáció terminológiai helyzetben van-e. Egyértelműsítést nem végeznek, vagyis a pontos terminológiai szerepet nem állapítják meg.

A terminológiakereső eljárások általában szabályalapúak, bár a korpusznyelvészet számos statisztikai eljárást (asszociációs mértékeket) alkalmaz kollokációk keresésére. [6][10][11] A terminológiakivonatolás bemeneti szövegei azonban általában túlságosan kis terjedelműek (< 200 000 szövegszó) ahhoz, hogy a statisztikai eljárások megbízható eredményhez vezessenek. Ugyanakkor egyes esetekben a nagy korpuszokra alkalmazott statisztikai számítások hasznosak lehetnek a szabályok előkészítésében.

2.1 Mintakereső eljárások

A mintakereső eljárások a terminus technicusok morfológiai-morfoszintaktikai jellemzőit próbálják megállapítani, s az ilyen jellemzőkkel rendelkező szavakat és kollokációkat keresik a szövegben. [5][7][11] A jelenlegi megvalósításunkban ezek morfoszintaktikai címkékből álló minták, amelyekben az egyes címkék felszíni sorrend szerint következnek. Példa angol nyelvű szöveg kivonatolásához használatos mintákra:

```
UNKNOWN+ADJ+N
ADJ+ADJ+N
ADJ+NUM+N
```

Ez természetesen nem a teljes mintasorozat, a jelenlegi terminológiai modellek 20-24 mintát alkalmaznak. Ez a fajta kivonatolási eljárás azonban túl sok zajt eredményez [7], ezért rendszerbe állítottunk egy heurisztikus utószűrő modult, amellyel egyes szisztematikus hibákat kerültünk el. Ezek a hibák általában gyakori, produktív kollokációkat generáló szavak formájában jelennek meg.

Az utószűrés valójában két lépésből áll:

(1) Mivel a szöveg szószintű elemzéséhez egyelőre nem egyértelműsítő szófaji jelölőt (POS-tagger), hanem morfológiai elemző programot alkalmazunk, annak eredményét is utószűrjük. Ez a szűrés feladatspecifikus, a morfológiai elemző által visszaadott egyes elemzések alkalmazását megtiltjuk.

(2) A primér mintakereső modul által visszaadott mintákat lexikális összetételük alapján szűrjük. Itt azokat a mintákat szűrjük ki, amelyek meghatározott szavakkal kezdődnek vagy végződnek.

Ezzel a módszerrel a kivonatolás pontossága (precision) kb. 20 százalékponttal javítható, ezzel 60-80%-os pontosság érhető el. Az eljárást egyelőre angol és magyar forrásnyelvű szövegek esetén alkalmaztuk. [11]

A módszert azzal fejlesztettük tovább, hogy a morfoszintaktikai minták formalizmusában eleve megengedjük lexikális korlátozások alkalmazását, hasonlóan a korábban tartalomelemzéshez alkalmazott mondatelemző rendszerhez (vö. [8]). Az utószűrést pedig teljes körűvé tettük, vagyis a javított változatban nemcsak a minták első és utolsó elemét lehet vizsgálni.

Ez utóbbi a módszerrel tovább javítható a kivonatolás pontossága, de ehhez a minták számát és bonyolultságát is növelni kell. Az elképzelt ideális felhasználói munkamódszer iteratív jellegű, vagyis a kezdeti kivonatolás után a felhasználó manuális utószűrést végez, de eközben szisztematikus szűrési utasításokat is kiad. Ezek az utasítások új utószűrési szabályok létrehozását eredményezik; emiatt olyan módszer alkalmazására lesz szükség, amellyel automatikusan is létrehozhatók ilyen szabályok.

2.2 Szótáras eljárások

A mintakiemelő eljárások alkalmasak ismeretlen többszavas terminus technicusok megkeresésére. Azonban célszerű kihasználni, hogy a legtöbb témakörben rendelkezünk kiinduló terminológiával, vagyis a forrásnyelvi szöveg terminológiája nem ismeretlen egészében.

Két szótáras eljárást alkalmazunk:

(1) Induktív terminológiakeresés [1][5]: ismert – szótárban tárolt – terminus technicusok összes alakjának előfordulásait keressük a forrásnyelvi szövegben, és felírjuk azon kollokációikat, amelyek megfelelnek a 2.1. fejezetben – a mintakereső eljárásoknál – leírt mintáknak. Amennyiben a mintakereső eljárások mintái kellően megengedőek, ez a szótáras eljárás mindenképpen szűkebb halmazt eredményez, mint az általános mintakeresés. A két módszer együttes alkalmazása esetén lehetőségünk van a terminusjelöltek bizonyos fokú automatikus értékelésére, mivel a szótáras induktív eljárás által megtalált terminus technicusok érvényessége valószínűbb. Magasabb pontszámot rendelhetünk azokhoz a találatokhoz, amelyek megtalálhatók voltak a szótárban, alacsonyabbat kaphatnak azok a jelöltek, amelyek egy szótári terminus technicus és egy vagy több szövegszó kollokációjából állnak, a legalacsonyabbat pedig azok a jelöltek kapják, amelyeket az általános mintakeresés eredményeként kaptunk.

(2) Az általános szókincshez nem tartozó egyszavas terminus technicusok keresése. Az eddig említett eljárások csak a többszavas terminus technicusok megkeresésére alkalmasak. Ez a szótáras eljárás azokat a szavakat keresi meg a szövegben, amelyek nem szerepelnek egy kellőképpen szűk alapszókincsben. Az alapszókincset szótárral reprezentáljuk, ez a szótár az angol és a magyar nyelvű kivonatolás esetén kb. 20 ezer címszót tartalmaz. Szakmai szövegekben valószínű, hogy az alapszókincsben nem szereplő szavak a terminológiához tartoznak, azonban sok alapszókincsbeli szó is megjelenhet terminológiai szerepben. Ez az eljárás nem alkalmas az utóbbiak felismerésére.

A második eljárás kiegészíthető kollokációkereséssel is, vagyis kereshetjük az alapszókincsben nem szereplő szavak azon kollokációit, amelyek megfelelnek a 2.1. részben – a mintakereső eljárásoknál – leírt mintáknak. Ebben az esetben az általános mintakereséssel nyert egyes jelölteket „erősíthetünk”.

A fenti eljárások közül a terminológiakivonatoló alkalmazásba egyelőre csak a másodikat integráltuk, az első eljárás egyelőre külön modulként létezik.

2.3 Környezetvizsgáló eljárások

A környezetvizsgáló eljárások nem a jelöltek attribútumait (belső szerkezetét), hanem a környezetük jellemzőit vizsgálják. E módszerek épülhetnek a környezet (és a jelölt) grammatikai tulajdonságaira, illetve a forrásnyelvi szöveg formázására (ha elérhető). Ilyen eljárásokat egyelőre nem implementáltunk. A lehetséges módszerek:

(1) Keressük a forrásnyelvi szövegben előforduló definíciókat, s ezek alanyát emeljük ki mint terminusjelöltet. Az eljárás megvalósításához sekély szintaktikai elemző program szükséges, amellyel egyfelől a szövegben levő főnévi csoportok határainak megkeresésére, másrészt pedig a definícióra utaló felszíni jegyek felismerésére használunk.

(2) A címek megkeresése. A szakmai szövegek belsejében szereplő címek főnévi csoportjai nagy valószínűséggel terminus technicusok, ezért egy olyan eljárás, amely felismeri a szövegbeli címeket [15], és megkeresi bennük a főnévi csoportokat, igen nagy pontosságú jelöltlistát eredményez.

2.4 Statisztikai módszerek: az egyszavas terminus technicusok megtalálása

Az eddig alkalmazott eljárások megfelelőnek bizonyultak többszavas terminus technicusok megkeresésére – abban az értelemben, hogy gyakorlati feladatokhoz jól használhatók, bár pontosságuk jelentősen növelhető.

Az egyszavas terminus technicusok azonban gyakran rejtve maradnak a mintakereső eljárások előtt, mivel azok szabályai közé az egyszavas mintákat általában nem vesszük fel. Így azok az egyszavas terminus technicusok, amelyek részei az alapszókincsnek, nem jelennek meg a jelöltlistán. Ezek megkeresésére alkalmazhatunk statisztikát: egyfajta „deviáns gyakoriság" módszert, amely az egyes szavak relatív gyakoriságát vizsgálja. Egyfelől szükség van egy nagy terjedelmű köznyelvi korpuszból nyert szóstatisztikára, másfelől pedig ki kell számítani a vizsgált forrásnyelvi szöveg szavainak relatív gyakoriságát. Ha egy szó relatív gyakorisága egy meghatározott küszöbértékkel meghaladja a köznyelvi korpuszbeli adatot, jó jelöltté válhat a terminus technicusok listáján.

A fenti módszer implementálása folyamatban van. Feltételezésünk szerint elsősorban 10 000 szövegszót meghaladó terjedelmű forrásnyelvi szövegeken használható majd megfelelően.

3 Kétnyelvű terminológiakivonatolás

A munkafolyamatot tekintve a kétnyelvű terminológiakivonatolást két lépésben valósítottuk meg:

(1) Automatikus egynyelvű terminológiakivonatolás

(2) A jelöltlista célnyelvi megfelelőinek megkeresése

A kétnyelvű terminológiakivonatolást ideális esetben szinkronizált párhuzamos szövegeken végezzük. [3][4][17] Ilyen párhuzamos szöveg a fordítómemória, amely eredeti rendeltetése szerint korábbi fordítások újrafelhasználására szolgál. Azonban a konkrét terminológiakivonatolási feladat számára – amennyiben fordítások előkészítésére használjuk – csak a forrásnyelvi szöveg ismert, mert a folyamat elvárt kimenete éppen a célnyelvi fordítás. Emiatt a kétnyelvű terminológiakivonatolás első és második lépése különböző bemeneti adatokkal működik.

A kétnyelvű terminológiakivonatolás első lépésében a forrásnyelvi szövegen egynyelvű terminológiakivonatolást végzünk, majd meghatározzuk a végleges terminuslistát. Megjegyezzük: ha a kivonatolási feladat nem fordítás előkészítésére, hanem például szótári címszavak kiválasztására, illetve szócikkek építésére szolgál, akkor az egynyelvű kivonatolás futhat párhuzamos szöveg (fordítómemória) forrásnyelvi oldalán is.

A forrásnyelvi terminuslista célnyelvi megfelelőinek meghatározása egyrészt szótárral, másrészt pedig párhuzamos szövegeken végezhető. Főnévi csoportok fordítá-

sának azonosítására alkalmazhatók már kidolgozott speciális módszerek [14]. Amennyiben rendelkezésre állnak korábbi projektekből terminológiai szótárak, egyes forrásnyelvi terminus technicusok megfelelői abban is megkereshetők. Azonban a legtöbb fordítási feladatban van új terminológia, ezért e módszer fedése sohasem 100%. A fordítási feladatokhoz ritkán készítik elő a terminológiát, ezért azokhoz fordítómemória gyakran rendelkezésre áll, korábbi terminológiai szószedet azonban alig. Ezért a piacon rendelkezésre álló fordítástámogató eszközök majdnem mindegyike nyújt konkordanciaszolgáltatást is, amely fordítómemóriák forrásnyelvi oldalán keresi meg szavak, kollokációk előfordulását, és megjeleníti az ezeket tartalmazó mondatok (szegmensek) fordítását – a konkrét terminus technicus célnyelvi pozícióját azonban már nem.

A szótáras módszer triviális, ezért a továbbiakban a célnyelvi megfelelő párhuzamos szövegből való kinyerésére összpontosítunk. A konkrét feladat olyan módszer megalkotása, amely nagy pontossággal megtalálja a forrásnyelvi terminusoknak megfelelő célnyelvi szavakat vagy kollokációkat a párhuzamos szöveg célnyelvi oldalán. Ennek előfeltétele a párhuzamos szöveg megfelelő mondatszintű szinkronizálása.

E keresés kiindulópontja valamiféle fordítási modell felállítása, amely a forrásnyelvi szavaknak célnyelvi szavakat feleltet meg. Ez általában olyan valószínűségi modell, amely annak valószínűségét határozza meg, hogy adott célnyelvi szó fordítása-e adott forrásnyelvi szónak: $P(w_T|w_S)$.

A fordítási modell lehet teljes: ez azt jelenti, hogy elvégezzük a párhuzamos szöveg teljes szószintű szinkronizálását, amelynek során a maximális $P(w_T|w_S)$ valószínűségű szópárokat keressük [3][4][14][17]. Ezeket az eljárásokat általában a statisztikai gépi fordítással összefüggésben alkalmazzák. Ebben az esetben a párhuzamos szöveg feldolgozása függetlenül végezhető az egynyelvű terminológiakivonatolástól. Az egynyelvű terminológiakivonatolás eredménylistáját a szószinten szinkronizált párhuzamos szövegen futtatjuk végig, ahol a forrásnyelvi terminus technicusok szavainak célnyelvi megfelelőit keressük ki. Ebben az eljárásban továbbra is feladat marad a többnyelvű terminus technicusok esetleg szintén többszavas fordításainak megtalálása és a megfelelő célnyelvi kifejezés (morfoszintaktikai/szintaktikai szerkezet) helyreállítása.

A rendelkezésre álló párhuzamos szöveg azonban gyakran túl kis terjedelmű ahhoz, hogy teljes fordítási modell felállításával jó minőségű eredményhez jussunk. A teljes fordítási modellre vezető algoritmusokra, illetve a statisztikai gépi fordítás eljárásaira általában is jellemző, hogy rendkívül nagy mennyiségű (több millió, több tízmillió szövegszónyi) párhuzamos szöveget igényelnek a megfelelő működéshez. Ez a mennyiség azonban a konkrét fordítási feladat esetén gyakran nem áll rendelkezésre.

A statisztikai gépi fordításban általában megengedhető, hogy különböző forrásból származó, különböző tárgykörökhöz tartozó szövegek együttes felhasználásával érjék el a „kritikus tömeget”. Ez azonban épp a terminológia meghatározása esetén ronthatja az eljárás minőségét, mert a terminológiával szemben elvárás, hogy meghatározott témakörnek, illetve szövegtípusnak megfelelően egy bizonyos fordítást alkalmazzunk, ennek megtalálására pedig nagyobb az esélyünk, ha szűrjük a kétnyelvű kivonatoláshoz felhasznált párhuzamos szövegeket.

A kétnyelvű terminológiakivonatolásban ezért „részleges” fordítási modellekkel is kísérletezünk. Ez azt jelenti, hogy a kiindulási párhuzamos szövegeink csak mondatszinten vannak szinkronizálva [13][16], és ebben keressük a forrásnyelvi (szűrt) ter-

minuslista elemeinek előfordulásait, illetve azok célnyelvi megfelelőit. A fordítási modell felállításának legfontosabb előfeltevése, hogy a terminológia fordítása konzisztens – vagyis arra számítunk, hogy a terminológiai szerepben megjelenő szavak/kifejezések fordításának lexikális összetétele mindig ugyanaz lesz. Ugyanakkor nem beszélünk a konkrét nyelvi megvalósításról, mert az mindig más lehet – ezért a forrás- és célnyelvi szegmensek szavait mindig szótő-visszaállítón és szűrőszólistán keresztül nézzük.

Árnyalja a modellt az is, hogy ugyanaz a terminus technicus – különösen, ha egyszavas – megjelenhet terminológiai helytzetben és azon kívül is, illetve interdiszciplináris szakmai szövegekben egyes terminus technicusok a terminológiai helyzetben maradva is lehetnek többértelműek. Ezért nem alkothatunk kizárólagos modellt.

Egyelőre csak kísérletek folynak e módszerek implementálására. A módszer hasonló az asszociációs mértékek számításához: azokat a célnyelvi szavakat keressük, amelyek szignifikánsan nagyobb valószínűséggel fordulnak elő olyan célnyelvi szegmensekben, amelyek forrásnyelvi oldalán a nekik megfelelő forrásnyelvi terminus technicus megtalálható. Amennyiben ez egyes szavakra nem bontható le, a célnyelvi kereshetünk két- és háromelemű kollokációkat is, amelyek esetén – mivel a terminus technicusok megfigyelésünk szerint erősen összefüggő struktúrát alkotnak – kihasználhatjuk, hogy a többszavas terminus technicusok elemei a felszínen valószínűleg szomszédosak lesznek egymással.

A fenti kísérletekre azért van szükség, mert a teljes fordítási modellek csak olyan terminus technicusok célnyelvi megfelelőinek megtalálására alkalmazhatók biztonságosan, amelyek legalább négyszer-ötször előfordulnak a forrásnyelvi szövegben. Bár a terminológiahasználat alapvető követelménye a konzisztencia, a konzisztencia pedig feltételezi az ismétlődést (tehát a többszöri előfordulást), a konkrét forrásnyelvi szövegekben a kivonatolás utáni utószűrés során elfogadott terminus technicusok 30-60%-a csak egyszer fordul elő. Mivel pedig a kétnyelvű kivonatoláshoz rendelkezésre álló párhuzamos szövegek terjedelme gyakran nem haladja meg nagyságrendelkkel a forrásnyelvi szövegét, ezért ott is nagy számban lesznek olyan terminus technicusok, amelyek a korpuszban csak egyszer fordulnak elő. Emiatt a részleges fordítási modellt érdemes szótárral támogatni, vagyis az ismert terminológiai megfeleltetéseket – a korábbi szószedeteket – felhasználni szószintű horgonyok kialakítására.

4 Alkalmazási példa

Az automatikus terminológiakivonatolást egy angol nyelvű szakkönyv lefordításának előkészítésére használtuk. A könyv terjedelme 151 738 szövegszó. A terminológiakivonatoláshoz olyan alkalmazást használtunk, amely a 2.1. részben, illetve a 2.2. részben leírt eljárásokat alkalmazza együtt. Az autentikus terminuslista az automatikus kivonatolás eredményének manuális utószűrésével állt elő.

A terminológiakivonatolás 12 094 jelöltet adott vissza, ebből a manuális utószűrés során 1814 (!) terminus technicust fogadtunk el. Nagyon fontos megjegyezni, hogy a manuális utószűrés eredménye nem tükrözi az eljárás pontosságát, mivel utólagos szerkesztőségi döntés alapján kb. 4000 programnyelvi kulcsszót töröltünk.

A manuális utószűrés ebben az esetben kb. 4 órát vett igénybe. Ezt az időt a könyv teljes szövegének végigolvasásához és a terminus technicusok manuális kijelöléséhez szükséges idővel kell összevetni.

A fordítási terminológia lényege azonban az, hogy a forrásnyelvi szöveg terminológiájához egyértelmű fordításokat rendel. Mivel ebben a munkában csak a forrásnyelvi szövegből nyertünk ki automatikusan terminus technicusokat, a fordítások meghatározása a manuális utómunkához tartozik. Ezt a jelen esetben a projekt terminológusa végezte, egy korábbi, hasonló témájú fordítási projekt terminológiai szótárának felhasználásával. [12]

5 A továbbfejlesztés irányai

Pillanatnyilag egy mintakereső és egy szótáras eljárást használunk, egy alkalmazásba integrálva. További egy szótáras eljárás megvalósítása megtörtént, a statisztikai módszer, illetve a részleges fordítási modellt alkalmazó kétnyelvű kivonatolási eljárás implementálása folyamatban van.

A továbbfejlesztés során meg kell valósítani az iteratív munkát lehetővé tevő felhasználói felületet, illetve az utószűrési szabályok (fél)automatikus generálását. Amikor pedig minden fentebb vázolt kivonatolási eljárás megvalósítása megtörtént, további kísérleteket kell végezni a pontosság növelése végett.

Az alkalmazott kivonatolási eljárások nyelvfüggetlenek, pontosabban adatvezéreltek: működésükhöz forrásnyelvi szótő-visszaállító és morfológiai elemző program (illetve, ha rendelkezésre áll, szófaji címkéző program) szükséges, emellett pedig a kivonatolási szabályokat nyelv- és néha szövegtípus-függő módon kell összeállítani. Utóbbiak azonban hozzáférhetők és szerkeszthetők a felhasználó számára.

A jövőbeli feladatok közé tartozik az is, hogy az itt kidolgozott eljárásokat további nyelvekre is kipróbáljuk.

6 Köszönetnyilvánítás

A szerzők szeretnének köszönetet mondani Prószéky Gábornak (MorphoLogic Kft.) a módszertani tanácsadásért, Ugray Gábornak (Kilgray Kft.) a statisztikai és az utószűrési módszerek kidolgozásában nyújtott segítségéért, Chris Callison-Burchnek (Linear B) pedig azért, hogy lehetővé tette az általa kidolgozott kereshetőfordítómemória-technológia kipróbálását.

Ez az írás közvetlen eredménye az IKTA-00181/2003. számú, a Magyar Köztársaság Oktatási Minisztériuma által támogatott projektnek.

Bibliográfia

1. Castellví, M. T. C., Bagot, R. E. and Palatresi, J.(2001), Automatic term detection: A review of current systems, in D. Bourigault, C. Jacquemin and M.-C. L'Homme (eds), Recent Advances in Computational Terminology, John Benjamins, Amsterdam–Philadelphia, 53–88.
2. Hodász Gábor, Pohl Gábor (2005): MetaMorpho TM: a linguistically enriched translation memory. In: International Workshop, Modern Approaches in Translation Technologies (ed. Walter Hahn, John Hutchins, Cristina Vertan), Borovets, Bulgaria.

3. I. Dan Melamed (2000), Models of Translational Equivalence among Words, Computational Linguistics 26(2), 221-249.
4. I. Dan Melamed (2001). Empirical Methods for Exploiting Parallel Texts. MIT Press.
5. Jacquemin, C.(2001), Spotting and Discovering Terms through Natural Language Processing, MIT Press, Cambridge (Mass.).
6. Kilgarrif, A. and Tugwell, D.(2001),Word sketch: extraction and display of significant collocations for lexicography, Proceedings of the 39th ACL and 10th EACL Workshop 'Collocation: computational extraction, analysis and explotation', Toulouse, 32–38.
7. Kis Ádám–Kis Balázs–Pohl Gábor (2004), A számítógépes terminológiakivonatolás új megközelítése. In: A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, 63-72.
8. Kis Balázs – Naszódi Mátyás – Prószéky Gábor (2003), Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. Az I. Magyar Számítógépes Nyelvészeti Konferencia előadásai, Szeged, 145-152.
9. Kis, Ádám–Kis, Balázs (2003), A prescriptive corpus-based technical dictionary. development of a multi-purpose technical dictionary, Papers in Computational Lexicography: Proceedings of COMPLEX 2003, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 47–56.
10. Kis, B., Villada, B., Bouma, G., Bíró, T., Nerbonne, J., Ugray, G. and Pohl, G. (2004), A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-Word Lexemes, Proceedings of LREC 2004, Lisbon.
11. Kis, Balázs–Villada Moirón, Begoña–Bíró, Tamás–Bouma, Gosse–Pohl, Gábor–Ugray, Gábor–Nerbonne, John (2004): Methods for the Extraction of Hungarian Multi-Word Lexemes. In: Proceedings of CLIN-2003. University of Antwerp.
12. Lengyel István–Kis Balázs–Ugray Gábor (2004), MemoQ – Új megközelítés a fordítástámogatásban. Infrastruktúratanulmány. In: A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, 100-107.
13. Pohl Gábor (2003): Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás. In: Az I. Magyar Számítógépes Nyelvészeti Konferencia előadásai, Szeged, pp 254-259.
14. Pohl Gábor (2005): Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás. In: III. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete. Szeged.
15. Pohl Gábor–Ugray Gábor (2004): Angol címek felismerése. In: II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szeged, pp 155-160.
16. Robert C. Moore (2002): Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th AMTA Conf: Machine Translation: From Research to Real Users, pages 135-244, Langhorne, PA. Springer.
17. Chris Callison-Burch–Colin Bannard–Josh Schroeder (2005): A compact data structure for searchable translation memories. In: Practical Applications of Machine Translation. Proceedings of the 10th EAMT Conference, Pázmány Péter Catholic University, Budapest.

V. Szintaxis és szemantika

Többszavas kifejezések kezelése MT szótárban

Váradi Tamás

MTA Nyelvtudományi Intézet
1068 Budapest Benczúr u 33
varadi@nytud.hu

Kivonat: A dolgozat a számítógépes alkalmazásokban, elsősorban gépi fordítórendszerekben használt szótárak felépítésének elveit vizsgálja. Az egyedi szavak megfeleltetése helyett a többszavas kifejezések minél nagyobb számú használata lényegesen csökkenti a többértelműséget. A többszavas kifejezések egy folytonos skála mentén helyezhetők el, melynek egyik végpontján a teljes mértékben rögzített szerkezetek, a másikon pedig a teljesen nyílt, azaz minden elemükben megváltoztatható szerkezetek vannak. A dolgozat bemutatja azt, hogyan lehet lokális grammatikákkal nyitott lexikai osztályt tartalmazó kifejezéseket kezelni egy kétnyelvű gépi fordítórendszer szótári komponensében.

1. Bevezetés

A legtöbb számítógépes alkalmazás használhatóságának kulcskérdése a szótári komponens. Robosztus számítógépes nyelvészeti lexikon építése nem képzelhető el úgy, hogy egyszerűen átültetjük a hagyományos szótár tartalmát elektronikus formára. Hagyományos szótáron olyan szótárat értünk, amelyik közvetlen „emberi fogyasztásra" készült, függetlenül attól, hogy könyv alakban vagy elektronikus adathordozón jelent meg. Amint látni fogjuk, az olvasó szerepének feltételezése alapvetően megszabja a szótár tartalmát és tálalásának módját.

A jelen dolgozatban a számítógépes alkalmazásokhoz készülő kétnyelvű szótár szerkezetét tárgyaljuk. A gépi fordító rendszerek és egyéb kétnyelvű számítógépes alkalmazások lexikai megfeleléseinek kidolgozásában alapvető probléma a többértelműség feloldása. Valamirevaló szótárban egy-egy szóhoz általában számos célnyelvi megfelelést társítanak, amelyek az adott szó különböző jelentéseinek felelnek meg. Néha azonos jelentést egy szinonima halmazzal értelmeznek, hasonlóan ahhoz, ahogy azt a Wordnet-ben találjuk. A szótári jelentések tipikusan kontextus nélküli megfeleltetéseket tartalmaznak. Annak eldöntése, hogy egy bizonyos nyelvi kontextusban az adott szó melyik jelentésével szerepel, az ún. WSD *(word sense disambiguation)* igen nehéz feladat, amelyik online alkalmazásban robosztus méretekben még nem vethető be.

Jelentősen könnyíthetjük a kétnyelvi megfeleltetések problémáját, ha nem egyes szavak, hanem többszavas kifejezések megfelelőit keressük. A többtagú kifejezések ugyanis gyakran mintegy magukban foglalják a többértelműsítésükhöz szükséges kontextust is. A „fogás" szó többek között jelentheti egy étkezés részét illetve szorítást, angolul ennek megfelelően a 'course' illetve a 'grip' főnevekkel fordíthatjuk. Kontextuson kívül a „fogás" szó ineherensen többértelmű, viszont a „szoros fogás" kifejezésről önmagában véve is egyértelműen eldönthető, hogy a szorítás értelmezéséről van szó.

2. A hagyományos szótárak korlátai

Kézenfekvőnek mutatkozik, hogy a számítógépes alkalmazások számára felhasználjuk a hagyományos lexikográfia eredményeit. Az embereknek szánt kétnyelvű szótárak azonban minden gazdagságuk ellenére súlyos fogyatékosságokat mutatnak. Amint látni fogjuk, ezek a hiányosságok nem egyes szótárak vagy szótárírók tökéletlenségét, vagy hanyagságát mutatják, hanem a szótárak rendeltetéséből fakadó elvi korlátok. A hagyományok szótárak rendeltetésének kitűnő összefoglalását adja Bolinger [1] alábbi meghatározása:

> Dictionaries do not exist to define, but to help people grasp meanings, and for this purpose their main task is to supply a series of hints and associations that will relate the unknown to something known.
> The dictionary has done its job when it gives the reader a handhold in his own experience — a pair of synonyms, a diagram, a context, a comparison, tied to any convenient reference post.

Nézzük röviden, melyek azok a jelenségek, amelyeket a hagyományos lexikográfia a fenti elvet követve egyszerűen kiiktat a szótárak érdeklődési köréből.

2.1 Hiányzó lexikai egység

A szótárak címszavai egyáltalán nem, vagy csak szórványosan tartalmaznak tulajdonneveket és egyéb enciklopédikus tudást hordozó lexikai elemeket, amelyek azonban nagy gyakorisággal fordulnak elő válogatatlan szövegekben, különösen hírekben. Részben az ilyen elemek pótlására szolgál az ún. nyílt tokenosztályú kifejezések felismerésére kifejlesztett technológia.

De nemcsak a szótári egységek hiányoznak, a meglévő szavak feldolgozása is hagy kívánnivalót. A gazdaságosság jegyében ugyanis a szótárírók nem tüntetik fel kimerítően az egyes címszavakból képzés vagy összetétel által előállítható valamennyi alakot. Az "automatikusan", azaz szabályosan előállítható alakokat általában mellőzik, feltételezve, hogy az anyanyelvi kompetenciával bíró szótárolvasó ezeket szükség esetén mind elő tudja állítani.

További korlát a hagyományos szótárak teljességében az az elv, hogy a lexikográfia a nyelvhasználat „időtálló" elemeinek rögzítését tűzi ki célul. Egy-egy lexikai egység szótárba felvétele előtt a szótáríró intuitív alapon, vagy újabban egyre gyakrabban korpuszgyakorisági adatok alapján, mérlegeli, hogy az adott lexikai elem „megérett-e" arra, hogy bekerüljön a szótárba. A számítógépes alkalmazások azonban

általában a nyers szövegekkel szembesülnek, melyekben jócskán fordulnak elő egyedi, esetleg kérészéletű kifejezések. A nyelvfeldolgozó programok nem engedhetik meg annak mérlegelését, hogy létezőnek ismernek-e el bizonyos szavakat, kifejezéseket.

Ez utóbbi probléma elvi jellegű és feloldhatatlan, hiszen itt az élő, folytonosan változó nyelvhasználat és az azt valamilyen szempont (érték vagy gyakoriság) alapján rögzíteni kívánó norma viszonyáról van szó, amely más fogalmakkal a type/token, vagy a langue/parole, competence/performance ismert dilemmáit veti fel.

2.1.1 Hiányos vagy homályos (fuzzy) információ

Gyakran a hiányolható információ nem maga a lexikai elem, hanem azzal van a gond, hogy definíciójának terjedelme nem határozható meg egyértelműen: vagy teljesen hiányzik vagy homályos annak meghatározása, hogy az adott lexikai egység a jelenségek mely körére alkalmazható.

> **grater** a kitchen UTENSIL (= a tool) with a rough surface,used for grating food into very small pieces: a **cheese/nutmeg grater** (OALD7) [10]
>
> A **grater** is a kitchen tool which has a rough surface that you use for cutting food into small pieces (COBUILD)[7]
>
> **grater** a tool used for grating food: a **cheese grater** (LDCE)[8]

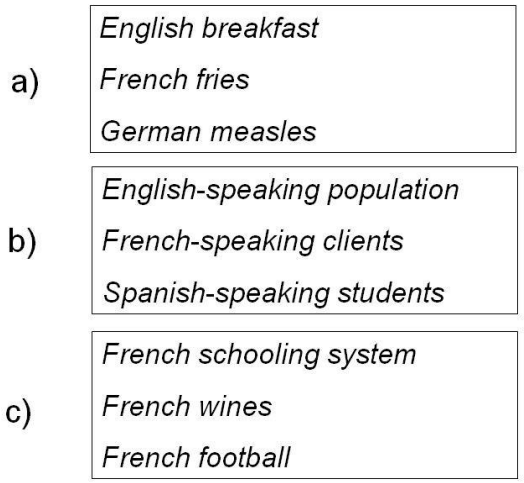
A fenti három, élvonalbeli angol értelmező szótárakból vett idézetből az még talán kitalálható, hogy ételreszelőről van szó, de az már korántsem egyértelmű, hogy a két említett reszelőfajtán kívül milyen további reszelők vannak? Különösen kétnyelvi kontextusban, ahol a szótár olvasó nem tagja a célnyelvi (kulináris) kulturális közösségnek, nem tételezhető fel az (étel)reszelő használati körének egyértelmű ismerete. Számítógépes alkalmazásokban pedig jelenleg reménytelen ilyesfajta mindennapi tudásra apellálni.

2.2 A többszavas kifejezések típusai

Miután beláttuk annak fontosságát, hogy a számítógépes alkalmazásba épített szótár a lehető legexplicitebb legyen, és láttuk a hagyományos szótárak elvi és gyakorlati korlátait e tekintetben, vizsgáljuk meg a többszavas kifejezések néhány fontos jellemzőit.

Az 1. ábra mindegyik példájában nemzetiségneveket találunk. Az a) alatti példák olyan kifejezéseket tartalmaznak, amelyek maximálisan kötöttek, egyik elemük sem változtatható, jelentésük nem teljesen kompozicionális (legalábbis a szinkrónia síkján). A spanyol nátha mellett nincs francia vagy román nátha. A többszavas kifejezéseket tárgyaló szakirodalom elsősorban ilyen rögzült kifejezésekkel foglalkozik lásd pl. [4] magyarra [3]. Ezzel szemben a c) alatti kifejezések mindkét eleme tetszőlegesen változtatható: itt a névszói kifejezés főtagját változtattuk, de a „francia iskolarendszer” helyett vehettünk volna bármilyen tetszőleges nemzetnevet. Ezeknek a kifejezéseknek teljesen áttetsző a szintaktikai és a szemantikai szerkezetük, kompozicionalitásuk mindkét tekintetben maximális.

A kétnyelvű gépi alkalmazások számára mindkét típusú kifejezés kezelése viszonylag egyszerű esetet jelent. Az a) tipusú, teljesen kötött kifejezéseket a szótár szócikkei között fel kell sorolni a hozzá tartozó egy vagy többszavas megfelelővel együtt. Itt legfeljebb csak a minél nagyobb lefedettséghez szükséges ráfordítás jelenthet gondot. A c) tipusú, teljesen nyitott kifejezések viszont nem igényelnek szótári bejegyzést, szó szerint fordíthatók.



1. ábra Többszavas kifejezések a kompozicionalitásuk foka szerint

A jelen dolgozatban a 1b) típusú példákra hívjuk fel a figyelmet. Az itt található kifejezések elemei egyértelműen szerkezetet alkotnak. Figyeljük meg, ha elhagyjuk a jelzőt vagy rosszul képzett kifejezést kapunk vagy lényegesen módosul a jelentés **speaking population*. A kifejezések jelentése sem kompozicionális, hiszen a *speaking* 'beszélő' szó nem beszédtevékenységre, hanem nyelvi kompetenciára vonatkozik. Ugyanakkor a kifejezés elemei nem olyan mértékben rögzítettek, mint azt az a) típusú példáknál láttuk: a példákban szereplő jelzőket tetszőlegesen kombinálhatjuk a többi kifejezés főtagjával (*French speaking population, Spanish speaking students* stb.). Azt találjuk tehát, hogy a b) csoportban a kifejezések részben kötöttek, részben nyitottak. Felmerül tehát a kérdés, hogyan kezelhetjük őket egy számítógépes alkalmazásban?

Először az a kérdés merülhet fel, hogy egyáltalán kell-e velük külön foglalkozni, azaz nem járhatunk-e úgy, mint akár az a) vagy a c) csoport tagjaival. Listába foglalni őket és a célnyelvi megfeleléseket az egyes elemekhez hozzárendelni, ha egyáltalán lehetséges, rendkívül veszteséges eljárás, hiszen ez azt jelentené, hogy a kombinálható elemek teljes permutációját elő kellene állítani, a megfelelőkkel együtt. Ugyanakkor a szó szerinti fordítás sem kivihető, mivel ugyan ebben az esetben történetesen létezik szó szerinti megfelelés *angolul beszélő népesség* stb., annak jelentése kétértelmű (tevékenységet és kompetenciát egyaránt takar.) Más esetben azonban, mint

például a 2a és 2b példákban, sem listába foglalni, sem alkotó elemenként szó szerint fordítani nem tudjuk őket.

2a) a twelve year old boy

2b) egy tizenkét éves fiú

Ha viszont a megfelelés nem áttetsző a két nyelv között, és ennek következtében fel kell vennünk a kétnyelvű szótárban, akkor a gépi alkalmazás számára nem folyamodhatunk a „példálódzás" módszeréhez, ami alkalmanként tökéletesen kielégítő a hagyományos szótárak esetében. A hagyományos szótáraknál elégséges megadni egy megfelelést a maga felszíni alakjában, és a minta kivonását rábízni a szótárolvasó (nyelvi) intelligenciájára. 2a) és 2b) alapján az olvasó szó szerint számtalan hasonló kifejezést tud alkotni mindkét nyelven. A számítógépes rendszer azonban erre pusztán a felszíni alakok alapján képtelen.

2.3 A többszavas kifejezések szerkezete

Az 1b) alatt található kifejezéseket tehát szó szerint fordítani nem lehet, felsorolni pedig vagy nem érdemes vagy nem is lehet. A feladat tehát annak a mintának a megalkotása, amelynek segítségével tetszőleges számban képezhetők illetve megfeleltethetők az ilyen szerkezetű elemek. Szerencsére a kifejezéseken belüli kombinációs megszorítások jól kezelhetők véges állapotú lokális grammatikákkal. A Maurice Gross [2] nevéhez fűződő lexikalizált véges állapotú nyelvtanok implementálására Max Silberstein hatékony eszközt fejlesztett [5],[6], melynek magyar nyelvű alkalmazásáról lásd [9].

A kifejezések szerkezetét elemezve azt találjuk, hogy egyes elemeik konkrét szóalakok (mint pl. *éves*, *nyelvű*) mások viszont vagy teljesen nyílt osztályt alkotnak (számkifejezések, mint pl. *huszonhat, kilenc és fél*) vagy egy terjedelmes, bár felsorolható listát (nemzetiségnevek). A 2. ábra az 1b) kifejezéseket előállító lokális grammatikát mutatja, amely listás felsorolást tartalmazó beágyazott gráfot (Language.grf) használ. Ezzel ekvivalens megoldás, ha a szótárban szemantikai jegyekkel különböztetjük meg a Language.grf elemeit, és a fő gráfban a szemantikai jegyekre hivatkozva definiáljuk a gráf megfelelő helyén előforduló elemek körét (lásd. 3. ábra)

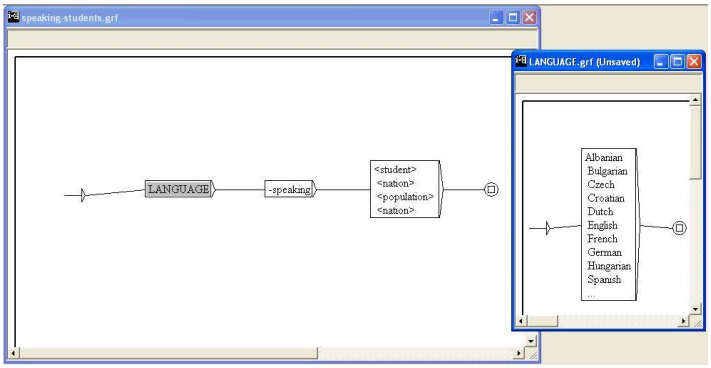
Azt látjuk tehát, hogy a többszavas kifejezések egy része leírható/előállítható egy olyan véges állapotú lokális grammatikával, melynek bizonyos csomópontjait egyedi lexikai elemek töltik ki, a szerkezet más pontjain pedig a lexikai egységek valamilyen osztálya szerepel. Hagyományosan a szófajokat használjuk az egy adott szintaktikai pozícióban behelyettesíthető szavak halmazának képviseletére. A szintaktikai szerkezeteket, például a főnévi csoportot, a hagyományos grammatikában ismert szófajok kombinációjaként jellemezzük. A feltevés az, hogy miután ezen az általános síkon meghatároztuk a végső nem-terminális elemek kapcsolódásait, a szóosztályokat bármely azonos kategóriájú lexikai elemmel helyettesíthetjük.

Teljesen nyitott szerkezetek leírására ez az eljárás meg is felel. A fenti példák azt bizonyítják, hogy a kifejezések egy részére ez a modell nyilvánvalóan alkalmatlan. A „spanyol nyelvű lakosság" kifejezést nem elégséges „Adj Adj N" kategóriák szekvenciájaként jellemeznünk, hiszen ennek számtalan rosszul képzett, vagy értelmezhetetlen alak is megfelel, mint pl. a 3a) vagy a 3b).

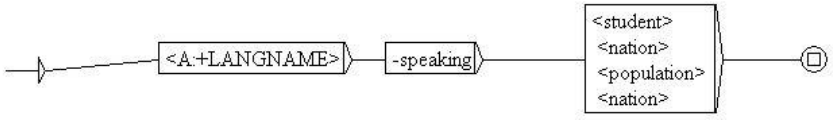
3a) *nyelvű spanyol lakosság

3b) *spanyol nyelvű kardántengely

A szintaxis autonómiája nevében természetesen mondhatjuk, hogy a „spanyol nyelvű kardántengely” kifejezés teljesen grammatikus, értelmezést tulajdonítani neki nem is lehetetlen (pl. spanyol nyelvű felirattal ellátott kardántengely?), még ha az nem is lesz egyező a személyek esetében alkalmazott ’spanyol anyanyelvű’ értelemmel. Ugyanakkor azonban a 3a) példa nem szemantikailag, hanem szintaktikailag rosszul formált, pedig ez is „Adj Adj N” szekvencia. Érvelésünk szempontjából közömbös, hogy szintaktikai vagy szemantikai jellegű megszorításról van szó. A lényeg az, hogy a szófajokkal megadott struktúraleírás messze nem kielégítő, mert a rendkívül durva kategóriákkal nem tudunk számot adni a szerkezet elemei között fennálló finom megkötésekről. Természetesen a nyelvtechnológiai alkalmazásokban annak a kérdésnek, hogy a kifejezés elemei között fennálló kombinációs megszorítások szintaktikai vagy szemantikai/pragmatikai jellegűek-e semmi jelentősége nincs.



2. ábra Az 1b) kifejezés szerkezetét leíró lokális grammatika



3. ábra Az 1b) kifejezések lokális grammatikája szemantikai jegyek használatával

Az a kérdés sem kell, hogy különösebben foglalkoztasson minket, hogy vajon a szótárba vagy a nyelvtanba tartozó jelenségről van szó. Egyrészt ez részben implementációs kérdés, részben pedig nem kategória hanem fokozatos átmenet (kontinuitás) kérdése. Ugyanazt a technológiát (reguláris grammatikát) alkalmazhatjuk a részben, mint a teljesen nyílt kifejezések leírására. A különbség csupán abban rejlik, hogy a szerkezet elemei csak egyedi lexikai elemekből állnak-e (kötött idiómák) illetve, ha

a lexikai elemek csoportjára utalnak, azt hogyan teszik. Ez utóbbi tekintetben a néhány elemű listától a szófaji kategóriával a lehető legáltalánosabban jellemzett osztályig terjed a skála.

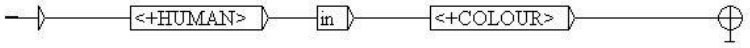
Az itt tárgyalt kifejezésekhez olyan lokális grammatikákat szükséges alkalmazni, amelyek vegyesen tartalmaznak egyedi lexikai elemeket, listákat és szóosztályokat. A véges állapotú gráfok csomópontjain mindig olyan szintű megkötést kell alkalmaznunk, amely a lehető legpontosabban jelöli ki azon lexikai elemek körét, amelyek abban a helyzetben előfordulhatnak. Ha ez egyetlen, adott szóalak, akkor a toldalékolt felszíni alak szerepel, ha a toldalék tetszőleges lehet, de a lexéma nem, akkor a lexéma szintjén specifikáljuk a szerkezeti csomópontot. Másik végletként a szófaji kategóriát alkalmazzuk.

A két véglet (egyedi szóalak, szófaj) között található lexikai csoportok specifikálása különösen érdekes lehet. Akár szemantikai jegyekkel illetve diszjunktív listával, akár véges állapotú automatával határozzuk is meg a minta nyíltvégű elemeit, azok általában intuitíve egy természetes szemantikai osztályt alkotnak. Ilyenek az 1) – 3) példákban is szereplő nemzetnevek és számok. Hasonló szerepet játszhatnak színnevek, testrészek, lásd 4) és 5) példáit.

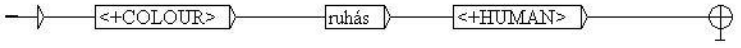
4a) the lady in black
4b) A fekete ruhás hölgy

5a) baby-/poker-/hard-faced boy
5b) baba-/póker-/kemény arcú fiú

Például a 4a) és 4b) szerkezetét informálisan meghatározhatjuk a 4. és 5. ábrán látható módon, ami közvetlen implementálható is a Nooj rendszerben (URL: `perso.wanadoo.fr/rosavram`) Ez a megvalósítás feltételezi, hogy a lexikonban szemantikai jegyeket használunk a színeket valamint a személyeket jelölő szavakra.



4. ábra Lokális grammatika a 4a) kifejezésre szemantikai jegyek használatával



5. ábra Lokális grammatika a 4b) kifejezésre szemantikai jegyek használatával

A 4. és 5. ábrán látható szemantikai jegyek használata csak az első lépcső a lexikon kiépítésében. Egy magasabb szervezettségű lexikonban a szavak megfelelő csoportját nem egyedi szemantikai jegyek vagy azok halmaza jelöli ki, hanem a jegyek típushierarchiába szervezett szerkezete, amely öröklődést is lehetővé tesz.

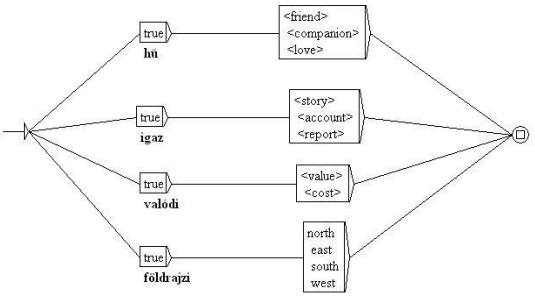
Az eddigi példák szándékosan egyszerűek voltak, de az az állításunk, hogy az általuk illusztrált korlátozottan nyitott szerkezetű kifejezések korántsem marginális szerepűek a nyelv lexikai szerkezetében és szerkezetük sem olyan triviális, mint az eddig említett példák. Tekintsük például a 7. ábrán szereplő lokális grammatikát, amely az időkifejezések szerkezetének legfelső szintjét mutatja be. A 8. ábra az „oraperc" nevű algráf egy részét tartalmazza[69].

3. A lokális grammatikák alkalmazásai

Végezetül a fent bemutatott, korlátozottan produktív kifejezések lokális grammatikáinak három nyelvtechnológiai alkalmazását mutatjuk be.

3.1 Szemantikai egyértelműsítés

A lokális grammatika kiváló eszközül szolgálhat egy-egy kifejezésben szereplő lexikai egység egyértelműsítésére, amit a 6. ábrán szereplő transzdúcer példáz. Az angol „true" szónak az adott kontextusban érvényes magyar megfelelőit a gráf kimenete szolgáltatja, melyet a csomópont alatt találunk.

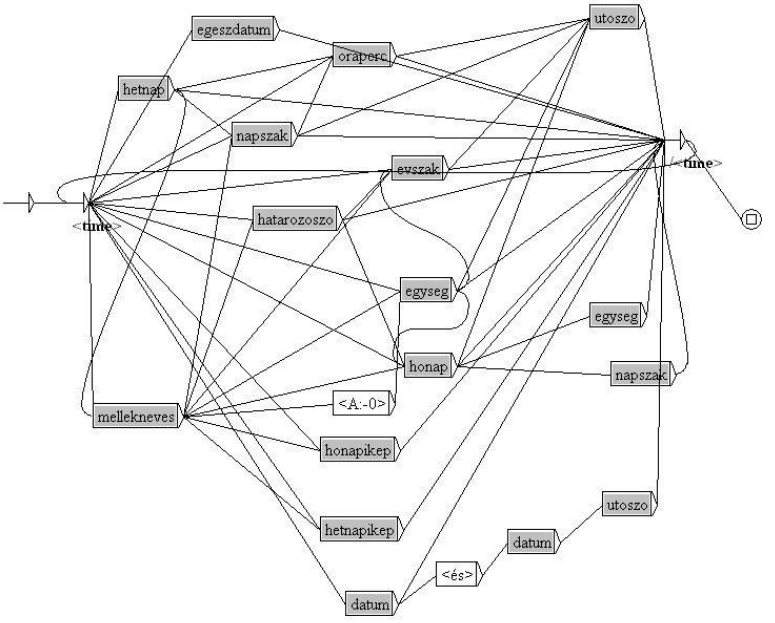


6. ábra Szemantikai egyértelműsítés lokális grammatikával

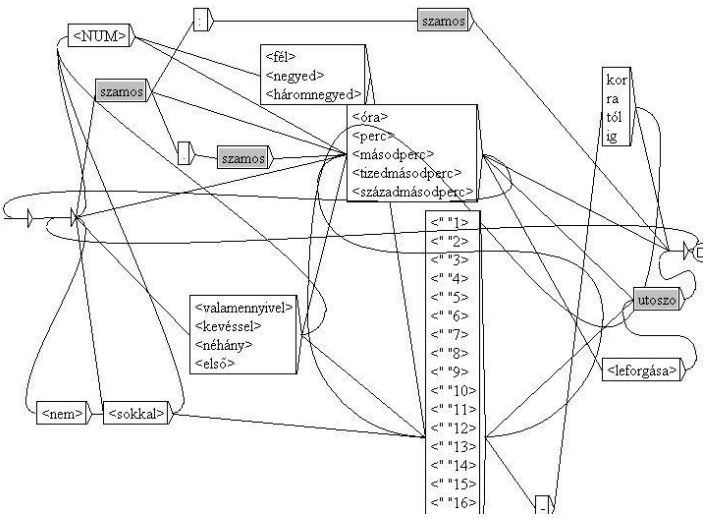
[69] A gráfokat Gábor Kata készítette munkatársaival az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztályán.

3.2 Részleges felszíni gépi fordítás

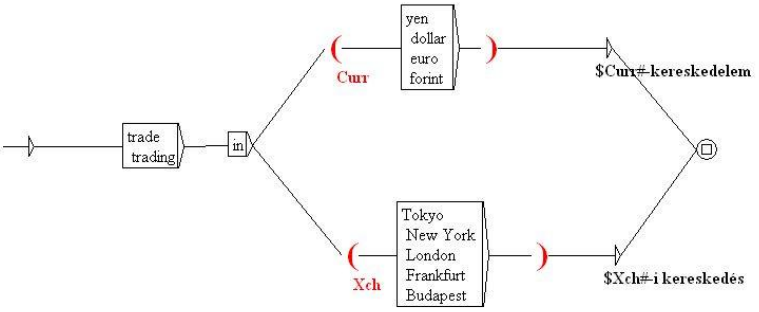
A felszíni elemzés (shallow parsing) mintájára szerkeszthetünk olyan transzdúcert is, amely a felszíni forrásnyelvi minta kimeneteként annak célnyelvi megfelelőjét adja. Erre látunk példát a 9. ábrán, ahol egy olyan grammatika látható, melynek segítségével a „trade in dollar” kifejezéseket a megfelelő „dollárkereskedelem” kifejezéssel, a „trade in London” stb. kifejezéseket pedig a „londoni kereskedés” megfelelővel fordíthatjuk le.



7. ábra Időkifejezések lokális grammatikája



8. ábra Részlet a 7. ábrán szereplő „oraperc" algráfból

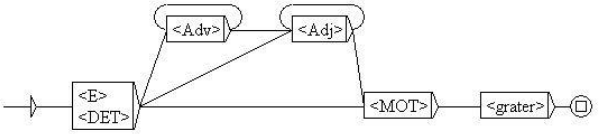


9. ábra Mintaillesztéses fordítás változókat használó transzdúcer segítségével

3.3 Automatikus lexikonépítés

A lokális grammatikák nemcsak elemzésre alkalmasak, hanem arra is, hogy segítségükkel automatikus bővítsük a szótárat hasonló kifejezésekkel. Ha ugyanis a kifejezés lokális grammatikájában valamelyik csomópont meghatározását kellően tágra vesszük és azt nagy korpuszra alkalmazzuk, eredményként olyan felszíni kifejezéseket kapunk, amelyek megfelelnek a grammatikának. Az így begyűjtött kifejezések halmazát elemezve pedig esetleg megtalálhatjuk azt a szemantikai jegyet, amely az adott csomópontra illeszkedő szavak sajátja. Példaként tekintsük a 10. ábrán található gráfot, amely az angol „grater""szóval alkotott jelzős szerkezetek korpuszból való kinyerésére szolgál. A <MOT> szimbólum tetszőleges szóalakra illeszkedik, az <E>

az üres elemet jelöli. Az ábrán a <DET> szimbólummal diszjunktív kapcsolatban szerepel, ami opcionálissá teszi a csomópontot a grammatikában.



10. ábra Automatikus lexikai elsajátítás lokális grammatikával

4. Összefoglalás

A jelen dolgozatban a többtagú kifejezések egy kevéssé vizsgált fajtájára hívtuk fel a figyelmet, amelynek meghatározó jellemzője, hogy részlegesen produktívak, azaz szerkezetük elemei között szerepelnek olyanok, amelyek egy többé-kevésbé nyitott szóosztállyal jellemezhetők. A számítógépes alkalmazások, mint például a gépi fordítás lexikonja számára az ilyen kifejezések explicit jellemzése szükséges, nem elégedhetünk meg a hagyományos lexikográfia utalásos, példálódzós módszerével. Ez teszi szükségessé az ilyen részleges nyitott kifejezések lokális grammatikák segítségével történő kezelését, melyre több példát említettünk.

A kifejezések elemei között olyan finom szintaktikai, de túlnyomórészt szemantikai kombinációs megszorítások vannak, amelyekre a szófaji kategóriák túl durvának bizonyulnak. A természetes szemantikai osztályt alkotó szavak (mint pl. színnév, nemzetnév, testrész, szám stb.) csoportját a lexikonban definiált hierarchikusan szervezett, öröklődő jegyek segítségével lehet legjobban megragadni.

A dolgozat végén bemutattunk egy komplex példát az időkifejezések grammatikájából valamint a lokális grammatikák alkalmazását szemantikai egyértelműsítére, mintaillesztéses felszíni fordításra valamint automatikus lexikonfejlesztésre.

Bibliográfia

1. Bolinger, D. (1965). "The Atomization of Meaning." Language **41**: 555-573.
2. Gross, M. (1997). The Construction of Local Grammars. in Y. S. Emmanuel Roche (szerk.) Finite State Language Processing. MIT Press**:** 329-352.
3. Oravecz, Cs. et al. 2004 Többszavas kifejezések számítógépes kezelése in Alexin, Z., Csendes, D. (szerk.) II. Magyar Számítógépes Konferencia. Szeged. SZTE Informatikai Tanszékcsoport: 141-150

4. Sag, I. et al. 2002 Multiword Expressions: A Pain in the Neck for NLP. in Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics CICLING 2002): 1--15,
5. Silberztein, M. (1993). Dictionnaires électorniques et analyse automatique de textes: le systeme INTEX. Paris, Masson.
6. Silberztein, M. (1999). "Text Indexation with INTEX." Computers and the Humanities **33**(3): 265-280.
7. Sinclair, J., (szerk.) (2004). Collins Cobuild Advanced Advanced Learner's English Dictionary. Glasgow. HarperCollins Publishers
8. Summers, D. (szerk.) (2003). Longman Dictionary of Contemporary English. Harlow. Pearson Education Ltd.
9. Váradi, T., Gábor K., (2004) A magyar INTEX fejlesztésről in Alexin, Z., Csendes, D. (szerk.) II. Magyar Számítógépes Konferencia. Szeged. SZTE Informatikai Tanszékcsoport: 3-10
10. Wehmeier, S., (szerk.) (2005). Oxford Advanced Learner's Dictionary. Oxford, Oxford University Press.

Vonzatok és szabad határozók szabályalapú kezelése

Gábor Kata[1], Héja Enikő[1]
[1] MTA Nyelvtudományi Intézet, Korpusznyelvészeti osztály, Postafiók 701/518, H-1399 Budapest, Magyarország

{gkata,eheja}@nytud.hu

Kivonat: A cikkben bemutatjuk egy szintaktikai szabályrendszer kidolgozásának módszertanát, melynek segítségével elkülöníthetők a vonzatok és a szabad határozók, valamint megfogalmazhatók a szabad határozók mondatba illesztését végző szabályok.

1 Bevezetés

A cikkben tárgyalt munka célja egy olyan szabályrendszer kialakítása, mely az automatikus szintaktikai elemzés során az esetragos főnévi csoportokat funkciójuknak megfelelő annotációval látja el. A szabályrendszer az Intex/NooJ szövegfeldolgozó eszköz [7] magyar moduljába [8] épül. Az Intex, illetve legújabb változata, a NooJ magyar szintaktikai modulja részleges elemzést végez: felismeri a tagmondatok legfelsőbb szintű összetevőit, és meghatározza a köztük lévő szintaktikai függőségi viszonyokat. Ennek megvalósításához szükség van arra, hogy meg tudjuk különböztetni az igei vonzatkeret részeként előforduló főnévi csoportokat a szabad határozói (adjunktum) funkciót betöltő NPktől. Ebben a cikkben kizárólag az ige valamilyen típusú bővítményeként[70] előforduló legfelsőbb szintű NPk funkciójának meghatározásával foglalkozunk..

Ellentétben azzal az elterjedt vélekedéssel, mely az argumentumszerkezetet és az ige bővíthetőségét annak lexikai tulajdonságaként kezeli, mi a szintaktikai relációkat megjelenítő esetragokat választottuk kiindulópontul. Az esetragoknak saját funkciót tulajdonítunk, és az esetragos NPk szintaktikai szerepét az esetrag funkcióiból kívánjuk levezetni. A megközelítésünk azon a gondolaton alapul, hogy a vonzatszerep és a szabad határozói szerep közti különbség valójában a bővítményt tartalmazó szerkezet kompozícionalitásának és produktivitásának mértékében rejlik. Míg a vonzatok előfordulásai azokra a tagmondatokra korlátozódnak, melyek a vonzatot előíró igét tartalmazzák, az adjunktumok ugyanazt a szerepet igék egy tágabb csoportja, egyes esetekben valamennyi ige mellett betölthetik. Eszerint a gondolatmenet szerint a vonzatság egy az esetragok lehetséges funkciói közül, mely abból a szempontból speciális, hogy csak nagyon megszabott környezetben tölthetik be. Az esetragok funkcióit szabályokkal írjuk le, melyek a főnévi csoporthoz nyelvtani szerepet ren-

[70] A "bővítmény" szót összefoglaló névként használjuk a vonzatokra és a szabad határozókra.

delnek. Az esetragos főnevek azon előfordulásai, melyek leírhatók anélkül, hogy a szabály bemenetében hivatkoznánk a főnévi csoport tagmondatában előforduló ige lemmájára, adjunktum szerepűnek tekinthetők.
A szabályok által kiosztott címkék szemantikai tartalommal rendelkeznek, de magukat a szabályokat szintaktikai műveletként fogjuk fel, melyek lehetővé teszik az adott NP használatát az aktuális kontextusban.
Cikkünkben ismertetjük a módszert, melynek segítségével elkülönítjük az esetragok funkcióit, valamint leírjuk és az automatikus szintaktikai elemzésben megvalósítjuk az adjunkciós szabályok rendszerét. Munkánk eredménye egy kritériumrendszer, melynek segítségével a vonzatok elkülöníthetőek a produktívan használt, kompozícionális szerkezeteket alkotó adjunktumoktól.
A következőkben bemutatjuk a magyar szintaxis néhány fontos jellemzőjét, valamint leírjuk, miért nehéz a magyar nyelvre alkalmazható vonzattesztet találni [2], majd bemutatjuk saját módszerünket [3]. Ezután egy konkrét példát ismertetünk [4], végül bemutatjuk a kiértékelés eredményét [5].

2 Vonzatok és szabad határozók a magyar mondatban

A magyar mondatban az összetevők felszíni szórendje nem tükrözi szintaktikai szerepüket. A komplementumok és adjunktumok csaknem bármilyen felszíni sorrendje elfogadható, bár egyes szerkezeti pozíciók különböző diskurzus-funkcióknak feleltethetők meg [1]. Fókuszálás vagy topikalizáció által bármilyen funkciójú összetevő az ige elé mozgatható. Ezen kívül az igemódosítók (névelőtlen NPk, igekötők, adverbiumok) is megelőzhetik az igét. Magyar szövegek automatikus elemzésekor azzal kell tehát szembesülnünk, hogy a felszíni sorrendet legfeljebb a diskurzus-funkciók feltérképezésére használhatjuk, a szintaktikai függőségi viszonyokat azonban nem határozhatjuk meg a sorrend alapján. Mivel a felszíni sorrendet nem használhatjuk, egyéb konfigurációs információ pedig nem áll rendelkezésünkre, nem használhatjuk a transzformációs, konfigurációs nyelvtanok tesztjeit a vonzatok és adjunktumok elkülönítésére. Ezt támasztja alá az is, hogy Radford [6] (angol nyelvre vonatkozó) tesztjei közül egyik sem használható a magyarra:

a) Passziválás: A vonzatszerepű PP-ből kiemelt NP passziválható, az adjunktum PP főnévi csoportja nem:

[This job] needs to be worked at by an expert.
*[This office] is worked at by a lot of people.

b) Pronominalizálás: A *'do so'* szerkezet, ami a V' kategóriát helyettesíti, tartalmazhat újabb V'-t képző adjunktumot (i), de el is hagyható (ii), míg a komplementum kötelezően benne foglaltatik a V'-ban (iii), nem hagyható el (iv).

1. John will [buy the book on Tuesday] and Paul will *do so* as well.
2. John will [buy the book] on Tuesday and Paul will *do so* on Thursday.
3. John will [put the book on the table] and Paul will *do so* as well.

4. *John will [put the book] on the table and Paul will *do so* on the chair.

c) Felszíni sorrend: A vonzatok közelebb vannak az igéhez, mint az adjunktumok, mert a szintaktikai fában előbb csatlakoznak hozzá, és a fa élei nem keresztezhetik egymást.

d) Ellipszis: Bármilyen frazális kategória elliptálható. V' kategóriájú összetevő akkor elliptálható, ha tartalmazza az igei fejet vonzataival és adjunktumaival (i), a fejet vonzataival, adjunktumok nélkül (ii), de a fej egyik vonzatával, de a másik nélkül nem alkot elliptálható összetevőt (iii):

i) – Who might be going to the cinema on Tuesday?
– John might be ____ .
ii) – Who might be going to the cinema when?
– John might be ____ on Tuesday.
iii) – Who will put the book where?
**John will ____ on the table.*

Az a) és b) tesztek azért nem alkalmazhatók, mert a magyarban nincs hasonló passziválás, illetve pronominalizálás. A c) teszt nem teljesül a magyar mondatok egy részében:

A gyerekek nyírják a kertben a füvet.

A fenti mondatban a szabad határozó az ige és vonzata közé ékelődik, tehát az ige nem közvetlenül szomszédos a vonzataival.

Ha feltételezzük, hogy a magyar '*tesz*' ige az angolhoz hasonlóan három vonzattal (alany, tárgy, lokatívusz) rendelkezik, megmutathatjuk, hogy a d) teszt sem használható a magyarra:

– *Ki megy hová kedden? – János __ moziba __.*
– *Ki tette a könyvet hová? – János __ __ az asztalra.*

A második mondatpár, melyben a *tesz* ige két vonzata szerepel, és a harmadik az igével együtt elliptálódik, megmutatja, hogy a vonzatok nem csak közösen hagyhatók el. Azaz egyik teszt sem alkalmazható a magyar nyelvre.

Komlósy [3] a magyar igei argumentumszerkezetről írott tanulmányában azt állítja, hogy a vonzatok és szabad határozók elkülönítése olyan művelet, melynek elvégzéséhez a nyelv egész nyelvtanának ismeretére szükség van. A vonzatot olyan összetevőként határozza meg, melynek szintaktikai és szemantikai tulajdonságait az őt kormányzó ige írja elő. Három tesztet javasol a szerkezetek elkülönítésére, melyek, amint a szerző is vállalja, nem elégségesek az összes szerkezet meghatározásához:

e) ha egy bővítmény kötelező, akkor vonzat;

f) ha egy opcionális bővítmény kitétele lehetővé teszi a szerkezet kibővítését egy másik bővítménnyel, mely mellől az első bővítmény már nem hagyható el, akkor az első bővítmény vonzat;

g) ha X szónak bővítménye Y, és van olyan Z szó, ami szisztematikusan helyettesítheti X+Y szerkezetet, valamint helyettesítheti X -et, amikor Y nincs jelen, de nem helyettesítheti X -et, ha Y jelen van, akkor Y X opcionális vonzata.

A kötelezőségre hivatkozó e) teszt használatát mindenképp mellőzni szeretnénk, mivel ellipszis vagy egyéb műveletek által gyakorlatilag bármit elhagyhatunk a mondatból, és nehézségekbe ütközhet annak eldöntése, hogy egy adott mondat tartalmaze kötelező, de elhagyott összetevőt. Az f) és g) tesztek megbízhatóságát nem vitatjuk,

ám használhatóságuk korlátozott, így mindenképp szükségünk van más kritériumokra is.

Bár a GB vonzattesztjeit elvetettük, az összetevők szintaktikai szerepeinek azonosítására használt koordinációs teszt a mi munkánkban is fontos szerepet játszik. Ha feltételezzük, hogy csak azonos szerepű összetevők koordinálhatók, a vonzat – adjunktum ellentét segítségével kell számot adnunk az alábbi mondatról:

**János beszennyezte a szőnyeget sárral és a cipőjével.*

A *sár* és a *cipő* tehát különböző funkciót lát el ebben a mondatban, bár szemantikailag mindkettő a beszennyezés eszközének tekinthető.

Azt feltételezzük, hogy a fentihez hasonló mondatokban más szabály kapcsolja az igéhez a nem koordinálható, azonos esetragot viselő NPket, és ezek a szabályok az eltérő szintaktikai szerep mellett különböző szemantikai címkét is társítanak a főnévi csoporthoz. Az alábbi mondatban:

Párizsban még bíztam az apámban.

a két, azonos esetragú NP koordináció nélkül szerepel együtt, és nem is koordinálhatók. Ennek oka, hogy az egyik NP *('az apámban')* vonzat, míg a másik szabad határozó. A mondat szerkezetileg kétértelmű, de teljesen kizárja az olyan értelmezéseket, melyben a két NP azonos szerepet tölthetne be. Egy esetrag ugyanis csak egyszer jelölhet egy funkciót egy tagmondatban. Kérdés azonban, hogy a vonzat-adjunktum szembeállítással hogyan magyarázzuk meg az alábbi mondat helyességét?

2005-ben Párizsban még bíztam az apámban.

Ebben a mondatban két adjunktumszerepű NP-t találunk, melyek koordináció nélkül is jólformált szerkezetet alkotnak. Ezt a jelenséget úgy próbáljuk magyarázni, hogy a vonzat – szabad határozó szembeállítás helyett egy többfokozatú skálát alkalmazunk, melyben minden adjunkciós szabály különböző funkcióért felelős.

3 Kompozícionalitás és produktivitás

A NooJ magyar moduljában véghezvitt szintaktikai elemzés célja, hogy a szöveg valamennyi legfelsőbb szintű esetragos főnévi csoportját nyelvtani szerepe szerint annotálja. Ehhez a tagmondatok konfigurációs szerkezete helyett az esetragok szintaktikai szerepjelölő funkcióját kívánjuk használni. A dependencia-nyelvtan terminológiájával élve az esetragos NP szerepét a predikátumhoz való viszonyában határozzuk meg. Mindazonáltal a szerepek leírásakor kerülni akarjuk a predikátum tulajdonságaira való hivatkozást, és minél több esetrag-funkciót szeretnénk általános szabályokkal megragadni. A predikátum-vonzat relációt is az esetragok egyik funkciójának tekintjük.

Esetragnak azt a toldalékot tekintjük, ami a magyar főnév jobb szélén jelenik meg, másik toldalék nem követheti, és egy főnévnek csak egy esetragja lehet. Ezek alapján a magyarban 19 esetragot sorolhatunk fel. Feladatunk, hogy valamennyi esetrag lehetséges funkcióit szabályokkal leírjuk, szabályokkal nem kezelhető szerkezeteket pedig vonzatként felsoroljuk.

Elsőként megállapíthatjuk, hogy az alany- és a tárgyeset nem rendelkezik default jelentéssel: minden előfordulásukban az igei argumentumszerkezetet részét képezik[71].

[71]Természetesen ez nem vonatkozik a névutós frázisokban, illetve a főnevek vagy melléknevek vonzataként előforduló, nem legfelsőbb szintű NPkre.

A többi esetragról azt feltételezzük, hogy rendelkeznek saját szintaktikai és szemantikai tulajdonságokkal, melyek szabályokkal leírhatók. Ezeket az általános szabályokat, melyek az esetragok alapértelmezett funkcióját/funkcióit definiálják, default szabályoknak nevezzük. A default szabályok bemenete utalhat az őt tartalmazó főnévi csoport fejének szemantikai vagy morfoszintaktikai tulajdonságaira, de soha nem utalhat annak az igének a lemmájára, amelyik az esetragos főnévi csoport tagmondatának állítmánya. Ennek értelmében egy esetragnak egynél több default funkciója is lehet, bár a funkciókat leíró szabályok közül szigorúan véve csak egy szabály valódi „default", ami az esetrag összes olyan előfordulását lefedi, melyre a többi szabály nem illeszkedik. A szabályok kimenete a főnévi csoport szerepét leíró címke. Annyiféle nem vonzat szerepet különböztetünk meg esetragonként, ahány szabályt használunk az egyes esetrag funkcióinak leírásához (különböző esetragok funkciói viszont egybeeshetnek). Mivel a szerepeket az NPhez társító szabályokat szintaktikai (adjunkciós) szabálynak fogjuk fel, a szabályok kimenetében megjelenő szerepcímkék is szintaktikainak tekinthetők. Itt azonban fontos megjegyezni, hogy a szerepek erős szemantikai tartalommal bírnak, valamint a szabályok jellegéből is kiderül, hogy egyes adjunkciós műveletek szemantikailag megszorított bemeneten működnek. Ezek alapján úgy tekintettük, hogy a szintaxist és a szemantikát nem kezelhetjük külön modulban.

Például a *-ban* esetrag alapértelmezett jelentése attól függ, hogy milyen szemantikai jegyekkel rendelkező NPn jelenik meg: az időt kifejező főnévi csoportnak időhatározói szerepet ad (*'januárban találkozunk'*), míg egyéb esetekben szabályos helyhatározói funkciót ad az NPnek (*'a hordóban találtam'*). A szabályok, melyek az alapértelmezett szerepeket osztják ki, természetesen a vonzatfunkciók megállapítása után futnak le, mivel bemenetük kevésbé specifikus: a kontextustól függetlenül működnek.

Azok az [ige + NP + esetrag] szerkezetek, melyek nem írhatók le általános szabályokkal, [ige + vonzat] szerkezetként elemzendők. Azért nem rendelhető hozzájuk default szabály, mert ezek a szerkezetek nem kompozícionálisak: az NP igéhez képesti szerepét nem lehet olyan szemantikai címkével ellátni, mely nem utal az ige jelentésére. Például:

A közönség elhalmozta az előadót kérdésekkel.

Ha a fenti mondatban az [ige + NP + -val] szerkezet kompozícionális lenne, az NPhez tudnánk olyan absztrakt címkét társítani (pl. hely, idő, mód stb.), ami leírja az igéhez való viszonyát anélkül, hogy az ige jelentésére bármilyen módon hivatkozna. Hogy ez nem lehetséges, az abból is látszik, hogy természetes nyelven sem találunk hozzá olyan parafrázist, amely kifejezi az ige és az NP viszonyát, de nem tartalmazza sem az igét, sem annak szinonimáját.

Ezek alapján a 4) mondat az ige + vonzat szerkezet példájának tekinthető – tehát az *elhalmoz* ige lexikai tételébe fel kell vennünk .

Vannak azonban olyan esetragos szerkezetek is, melyek köztes kategóriát képviselnek az adjunkció teljes produktivitása és a vonzatság teljes lexikalitása között. Az esetragok ezen használatai csak egyes *szemantikai* igeosztályok mellett mondhatók produktívnak. Például a *-tól* esetragnak ilyen módon megkülönböztethetjük két funkcióját: a mozgást jelentő igék mellett a mozgás kiindulópontját jelentő NPn jelenik meg, míg állapotváltozást jelentő igék mellett az állapotváltozás közvetlen okát jelentő NPt azonosítja. Az esetragnak ezt a két funkcióját két szabállyal tudjuk leírni, melyek közül mindkettő utal a tagmondat állítmányának szemantikai osztályára. Azt

állítottuk, hogy az adjunkció teljesen produktív művelet, mely a tagmondat állítmányától függetlenül alkalmazható (az egyetlen követelmény, hogy a tagmondatnak *legyen* állítmánya), míg a vonzatság az egyedi igei lemmák lexikális tulajdonságától függ. A szemantikai igeosztályokon működő műveleteket leíró *nem-default* szabályok kevésbé produktívak, mint a default-szabályok, így besorolásuk nem egyértelmű. Mindazonáltal érdekünkben áll, hogy ne tekintsük vonzatnak a nem-default szabályok által létrehozott szerkezeteket, mert így a bennük szereplő NPk szerepéről több információt tudunk adni, mintha csak vonzat-státuszukra hivatkoznánk. Emellett egy szintaktikai teszt is alátámasztja, hogy a főnévi csoportok funkciói szélesebb körűek a vonzat-adjunktum kettősnél. Ha feltesszük, hogy egy jólformált magyar tagmondat nem tartalmazhat kettő vagy több olyan NPt, melyek ugyanazt az esetragot viselik, ugyanazt a szerepet töltik be, és nincsenek koordinálva, akkor problémát okoz a 3) mondat helyessége, amelyben két *-ban* esetragos adjunktum szerepű NP van (az ugyanolyan esetragos vonzat mellett). Mivel mi azt feltételezzük, hogy az esetragos NP annyiféle szerepet tölthet be, ahányféle szabály alkalmazható az esetragra, vagyis minden szabály kimenete különböző címkével látja el az NPt, egyszerűen megfogalmazhatjuk a jólformáltsági feltételt: minden szabályunknak csak egy találata lehet tagmondatonként (a találat azonban koordinált NPt is tartalmazhat), így minden NP különböző szerepet kap.

4 A szemantikai igeosztályok meghatározása

Miután végigvettük azokat az általános vezérelveket, amelyek kutatásunk alapjául szolgálnak, egy konkrét példa részletesebb tárgyalásával folytatjuk. Az alábbiakban a *–val* esetrag előfordulásait vizsgáljuk meg. Azt feltételeztük, hogy a szóban forgó esetraghoz két default szabály tartozik, vagyis két olyan szabály, amely anélkül határozza meg a megfelelő főnévi csoportok mondatban betöltött szemantikai szerepét, hogy bármilyen formában is hivatkozna a predikátumra. (Mint később látni fogjuk ez a feltételezésünk nem igazolódott.)

Ezek közül az egyik a default társhatározói szabály, amely ASSOCIATE nevű címkével látja el a releváns főnévi csoportokat. A szóban forgó főnévi csoportok közös jellemzője, hogy a predikátum által jelölt eseményben betöltött szemantikai szerepük az alany szerepével egyezik meg (‘*ül*’). A szabály akkor alkalmazódik, ha a megfelelő főnévi csoport rendelkezik a +HUMAN szemantikai jeggyel.

János Marival ül a kertben.

A másik szabályunk a default eszköz szabály. Ez azokat az NPket jelöli meg, amelyek az ige által jelölt eseményt végrehajtására szolgáló eszközre referálnak.

János kocsijával hazavitt mindenkit.

Fontos kiemelni, hogy a fenti szabályok esetében nem hivatkoztunk a predikátumokra, legfeljebb a kérdéses NPk szemantikai vagy szintaktikai jegyeire. Ez összhangban van azzal a hagyományos nézettel, hogy az adjunktumok szinte bármilyen ige mellett megjelenhetnek, vagyis az adjunktumok jelentése független az igéétől.

A default-szabályok alkalmazásának megvan az az előnye, hogy ezáltal a szövegben szereplő minden megfelelő esetraggal rendelkező főnévi csoporthoz rendelünk szemantikai szerepet, így a lefedettség az ige felismerésétől függetlenül 100% lesz.

A tesztelés során azonban kiderült, hogy kezdeti feltevésünk nem volt helyes, amennyiben nemcsak default eszköz és default társhatározó szabályokat kell létrehoznunk. Ennek az az oka, hogy a *–val* esetragos főnevek egy további meglehetősen produktív használata a határozói használat. Az alábbiakban erre láthatunk példát:

Mari csökönyös és áhítatos erőszakkal ragaszkodik Bélához.

Bizonyos esetekben azonban meglehetősen problémás az ige által jelölt cselekvés módjára vonatkozó adverbiális és az eszközhatározói szerepet betöltő főnévi csoportok elkülönítése.

A gyermekem már késsel és villával eszik.

A probléma megoldására első lépésként létrehoztunk egy szabályt, amely azon az előfeltevésen alapul, hogy a melléknevekből és igékből képzett főnevek képesek betölteni az igemódosító pozíciót és gyakrabban is kerülnek ebbe a pozícióba, mint főnévibe. Így a harmadik default szabályunk bemenetét az –Ás illetve –sÁg végű főnevek alkotják. További szabályokat is létrehoztunk az adverbiális és eszköz típusú NP-k elkülönítésére. A szabályok mögött az a megfigyelés húzódott meg, hogy a cselekvés módjának és eszközének szétválasztása akkor igazán problematikus, amikor az ige nem egy konkrét eseményre, hanem egy esemény típusra referál. Ezt illusztrálja a fenti mondatpár második mondata. Ilyenkor a mondatban megnevezett eszköz nem egy konkrét eszköz lesz: inkább azt a módot jelöli, ahogyan az eseményt végre szoktuk hajtani, vagy ahogyan az esemény általában végbemegy. Ezért a fent említett default eszköz szabály esetében figyelembe vettük, hogy az NP rendelkezik-e névelővel. A névelős, vagyis határozott főnévi csoportokat eszközként jelölték meg szabályaink, míg a névelőtleneket módként.

A *–val* esetragos főnévi csoportok vizsgálata során még egy default szabályt alkalmaztunk , amely a bemeneti szöveget a MEASURE címkével láthatja el. Ez a szabály szintén támaszkodik szemantikai jegyekre, a szóban forgó főnévi csoportok a szabály alkalmazásakor már rendelkeznek a TIME és MEASURE szemantikai jegyekkel. Az ilyen jegyű *–val* esetragos NPk feltételezésünk szerint az ige által jelölt esemény (változás) mértékét vagy két esemény között eltelt időt fejezik ki. *'[MEASURE Húsz évvel] ezelőtt'* vagy *'[MEASURE Három százalékkal] nőtt.'*

A default-szabályok kialakítása során nem hivatkozhatunk az igéhez kapcsolódó szisztematikus morfológiai és szintaktikai változásokra, mivel ezzel ellentmondásba kerülnénk a default szabályok definíciójával, amely semmilyen formában nem engedi meg az igére való hivatkozást.

Az alábbiakban a nem default szabályokat tekintjük át. Ebbe a szabályosztályba is felvettünk egy INSTRUMENTUM szabályt. Emlékeztetőül: ezek a szabályok azért nem default szabályok, mert a főnévi csoport egyes tulajdonságain kívül az igére is hivatkoznak. Ilyen predikátum például a már fent is említett '*beszennyez*', ahol az igének van egy eszköz típusú argumentumhelye. A koordinációs teszt segítségével megmutattuk, hogy a '*beszennyez sárral*' és a '*beszennyez a cipővel*' két különböző argumentum. Kérdésként merülhet fel azonban , hogy miért van szükség a default és nem-default eszköz szabályok megkülönböztetésére. Egyfelől láttuk, hogy szintaktikailag motivált az elkülönítésük. Másfelől egy nem-default szabály illeszkedése egy sztringre megakadályozza a default szabályok alkalmazását. Ez nyilvánvalóan akkor fontos, ha olyan default szabályok illeszkednének rá, amelyek más szemantikai szerepet tulajdonítanának az NPnek. Esetünkben pontosan ez lenne a helyzet, hiszen az ilyen NPk a fent leírt MODE szabály bemenetét képeznék.

Áttérve a nem-default társhatározói szabályra, szintén meg kell válaszolnunk a fenti kérdést. Míg a default szabály hivatkozott a releváns NP +HUMAN szemantikai jegyére, a default szabály nem használja fel ezt az információt. Ez a megkülönböztetés azt a tényt tükrözi, hogy létezik egy olyan igeosztály, amely esetében a *–val* esetragos főnév mindig társként viselkedik, vagyis szemantikai szerepe mindig megegyezik a mondat alanyának szemantikai szerepével. Ezt mutatja az alábbi mondat is.

2) *János veszekszik az autóval.*

A fenti mondat – 3)-mal szemben – csak úgy értelmezhető, hogy az *autó* is részese – és nem eszköze – volt a veszekedési eseménynek., annak ellenére, hogy az *autó* nem rendelkezik a +HUMAN jeggyel.

3) *János Marival ment moziba.*

A fenti mondatban *Mari* csak abban az esetben jelölhet társat, ha ember. Egyébként eszköz lenne.

A következő szabály az állapotváltozást kiváltó közvetlen okokat jelölő főnévi csoportokat látja el címkével. Akárcsak a többi nem-default szabály, ez is hivatkozik arra, hogy a főneves kifejezés környezetében található ige melyik igeosztályba tartozik. Idetartoznak például a '*megdöbbent*', '*felidegesít*', '*megrémít*' igék.

4.a) *János megdöbbentette Marit a hírrel.*

A fenti mondatban szereplő ige nézőpontunkból lényeges szemantikai tulajdonságait az alábbi kifejezés szemlélteti:

5) CAUSE(János, E), ahol E<hír, CHANGE(S, S')> és CAUSE(hír, S')

Eszerint János létrehozott (CAUSE) egy szituációt (E), ahol a szituációt egy olyan kétargumentumú predikátummal írhatjuk le, amelynek az első argumentuma (*hír*) állapotváltozást okoz (CAUSE) *Mari* mentális állapotában, vagyis átmenetet idéz elő S-ből S'-be. A következő felmerülő kérdés, hogy hogyan igazolhatnánk szintaktikailag a három metapredikátum jogosságát (i.e. CAUSE, MENTAL, CHANGE)?

Az alábbi tesztet használtuk annak eldöntésére, hogy egy adott ige tagja-e ennek az osztálynak:

4.b) *A hír megdöbbentette Marit.*

4.c) *Mari megdöbbent a hírtől.*

Feltételeztük, hogy egy ige akkor és csak akkor tartozik ebbe az osztályba, ha a 4.a.), 4.b) és 4.c) példamondatok szerkezetével egyaránt jól formált mondatot alkot. 4.a) és 4.b) alapján azt állíthatjuk, hogy az ilyen típusú igéknek rendelkezniük kell legalább egy olyan olvasattal, ahol az alany nem ágens. Ha ez nem teljesülne 4.b) agrammatikus lenne, hiszen az alany jelölete ebben az esetben nem képes egy cselekedet szándékos végrehajtására. Ebből következik, hogy a csoportba tartozó legtöbb ige – bár korántsem az összes – mentális állapotváltozásra vonatkozik. Azt látjuk tehát, hogy a MENTÁLIS metapredikátumunk ebben az esetben ekvivalens azzal a követelménnyel, hogy az alanynak legyen legalább egy nem ágenses olvasata az ige mellett. 4.c) illusztrálja a CAUSE és CHANGE metapredikátumok szükségességét. Elfogadtuk Komlósy [4] azon nézetét, mely szerint bizonyos igék Okozó szerepű argumentumai megjelenhetnek *–tól* esetragos főnévként. Ezenfelül azok az igék, amelyek mindhárom szerkezetben megjelenhetnek, feltételeznek két állapot közötti átmenetet is, ahol a hangsúly nem magán az átmeneten van, hanem a második állapot elérésén.. Ez a feltételezés párhuzamba állítható azzal a jelenséggel, hogy a szóban forgó igeosztály elemeit a általában perfektív igealakkal fordítjuk angolra. Ennek a jelenségnek az lehet az oka, hogy míg a perfektív igealakok az ige által jelölt ese-

mény bekövetkezése utáni állapotot hangsúlyozzák, az imperfektív igealakok magát a folyamatot. Egy másik érv a CHANGE metapredikátum szükségessége mellett pedig azon alapul, hogy vannak olyan igék, amelyek mellett ugyan megjelenhet *–val* esetragos Okozó szerepű főnévi csoport, így ezek az igék az Okozó metapredikátum alá tartoznak, de nincsen két meghatározott állapot közötti átmenet, így a CHANGE metapredikátum nem alkalmazható. Vegyük példaképp az alábbi mondatokat:

6.a) *Az igazgató Jánost terhelte a feladattal.*
6.b) *A feladat Jánost terhelte.*
6.c) *János terhelve van.*
6.d) **János terhelve van a feladattól.*

A CHANGE metapredikátum szükségességére vonatkozó szemantikai intuíciónkat explicit módon támasztja alá 6.d) helytelensége. Ha a CHANGE metapredikátum által jelölt jelentéskomponens is jelen van az igében, mind a három szerkezet jól formált. A fenti példában szereplő igéből azonban csak ez hiányzik. Ez azt támasztja alá, hogy a CHANGE metapredikátum is disztinktív és éppen ezért a CAUSE-tól függetlenül fel kell vennünk, ha meg szeretnénk adni az adott igeosztályba való tartozás szükséges feltételeit.

Egy másik igeosztályt alkotnak a faktitív igék. Az erre az igeosztályra hivatkozó szabályok egy AGENS2 nevű szemantikai címkét rendelnek hozzá a megfelelő főnévi csoportokhoz. Azért így neveztük el ezt a csoportot, mert a faktitív műveltetés alapigéje mindig ágenses, így a *–val* esetragos főnévi csoport az alapige ágensét fogja jelölni. Mivel a kauzatív igék is képezhetők a *–(t)At* műveltető képzővel és ezek mellett szintén megjelenhetnek *–val* esetragos főnevek (amelyek azonban ebben az esetben nem lehetnek ágensek), hivatkoznunk kell a releváns főnévi csoport szemantikai jegyeire is, azaz ki kell kötnünk, hogy a szabály csak akkor alkalmazódjon, ha az rendelkezik a +HUMAN jeggyel.

7) *János levágatja a haját a fodrásszal.*

Az előbbi illusztrációt használva a fenti példamondatot az alábbiak szerint írhatjuk le:

8) CAUSE(János, E), ahol E<fodrász, haj, ...> és AGENS2(fodrász, E)

Vagyis János létrehoz egy eseményt (E), amelynek legalább két szereplője van – hiszen csak a tranzitív igékből képzett műveltető igék mellett jelenhet meg az eredeti alany *–val* esetragos főnévként – és a fodrász az ágense annak az igének, amely E-t leírja. Következésképpen ezekben az esetekben a predikátum mellett megjelenő *–val* esetragos főnév az alapige ágense.

A fentiekben megmutattuk, hogy a default és nem default szabályok megkülönböztetése empirikus és elméleti nézőpontból egyaránt védhető.

Munkánk jelenlegi állapotában ezekkel az igeosztályokkal rendelkezünk. A többi igét , amely mellett megjelenhet *–val* esetragos főnévi csoport, vonzatként kezeljük. Értelemszerűen az ilyen környezetben megjelenő *–val* esetragos főnévi csoportokat nem tudjuk szemantikai címkével ellátni. A megfelelő igéknek ezt a tulajdonságát kódolni kell az igei szótárban.

5 Implementáció

A munkafolyamat első lépése az igei szókincs kiválasztása volt. A Magyar Nemzeti Szövegtár (MNSz) [9] 2,800 leggyakoribb igéjét választottuk. Ezen az igeosztályon definiáltuk az esetragok default jelentéseit, a szemantikai igeosztályokat, valamint a vonzatkeret kódolását. Az esetragokat aszerint vizsgáltuk, hogy milyen gyakran fordulnak elő ezekkel az igékkel. Négy gyakori esetragot tanulmányoztunk részletesen: *-val* (instrumentális), *-nak* (datívusz), *-tól* (ablatívusz) és *-ra* (szublatívusz). Először az esetragok default jelentéseit határoztuk meg, mivel a nem-default szabályok kidolgozása előfeltételezi a default jelentések ismeretét. A nem-default szabályok kidolgozása úgy zajlott, hogy csoportosítottuk a (nem default jelentésében szereplő) esetraggal előforduló igéket aszerint, hogy az esetragot viselő NP milyen szerepet tölt be mellettük. Ezek a csoportok megadták a szemantikai osztályokat, és várakozásunknak megfelelően nemcsak egy-egy esetrag jelentéseinek elkülönítésében játszottak szerepet. Utolsó lépésként azokat az igéket, amik mellett az esetragos NP szerepét egyik szabály sem fedi le, megjelöltük, mint az adott esetragot vonzatként előíró igét.

A szabályrendszer a NooJ magyar szintaktikai moduljában végrehajtott részleges elemzésre épül. Az elemzés bemenete nem egyértelműsített, text formátumú magyar szöveg, sebessége 240K/perc. A kimenet szintaktikai jellegű annotációt tartalmaz. Az általunk használt morfológiai elemzés szóalakokat felsoroló szótárakra épül: a szótárak az MNSz 900,000 leggyakoribb szóalakját fedik le. A szótári bejegyzés a szóalakot, a hozzá tartozó lemmát és morfológiai kódot, valamint a lemma esetleges szemantikai jegyeit tartalmazza. A morfológiai kód a HUMOR [5] elemzésén alapul.
A nyers szöveg nyelvtani elemzése magában foglalja a tokenizálást, a mondatszegmentálást, lexikai és morfológiai elemzést, többszavas kifejezések és tulajdonnevek felismerését (a tokenizálás részeként), valamint a részleges szintaktikai elemzést. Egyértelműsítést egyáltalán nem használunk, mivel azt a nyelvtani szabályok, elsősorban az NP-nyelvtan nagyrészt elvégzi[72]. A szintaktikai elemzést egymásra épülő, többlépcsős nyelvtanok végzik, melyek a korábbi nyelvtanok kimenetére hivatkoznak. A tagmondat központi elemének a finit igét tartjuk, melyről azt feltételezzük, hogy a legfelsőbb szintű frázisokkal, illetve ezek fejével lép dependencia-viszonyba, vagyis az elemzés alapja a frázisok megtalálása. A frázisok felismerése után a tagmondat állítmányának megtalálása következik. Ezután a hétlépcsős tagmondathatár-nyelvtanunk [2] bejelöli a határokat, amin belül az egyes igék vonzatait és szabad határozóit kereshetjük. Ettől kezdve a szintaktikai elemzés minden lépését a tagmondathatáron belül hajtjuk végre.
A szabályrendszer implementáláshoz szükséges további előfeldolgozás során a főnévi csoportokat és a finit igéket annotáljuk a szabályokban használt releváns tulajdonságaik szerint. Ehhez a NooJ-ban használt szótárakat kibővítettük a főnevek általunk használt szemantikai jegyeivel (pl. *time*, *human*, *measure*). A fej szótári jegyei alapján a felismert NPk annotációját kibővítettük ezekkel a jegyekkel, így a szintaktikai modul kimenet már tartalmazza őket. Az igei predikátumok pedig szintén szótári jegyként kapták meg a szemantikai csoportokat azonosító jegyeket.

A kidolgozott szabályrendszerek közül a -val rag szabályait implementáltuk és teszteltük. A szabályok a szöveg valamennyi főnévi csoportját annotálják szerep

[72]Az NP-nyelvtan fejlesztésekor [10] mért adatok alapján a magyar szövegek tokenjeinek kb. 69%-a valamelyik legfelsőbb szintű NP-be tartozik.

szerint. A szabályok alkalmazásának sorrendjét specifikusságuk foka szabja meg. Három szintet különböztettünk meg: 1) először a legspecifikusabb szabályok, vagyis a vonzatok azonosítását végző lexikális szabályok futnak le, 2) őket követik a predikátumosztályokra alkalmazott nem-default szabályok 3) végül az igére egyáltalán nem referáló default szabályokat alkalmazzuk. A csoportokon belüli alkalmazási sorrend tetszőleges, az egyetlen "minden más esetben" alkalmazandó default szabály kivételével, melynek az utolsónak kell lennie.

6 Értékelés

A kiértékeléshez Méray Tibor *Nagy Imre élete és halála* című művét használtuk. Ez a regény 12,545 mondatból áll, 130,027 szóalakot tartalmaz. A morfológiai elemzés során 1561 féle szóalak maradt ismeretlen.
A szövegben 29855 főnévi csoportot találtunk. Mivel a szabályaink pontosságát szeretnénk ellenőrizni, és a szabályok minden NPt annotálnak szerep szerint, csak a pontossági értékeket számoltuk ki, hiszen az egyes szabályok lefedettségének hiányai valamelyik másik szabály pontosságában is megmutatkoznak. Az értékelés során feltárt hibák – melyek nem az elemzés valamelyik korábbi lépésének hibás kimenetéből erednek – a szabályrendszer javításának módját is kijelölik.
Az eredmények manuális ellenőrzéséhez a Xaira (XML Aware Indexing and Retrieval Architecture) korpuszlekérdező eszközt [11] használtuk. Amint neve is mutatja, a Xaira jólformált XML dokumentumokból álló korpuszok indexálását és komplex lekérdezések megfogalmazását teszi lehetővé. A lekérdezések eredményéhez stíluslapok rendelhetők, így a szabályaink kimenetét emberek számára is könnyen olvashatóvá tudtuk tenni, ami megkönnyítette az eredmények manuális ellenőrzését.

1. Táblázat: Értékelés

Szabályok	Pontosság	Találatok száma
Vonzat	71.50%	179
Műveltető	100.00%	1
Okozás	36.00%	11
Társh. szabály	65.00%	76
Társh. default	61.60%	129
Eszköz (szabály és default)	42.37%	573
Mód	54.76%	168
Default idő/mérték	88.09%	42
Lexikalizált	100.00%	59
Összesen	59.57%	1238

A kiértékelés során azt találtuk, hogy a hibás szemantikai címkék nagy része két forrásból származik. Az egyik, hogy az ige, amelynek lemmájára vagy szemantikai

osztályára hivatkozni kéne nem szerepel a megfelelő listán. A hibás szemantikai szerepek másik fő oka a Mód szabály alkalmazásából adódó alacsony lefedettség. A Mód szemantikai szerepű összetevők ugyanis gyakorlatilag bármilyen ige mellett megjelenhetnek. Szerencsére munkánk jelenlegi szakaszában úgy tűnik, hogy a fent említett okok kiküszöbölésében még jelentős javulást lehet elérni, és ezáltal a pontosság nagy mértékben javítható.

Bibliográfia

1. É. Kiss, K.: The syntax of Hungarian. Cambridge University Press, 2002

2. Gábor K., Héja E., Mészáros Á.: Kötőszók korpuszalapú vizsgálata. In: *Alexin Z., Csendes D. (szerk..): A Második Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete, Szeged Egyetemi nyomda, 2004.* Szeged, pp. 305-306.
http://www.nytud.hu/oszt/korpusz/resources/gabor_heja_meszaros2004.ps

3. Komlósy, A.: Régensek és vonzatok. In: Kiefer F. (szerk.): Strukturális Magyar Nyelvtan I. Mondattan. Akadémiai Kiadó, Budapest, 1992. pp.: 299-528

4. Komlósy, A.: A műveltetés. In: Kiefer F. (szerk.): Strukturális Magyar Nyelvtan III. Morfológia. Akadémiai Kiadó, Budapest, 2000. pp.: 215-291

5. Prószéky G., Tihanyi L.: Humor -- a Morphological System for Corpus Analysis. Proceedings of the first TELRI Seminar in Tihany. 1996. Budapest, pp. 149-158.

6. Radford, A.: Transformational Grammar. Cambridge University Press, 1988. Cambridge

7. Silberztein, M.: Dictionnaires électroniques et analyse automatique de textes: Le systeme Intex. Masson, 1993. Paris

8. Váradi, T., Gábor, K.: A magyar INTEX fejlesztésről. In: *Alexin Z., Csendes D. (szerk): A Második Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete, Szeged Egyetemi nyomda, 2004. Szeged, pp. 3-11.*

9. Váradi, T.: The Hungarian National Corpus. Proceedings of the Third International Conference on. Language Resources and Evaluation, 2002. Las Palmas pp.385-389

10. Váradi, T.: Főnévi csoportok annotálása ClarK rendszerben. In: *Az Első Második Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete, Szeged Egyetemi nyomda, 2003.* Szeged, pp. 65-71.

11. www.xaira.org

Vonzatkeretek a Magyar Nemzeti Szövegtárban

Sass Bálint

MTA Nyelvtudományi Intézet, Korpusznyelvészeti Osztály
1068 Budapest, Benczúr u. 33.
joker@nytud.hu

Kivonat: Jelen munkálat célja a Magyar Nemzeti Szövegtár vonzatkereteinek felismerése, az MNSZ vonzatkeret-gyakorisági szótárának előállítása a rendelkezésre álló vonzatkeret-táblázat felhasználásával. Manuális vizsgálatok alapján megállapítható, hogy a készülő korpuszfeldolgozó eszköz egyszerű nyelvtanokkal is képes a vonzatkeretek felismerésére, elkülönítésére, gyakorisági viszonyaik hozzávetőleges megállapítására. A gyakorisági szótár visszahathat a vonzatkeret-táblázat fejlesztésére illetve a szótárírásban és szintaktikai elemzők készítésekor is felhasználható.

1 Bevezetés

Jelen munkálat távlati célja a Magyar Nemzeti Szövegtárban (továbbiakban MNSZ)található igei vonzatkeretek felismerése, azonosítása és az MNSZ *vonzatkeret-gyakorisági szótárának* elkészítése.

A vonzatkeretek felismerése az első lépés lehet a gépi szövegmegértés felé. Általa a korpuszból részleges szemantikai információt nyerhetünk. Lehetőség nyílik a vonzatkeret-táblázat empirikus adatokra támaszkodó továbbfejlesztésére. Hasonlóan, empirikus gyakorisági adatokra támaszkodó szótárak jöhetnek létre: segítségével a szócikkek gyakorisági alapú megválogatásán túl lehetőség nyílik a vonzatkeretek megválogatására gyakori keretek felvétele, ritka keretek elhagyása által. Ezen kívül a gyakoriság a szócikkbeli jelentéssorrend megállapításának is egyik szempontja lehet.

A 2005 nyarán indult munkálat kezdeti eredményeiről számolok be.

2 Eszközök

Három nyelvi erőforrás kerül felhasználásra: (1) az MNSZ-ben ismeri fel az alább ismertetett (2) vonzatkereteket a projekt keretében elkészülő (3) korpuszfeldolgozó program.

2.1 A vonzatkeret-táblázat

A Nyelvtudományi Intézetben 2001 és 2004 között készült lexikai adatbázis [2] a magyar nyelv alapszókincsét alkotó szavak szintaktikai és alapvető szemantikai tulajdonságait kódolja szófajonkénti elosztásban. A kiindulópont az igei argumentumszerkezetek kódolása volt (az adatbázis a vonzatkeret-táblázaton kívül a főnevek és melléknevek szemantikai tulajdonságait kódoló táblázatokból áll). Igevonzatként csak az olyan összetevők szerepelnak, ahol az igével szintaktikailag vagy szemantikailag nem teljesen kompozicionális szerkezetet áll elő. Az ezt a követelményt nem teljesítő gyakori igei kontextusok nem szerepelnek a táblázatban. Az adatbázis kezdetben az MNSZ leggyakoribb 20000 szavából indult ki, több lépésben bővült, jelenleg a Szeged Korpusz összes vonzatkeretét is tartalmazza.

Jelen munka során az igei vonzatkeretek forrása ennek az adatbázisnak gazdag igei vonzatkeret-táblázata, mely 9000 ige 18000 vonzatkeretét tartalmazza. A feldolgozás alapja az eredetileg Excel táblázatos forma XML-re konvertált egységesített változata.

2.2 Szintaktikai elemzés és vonzatkeret-illesztés

A vonzatkeretek felismerésének menete két részre tagolódik: a részleges szintaktikai elemzést és annotációt követi a vonzatkeretek tényleges illesztése.

Mindkét lépést a projekt keretében elkészülő korpuszfeldolgozó eszköz végzi. Az eszköz a vonzatkeretek felismerésére készül, de a későbbiekben egyéb, általános célú feldolgozó modulokkal kényelmesen kiegészíthető. A nyelvtant alkotó szabályokban távlatilag, szükség szerint több rendelkezésre álló tagmondatra bontó, tulajdonnévfelismerő [3] illetve frázisfelismerő [6-7] eljárást és eszközt fel kívánok használni. Ezeket egyesítve, ezekre építve alakítom ki a szabályrendszert.

A feldolgozás az MNSZ formátumának megfelelő, morfoszintaktikailag a Morphologic *Humor* elemzőjével [4] elemzett, egyértelműsített korpuszból indul ki. A kidolgozott morfológiai reprezentáció részletekbe menő lekérdezéseket tesz lehetővé, az elemzési lépésekben kihasználhatjuk a magyar nyelv morfológiája adta lehetőségeket. Az eszköz implementálja a többszintű reguláris nyelvtan (*cascaded regular grammar*) technológiát [1]: a nyelvtanokat egymásra épülő, tokenek feletti reguláris kifejezésekből alakíthatjuk ki.

Néhány hasznos kiegészítő funkció:

A szabályok megfogalmazásakor pozíciót is meg lehet adni, hivatkozhatunk például a mondat első szavára.

Tagadást is használhatunk, ezáltal könnyen megjelölhetjük például a nem-alanyesetű névszókat.

Öröklődés: a jelenlegi egyszerű formában az összes szerkezet automatikusan az utolsó tokenjének tulajdonságait örökli.

A többszintű annotációs tagek segítségével például a szabály számát bele lehet kódolni a tagbe hibakeresés céljából (pl. `X:1` és `X:2`), ugyanakkor a továbbiakban egységesen lehet hivatkozni a tagre `X`-ként; vagy az `NP:pred` annotációra hivatkozhatok ebben a konkrét formában, de általánosságban `NP`-ként is. (A szinteket kettőspont választja el egymástól.)

A szükségtelenné vált annotációkat törölhetjük. Ha adott szabály elől „el akarunk fedni” egy szerkezetet, ideiglenes cimkével annotáljuk, amit a szabály alkalmazása után törlünk.

A keretek illesztése során a program egyenként megnézi, hogy a mondat és a vonzatkeret egyes szerkezetei megfelelnek-e egymásnak, ha a vonzatkeret összes elemének talál megfelelőt, akkor a keret illeszkedik. Több illeszkedő keret esetén a legspecifikusabb keretet választja. Az illesztést nem befolyásolja, hogy esetleg a szintaktikai elemzés nem tudott teljesen lefedő elemzést nyújtani, csak a szükséges vonzatok megléte számít. Igementes mondatban természetesen nem lehet illeszkedő keretet találni.

3 Jelenlegi állapot

A jelenlegi rendszer a felismerési folyamat valamennyi lépését magában foglalja, a legtöbbet többé-kevésbé egyszerűsített formában. Az elemzett szövegtől eljutunk a nyers vonzatkeret-gyakorisági szótárig.

A nyelvtan egyszerű tulajdonnév felismerője lényegében nagybetűs szavak sorozatait keresi meg, kiegészítve azzal, hogy a mondatkezdő nagybetűs szót legtöbb esetben nem tekinti igazi nagybetűs szónak. Erre épül az főnévi csoportot felismerő, valamint az alany, az igei állítmány, a tárgy és a határozók azonosítására szolgáló nyelvtan.

A tagmondatra bontás problémáját egyelőre kikerülve, a tesztkorpuszba rövid, írásjelet nem tartalmazó, így jó eséllyel egy tagmondatból álló mondatokat választottam ki az MNSZ-ből. A tesztkorpuszt egész pontosan az MNSZ összesen 131682 darab 9-szavas mondata alkotta. Az irodalmi és a hivatalos nyelvben kissé ritkábbak a 9-szavas mondatok (átlagban 1500 szóra jut egy), mint a korpusz többi részében (ott 1200 szóra jut egy), azért nagyjából egyenetesnek tekinthető a megjelenésük. A tesztkorpusz így az MNSZ „nyelvét” képviseli, ami a sajtó többsége miatt leginkább a sajtónyelvhez hasonlítható.

A szövegben fellelhető hiányosságokkal, hibákkal (ismeretlen szavak, rossz mondatszegmentálás, elírások, szószéttöredezés, ékezetmentes részek, stb.) nem foglalkoztam.

Első lépésben csak a legegyszerűbb kereteket dolgoztam fel. Az egy tagmondat – egy vonzatkeret munkahipotézis alapján a vonzatkeretek közül elhagytam azokat, melyekben vonzatként tagmondat szerepelt. A szemantikai jegyeket nem vettem figyelembe, ezek kezelése a névszói táblázatok részletes feldolgozását kívánta volna meg. Figyelmen kívül hagytam az ige szintaktikai jegyeit és a főnévi igenévi vonzatokat is. Így az általam használt egyszerűsített vonzatkeretek lényegében a következőképpen épültek fel: adott igetőhöz névszói alany, tárgy és vonzatok tartoznak és minden egység esetében meg lehet adni, hogy milyen szófajú legyen, milyen esetet kíván illetve, hogy konkrétan mely szóalak, vagy lemma képviselje az adott pozíciót (pl. *részt vesz*, vagy *semmibe vesz*). A honnan? és hová? kérdésre felelő lokatívuszi vonzatok a táblázatban egybevonva szerepelnek, ezeknek a kezelése a megfelelő esetekkel illetve névutós kifejezésekkel történik. A keretekben meglévő opcionalitást úgy kezeltem, hogy önálló, csak kötelező elemeket tartalmazó alkeretet hoztam létre. Az aktuálisan feldolgozandó kategóriába így 16300 vonzatkeret (a 18000 keret 90%-a) került be.

Az esetlegesen előforduló azonos kereteket összevontam eggyé. Az azonosság oka legtöbbször nyilván az volt, hogy épp olyan tulajdonságokat hagytam el, melyek a keretek közötti különbséget adták, ugyanakkor voltak ténylegesen duplikált sorok illetve az opcionalitás használatának következetlenségéből adódó esetek is.

4 Eredmények

A rendszer a tesztkorpuszon végzett manuális vizsgálatok alapján megfelelően képes felismerni az egyszerű vonzatkereteket, el tudja különíteni adott ige különböző vonzatkereteit. A program jelen változatából látszik, hogy egyszerű nyelvtannal, nagy mennyiségű, nem hibátlan szövegen is képesek lehetünk körülbelüli gyakorisági viszonyok megállapítására. Ehhez ugyanis nem szükséges az összes vonzatkeret pontos felismerése, csak az, hogy a program a felismerésben ne egyoldalúan hibázzon.

Elkészült egy mag program, aminek a továbbfejlesztésével létre lehet hozni egy olyan rendszert, amely képes előállítani a bevezetőben említett gyakorisági szótárat.

4.1 Példák

Néhány jó példa:

egybevet vmit vmivel:
"Az önellenőrzés során a dolgozó egybeveti munkáját a követelményekkel."

utasít vkit vmire:
"A Közgyűlés utasítja a Polgármestert a szükséges intézkedések megtételére."

részt vesz vmiben:
"A Pénztárfelügyelet képviselője a közgyűlésen tanácskozási joggal vesz részt."

A jól felismert vonzatkeretek mellett, számos tanulsággal szolgáló eset is előfordult. Fény derült olyan vonzatkeretekre, melyek esetében a táblázat hiányos vagy nem teljesen helyes. Az *őrizetbe vesz* vonzatkeretet például explicit módon nem tartalmazza a táblázat, a *kisüt* esetén pedig mindig megköveteli a tárgyat. Az alábbi mondatban így a program megtalálta az alanyt, az állítmányt és a két határozót, illeszkedő vonzatkeretet viszont – helyes működéssel – nem talált:

"Az ország nagy részén hosszabb-rövidebb időre kisüt a nap."

Hasonlóan a *demonstrál* igének is csak különféle vonzatos (vmit, vmi mellett, vmi ellen) formái szerepelnek, de a vonzatnélküli változat nem, pedig az a leggyakoribb.

„Nemet mondott a kétfős frakcióra a Parlament Ügyrendi Bizottsága"

A fenti mondatra illeszkedett a *mond vmit vmire* keret, de megfontolandó lenne a *nemet mond vmire* keret felvétele. Utóbbi nem szerepel a Magyar Értelmező Ké-

ziszótárban [5] (továbbiakban ÉKSz.), viszont az MNSZ-ben hétszer több az előfordulása, mint az ÉKSz-ben szereplő *rosszat mond vkire* keretnek.

”Fizetéskiegészítést kapnak az év végén az ügyeletes egészségügyi dolgozók.”

A fenti mondatra – helytelenül – illeszkedett a *kap vmit vmin* (pl. lopáson) keret. A problémát az okozza, hogy a szabad határozó esetragja egybeesik a szükséges vonzat esetragjával.

4.2 Esettanulmányok

Két olyan igét választottam ki, melynek húsznál többféle vonzatkerete van, az illeszkedéseket megvizsgáltam, az alábbiakban foglalom össze a tapasztalataimat.

A *vág* szótő 90 mondatban fordul elő. Helyesen ismerte fel a program a *pofát vág*, *vág vmibe* (arcába, szavába, témába) és *vág vmit* kereteket. Utóbbi esetén azonban jól elkülöníthető volt három altípus: legtöbbször a specifikusabb *vág vmit vmire* (szeletekre, darabokra, karikákra stb.) keret fordult elő, (ez a táblázatban nem szerepel, csak a még konkrétabb *zsebre vág*), ritkábban pedig az igemódosítós *fát vág*, *grimaszt vág* illetve a *vág vmit rajta/belőle* keret.

A *vág vmit vmibe* legtöbb esetben a *nagy fába vágta a fejszéjét* specifikus keretként jelent meg, a kissé kétes értelmű *vág vminek* pedig a *neki* szó helytelen névmási elemzése miatt a *nekivág vminek* kerettel bíró mondatokat ismerte fel.

A ritkább keretek sok esetben helytelenül szabad határozót találtak meg. A *vág vkit vmin* általános keret helyett hasznosabb lenne a *pofon vág*-ot kiegészíteni a *kupán vág* kerettel. Fokozottan igaz ez az extrém ritka keretekre: a *vág vmiben* (ti. hónaljban) – minden esetben helytelenül – szabad határozóra illeszkedett.

Az elváló igekötő miatt előkerültek a vág igekötős formái is. Nem szerepel a táblázatban a *levág vmit vmiből* (szeletet, darabot), illetve a *kettévág vmit* (kelbimbót, szelet, uborkát, országot). Utóbbi 4 előfordulással a vág szótövet tartalmazó mini részkorpuszban 4%-ot képvisel! Az *elvág vmit* pedig összevonja az *elvágja a torkát* illetve *elvág vmit vmitől* nagyon különböző jelentésű kereteket.

A *vesz* szótő (1011 mondat) konkrétan lemmát megadó kereteinek a gyakoriságát és az ÉKSz-beli jelentéssorrendet vetettem össze az 1. Táblázatban.

1. Táblázat: A *vesz* lemmát megadó keretei az ÉKSz-ben

vonzatkeret	***db***	***ÉKSz. jelentés***
részt vesz vmi(be)n	210	5. jelentés
tudomásul vesz	18	16. jelentés
fordulatot vesz	8	18. jelentés (sajtó)
semmibe vesz	2	*nincs benne!*
feleségül vesz	1	11. jelentés

Megfigyelhető, hogy a jelentések sorrendje nem teljesen felel meg a gyakoriságnak: bizonyos keretek/jelentések (*tudomásul vesz*, *fordulatot vesz*) hátrébb, bizonyosak (*feleségül vesz*) előrébb vannak sorolva. Egy szótár írásakor megfontolandó lehet, hogy a *részt vesz*, mely 210 előfordulásával az összes(!) vesz igéjű mondat 21%-át adja, a jelentések között valahol legelöl szerepeljen.

Megvizsgáltam még két olyan igét, melyek többféle névutóval előfordulhatnak: a *fut vmi elől* és *fut vmi után* ÉKSz-beli megjelenése megfelel a gyakoriságnak. A *vádat emel vki ellen vmi miatt* (5 előfordulás) keret az ÉKSz-ben nem szerepel.

4.3 A vonzatkeret-gyakorisági szótár első változata

A 131682 mondatos tesztkorpuszon 142 perc alatt futott le a vonzatkeret-illesztés. 111522 mondatban talált kerettel rendelkező igetővet a program. A többi mondatban számos oka lehetett annak, hogy nem sikerült kerettel rendelkező igét azonosítani: eleve nem volt ige; képzett ige szerepelt; az ige minden kerete kiesett az egyszerűsítés során; hibásan vonta össze az igekötőt a program, stb. Végül ezek közül 102184 mondatban (92%) azonosított be a rendszer keretet.

A 2. Táblázatban egy szemelvényt látható a vonzatkeret-gyakorisági szótár első változatából, mely még minden bizonnyal külön féle forrásokból eredő számos hibát tartalmaz.

2. Táblázat: A lemmát is megadó keretek gyakorisági listájának kezdete

#	*vonzatkeret*	*db*
1	részt vesz vmiben	124
2	részt vesz vmin	103
3	kérdést tesz fel	27
4	tudomásul vesz	23
5	győzelmet arat	16
6	szert tesz vmire	16
7	figyelmet fordít vmire	13
8	hatást gyakorol vmire	12
9	világra jön	9
10	letartóztatásba helyez vkit	9

5 Alkalmazás

Mint a fentiekben láttuk, a vonzatkeret-gyakorisági szótár alkalmas a vonzatkeret-táblázat továbbfejlesztésének támogatására. A gépi használatra szánt lexikai adatbázist "kipróbálva" tisztázódnak azok a pontok, ahol talán változtatni érdemes a táblázaton: cél lehet a vonzatkeretek megfelelő finomítása, specifikussá tétele, leginkább a ritka keretek esetén, és főleg azokban az esetekben, ahol nem azokra a mondatokra illeszkedett egy-egy keret, amelyekre a táblázat szerzői gondoltak (pl. *vág vmiben*). Ha a vonzatkeret-táblázat célja valamiféle gépi megértés, akkor egyrészt valamilyen formában jelentéseket kell rendelni az egyes vonzatkeretekhez, másrészt az egymástól határozottan eltérő jelentésű kereteket külön kell kódolni akkor is, ha köztük sok formai hasonlóság van.

Egy vonzatkeret gyakran egy szótárbeli jelentésnek felel meg. A gyakorisági szótár alapján lehetőség lesz változtatni szótárak (pl. az ÉKSz.) "jelentéskincsén" illetve a gyakoribb jelentéseket előrevéve a jelentések sorrendjén. A szótárakban általában az a gyakolat, hogy először az alapszó jelentései vannak részletesen kidolgozva, csak aztán következnek a kifejezések. Így fordulhat elő, hogy a *kezébe/nyakába vesz* előrébb szerepel az ÉKSz-ben, mint a nagyságrendekkel gyakoribb *részt vesz*.

Nem utolsó sorban a megbízható vonzatkeret-felismerő hatékony szintaktikai elemző elkészítéséhez járulhat hozzá.

6 Fejlesztési lehetőségek

Egyértelmű, hogy a jelen dolgozatban bemutatott állapot csak egy kezdeti lépcsőt jelent. Nagyban fejleszthető a vonzatkeret-azonosítás megbízhatósága jobb nyelvtanokkal, a szemantikai jegyek tekintetbe vételével, a névszótáblázatok feldolgozásával, integrálásával. Fontos feladat tagmondatra-bontó modul beépítése, mely képessé teszi a rendszert összetett mondatok vonzatkereteinek megtalálására is. A biztosan szabad határozónak minősülő mondatrészeket megfelelő eszközzel ki kell szűrni.

A program korrekt kiértékeléséhez nagy mennyiségű manuális munkára vagy elég nagy kézzel annotált korpuszra van szükség.

Bibliográfia

1. Abney, S.: Partial Parsing via Finite-State Cascades. In *Proceedings of the ESSLLI`96 Robust Parsing Workshop*, 1996, pp. 1-8.
2. Gábor K.: Lexikai adatbázis dokumantációja, MTA Nyelvtudományi Intézet belső dokumentuma, 2004.
3. Gábor K., Héja E., Mészáros Á., Sass B.: Nyílt tokenosztályok reprezentációjának technológiája. `http://www.nytud.hu/oszt/korpusz/resources/ikta_ner.doc`
4. Prószéky G., Tihanyi L.: Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany,* Budapest, 1996, pp. 149-158.
5. Pusztai, F. (szerk.): *Magyar Értelmező Kéziszótár.* Akadémiai Kiadó, 2003.

6. Váradi T.: Főnévi csoport annotálása a CLaRK rendszerben. In *Alexin Z., Csendes D. (szerk.): MSZNY2003*, Szeged, 2003, pp. 65-71.
7. Váradi T., Gábor K.: A magyar INTEX fejlesztésről. In *Alexin Z., Csendes D. (szerk.): MSZNY2004*, Szeged, 2004, pp. 3-10.

Személyragos főnévi igeneves bővítményt megengedő predikátumok kinyerése a Magyar Nemzeti Szövegtárból

Bottyán Gergely, Sass Bálint

MTA Nyelvtudományi Intézet, Korpusznyelvészeti Osztály
1068 Budapest, Benczúr u. 33.
{bottyang, joker}@nytud.hu

Absztrakt

A szláv nyelvekben, valamint a legtöbb germán és újlatin nyelvben csak személyrag nélküli főnévi igenevek fordulnak elő. A magyarban azonban – néhány egymással különben nem rokon nyelvhez, például a walesihez és a portugálhoz hasonlóan – találhatunk személyrag nélküli és személyraggal ellátott főnévi igeneveket is [4].

A magyar főnévi igenév két típusát mutatja be az 1. táblázat.

1. táblázat: A magyar főnévi igenév két típusa

A kulcsot meg kell keresni.	(1)
Sikerült mindent időben befejeznie.	(2)

A főnévi igenév személyragos változata (1. táblázat, 2) az utóbbi időben több, a magyar mondatszerkezet generatív szemléletű leírásával foglalkozó nyelvész figyelmét is felkeltette. A vonatkozó kutatások igyekeztek azokat a mondatkörnyezeteket azonosítani, amelyekben személyragos főnévi igenevek megjelenhetnek. Ezen kívül célul tűzték ki a személyragos főnévi igeneveket tartalmazó kifejezések szerkezetének ábrázolását is [1,5,6]. A mondattan chomskyánus iskolája művelőinek többségéhez hasonlóan a kutatás során a témáról szóló dolgozatok szerzői saját nyelvi intuícióikra hagyatkoztak, és nem alkalmaztak szisztematikus adatgyűjtési eljárást. Ennek ellenére [5] például azt állítja, hogy kimerítő felsorolását adja a személyragos főnévi igenevek használatát megengedő mondatkörnyezeteknek, valamint, hogy empirikus anyagon alapul.

Jelen dolgozat annak a kutatásnak az eredményeiről számol be, amelyet a 153,7 millió szavas szótövesített, morfológiai annotációval ellátott és egyértelműsített Magyar Nemzeti Szövegtár [7] alapján végeztünk. Elsődleges célunk az volt, hogy megvizsgáljuk annak az állításnak az érvényességét, miszerint [5]-ben minden, személyragos főnévi igeneves bővítményt megengedő predikátum (ezeket ezentúl licenzornak fogjuk nevezni) szerepel. Ehhez a korpuszadatokból kívántunk licenzorlistát készíteni. Egy további cél volt annak megállapítása, hogy a megfelelő korpuszmondatokban mely licenzorok mely személyragos igenevekkel fordulnak elő. A nyelvtechnológiában általános módon félig automatikusan, félig manuálisan oldottuk meg a feladatot, iteratív módszerrel.

Először a morfoszintaktikai kód alapján automatikusan kigyűjtöttük a Magyar Nemzeti Szövegtárból a személyragos főnévi igenevet tartalmazó mondatokat. Munkahipotézisünk az volt, hogy a licenzorok a személyragos főnévi igenévvel azonos tagmondatban fordulnak elő. Mivel a korpuszban tagmondat-annotáció nincs, a tagmondatjelölteket közelítő módszerrel próbáltuk megfogni. A közelítő módszerhez a tagmondathatároló írásjelek és tagmondatkezdő kötőszavak [3] listáját használtuk fel. A kapott tagmondatjelöltekből automatikusan kiszűrtük azokat, amelyek [2]-ben azonosított licenzort tartalmaztak, az előforduló személyragos főnévi igenevet pedig feljegyeztük a megfelelő licenzorhoz. Ezután következett az eljárás manuális része: a megmaradt tagmondatjelöltekben új licenzorokat kerestünk, majd az új, kiegészített licenzorlistával újra alkalmaztuk az eljárást elölről.

Három iteráció után a következő eredményekre jutottunk. A szótövesített licenzorlista 197 tagú. Az ezeket tartalmazó tagmondatok a Magyar Nemzeti Szövegtárban szereplő 228367 személyragos főnévi igenév tokenből 223140-et (98 %) fednek le. Szótövesítve 17874 licenzor – személyragos főnévi igenév párt kaptunk.

Vizsgálatunk eredményei alapján az alábbi következtetéseket vonhatjuk le. (i) Bőven akad olyan mondatkörnyezet, amely [5]-ben nem került azonosításra, tehát az ott található licenzorlista nem kimerítő. (ii) Ha egy nyelvtani konstrukciónak legalább egy jegye (esetünkben: a személyragos főnévi igenév személyragja) megfogható a Magyar Nemzeti Szövegtárban vagy bármely más, hasonló méretű gazdagon annotált korpuszban, megéri a fáradságot félig automatikus, félig manuális módszerrel kinyerni a megfelelő adatokat a nyelvi erőforrásból, mert ezzel elejét vehetjük sietve levont téves konklúzióknak.

Bibliográfia

1. É. Kiss, K.: Agreeing infinitives with a case-marked subject. In The syntax of Hungarian. Cambridge: Cambridge University Press. (2002) 210-221
2. É. Kiss, K.: A személyragos alaptagú főnévi igeneves kifejezés. In É. Kiss, K., Kiefer, F. and Siptár, P., Új Magyar nyelvtan. Budapest: Osiris. (1999) 118-121
3. Gábor, K., Héja, E. and Mészáros, Á.: Kötőszók korpusz-alapú vizsgálata. In Alexin, Z. and Csendes, D. (eds.), MSZNY 2003 – I. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem. (2003) 305-306
4. Miller, D. G.: Where do conjugated infinitives come from? Diachronica, 20, 1. (2003) 45-81
5. Tóth, I.: Inflected infinitives in Hungarian. Ph. D. dissertation. Tilburg: University of Tilburg. (2000)
6. Tóth, I.: Can the Hungarian infinitive be possessed? In Kenesei, I. and Siptár, P. (eds.), Proceedings of the conference "Approaches to Hungarian". Budapest: Akadémiai Kiadó. (2002) 135-160
7. Váradi, T.: On Developing the Hungarian National Corpus. In Vintar, Š. (ed.), Proceedings of the Workshop "Language Technologies – Multilingual Aspects". Ljubljana: University of Ljubljana. (1999)

Szintaktikailag elemzett birtokos kifejezések algoritmizált fordítása adott formális nyelvre

Kardkovács Zsolt Tivadar, Tikk Domonkos

BME Távközlési és Médiainformatikai Tanszék, Média Labor,
H-1117 Budapest, Magyar tudósok körútja 2.
{kardkovacs,tikk}@medialab.bme.hu

Kivonat: Számos nemzetközi szakirodalom [5; 7; 10; 17; 20] foglakozott a birtokos szerkezetek szemantikai modellezésével, szemantikai sajátosságainak bemutatásával, azonban az eddig megalkotott modellek valamely konkrét birtokos szerkezetnek pontosan megfelelő formális mondat automatizált előállítását nem biztosítják. A cikkben megmutatjuk, hogyan lehet a problémát általános formában megoldani, illetve megmutatjuk, hogy az algoritmussal támogatott feldolgozásnak hol vannak a korlátai, melyek a még megoldandó feladatok.

1 Bevezetés

A birtokos jelzős szerkezetek nagyon sokféle szemantikai kapcsolatot fejezhetnek ki [5], ráadásul a birtokos- és birtokszerepek is felcserélődhetnek eltérő szövegkörnyezetben (pl. a könyv szerzője, a szerző könyve), így algoritmizált feldolgozásuk, formalizálásuk korántsem egyszerű feladat. A ma használatos általános célú válaszkereső rendszerek (Question Answering Systems, QAS), illetve internetkereső-motorok [3; 9; 4; 8] egyik jellegzetes hiányossága éppen ebből ered, hiszen természetes nyelvű bemenetek [2], és azon belül a birtokos szerkezetek feldolgozása nélkül, szótövezéssel, valamint a szóeloszlások statisztikai mutatói alapján azonos válaszokat kell kapjunk az alábbi kérdésekre:

- „Mikor látogatott az Egyesült Államok elnöke Oroszországba?”,
- „Mikor látogatott a Oroszország elnöke az Egyesült Államokba?”,
- „Oroszország melyik elnöke látogatott az Egyesült Államokba?”,
- „Kit látogathat meg Oroszország és az Egyesült Államok elnöke?”
- … (stb.)

A statisztikai modelleken alapuló megoldást azonban el kell vetnünk más okból is, hiszen a birtokos szerkezetnél – néhány idiomatikus kapcsolattól eltekintve – gyakori ismétlődésekre nem készülhetünk fel [5; 21], mélyreható szemantikai analízis nélkül pedig az egyes szerkezetekben rejlő sajátosságokat nem is azonosíthatjuk be. Sőt, a mondatstruktúra általános szintaktikai jellemzői sem feltétlenül segítenek a tájékozódásban, hiszen nagyon hasonló morfológiai, mégis lényegesen különböző mondattani szerkezettel rendelkeznek pl. az alábbi mondatrészletek is:

Józsinak az Írott-kő megmászása…

Józsinak a ColorStar tévéje…
Editnek a váza összetörése… [7]
Editnek az arca visszatükröződése…

(Itt jegyezzük meg, hogy Chisarik állításának, miszerint egy magyar mondatban nem lehet jelen DAT és NOM birtokos is ugyanabban a főnévi frázisban[73], a második mondatrészlet azonban ennek ellentmondani látszik.)

A „Szavak hálójában" projekt[74] keretében arra tettünk kísérletet, hogy a már létező nemzetközi tapasztalatokat összegezve egy olyan magyar nyelvű válaszkereső rendszert hozzunk létre, amely – legalábbis magyar nyelven – szemantikai szinten képes feldolgozni a felhasználói kérdésben megjelenő birtokos szerkezeteket.

A feldolgozás két nagy lépésre bontható: egy szintaktikai és egyfajta szemantikai elemzésre. A továbbiakban – terjedelmi okokból – csak az utóbbiról lesz csak szó. Ennek megfelelően, a kiindulási állapotunkban feltételezzük, hogy adottak a frázisban a birtokos és a birtok szerepű tagok – függetlenül attól, hogy azok szemantikailag helyesen vannak-e összerendelve –, és ehhez a struktúrához keressük a megfelelő formális, SQL nyelvű leírást, ha ilyen létezik. Cikkünkben megmutatjuk, hogy hogyan juthatunk az 1. táblázatban látható módon, annak baloldali eleméből kiindulva a jobboldali megfelelőjéig, vagy annak ekvivalenséig.

1. Táblázat: Birtokos és a velük ekvivalens SQL-kifejezések

Birtokos kifejezés	SQL

Bizet Carmenje	`SELECT cim FROM operak`
	`WHERE szerzo = 'Bizet'`
	`  AND cim = 'Carmen'`

Shakespeare drámái	`SELECT cim FROM dramak` `WHERE szerzo = 'Shakespeare'`
Edit címe	`SELECT cim FROM cimek` `WHERE nev = 'Edit'`

könyvek szereplői	`SELECT szereplo FROM szerepek`
	`WHERE darab IN`
	`  ( SELECT cím FROM konyvek )`

vállalat vezetői	`SELECT fonok FROM vallalat`

Petőfi anyjának a neve	`SELECT nev FROM szemelyek`
	`WHERE nev IN`
	`  ( SELECT anya FROM csaladfa`
	`  WHERE gyermek = 'Petőfi' )`

[73] "Hungarian now presents an interesting puzzle: it is impossible to have both a genitive and a dative possessor in the same noun phrase." (Chisarik, 2001 – p.11)

[74] A Magyar Köztársaság Kutatás-fejlesztési Hivatala által támogatott, NKFP-0019/2002 jelű projekt

A következő szakaszban bevezetjük a megoldás bemutatásához szükséges alapvető fogalmakat, majd megmutatjuk, hogy a birtokos szerkezetek feldolgozása milyen három jellegzetes típusproblémára redukálható. A harmadik szakaszban bemutatjuk a feldolgozást végző algoritmust. Rövid példákkal és diszkussziókkal illusztráljuk az algoritmus működését és sajátosságait a negyedik szakaszban. Végezetül, az utolsó szakaszban összefoglaljuk az elért eredményeinket.

2 A fogalomhasználatról

Birtokos szerkezetek nagyon változatos szemantikai kapcsolatokat képesek kifejezni a természetes beszédben, ami új kihívásokat jelent a nyelvfeldolgozásban is (lásd 2. táblázat). Ez alatt azt kell érteni, hogy egy szintaktikai elemzés után rá kell ismernünk arra az aggregációra, hivatkozásra vagy valamilyen relációra egy tudásbázisban, amelyik pontosan az adott birtokos kapcsolatot jellemzi.

2. Táblázat: Birtokos szerkezetek típusai

Genitivus típusok	Példakifejezések (angol és magyar)
származás-, forrásleírás	Moszkva küldötte (men of Rome)
anyagleírás	– (ring of gold)
rész-egész viszony	a tanszék vezetője (head of department)
mennyiségi leírás	húsnak kilója (pound of beer)
jellemzés	a jövő embere (man of yesterday)
(állandósult) kapcsolat	Péter felesége (Pam's address)
birtoklás leírása	Sára sapkája (John's coat)
alanyiság	Verdi operája (dramas of Shakespeare)
tárgyiasság	Bocskai portréja (II. Erzsébet arcképe)
cél- és szándékleírás	dolgozók iskolája (girls' school)
láncolás	Ábel apjának barátja (name of Tom's wife)

Jelölje $X \Rightarrow Y$ azt a birtokos szerkezetet, amelyben X a szerkezetben, szintaktikai értelemben a birtokos szerepű tag, míg Y a birtok.

A szintaktikai feldolgozást követően, azaz ha adott $X \Rightarrow Y$, valamiféle jelentéstani feldolgozást kell végezni, majd ennek segítségével kell egy formális megfelelőt, pl. SQL lekérdezést belőle előállítani. Természetesen, a kifejezés pontos jelentése függ a benne található szavak értelmétől, ahogyan ezt a 2. táblázatban is láthattuk, ráadásul ezeknek a feldolgozása egyáltalán nem magától értetődő (lásd 1. táblázat). Az SQL nyelven lekérdezhető relációs adatmodellek nem rendelkeznek átfogó szemantikai leírással, így két lehetőség áll előttünk:

1.) ontológiai modelleket építünk, amely kölcsönösen egyértelműen leképezi az adatbázis struktúráját a modellre és ebből kiindulva próbáljuk meg a nyelvtani kifejezésben szereplő kapcsolat jelentését értelmezni,

2.) vagy olyan sémastruktúrát alakítunk ki, amelyből kinyerhető a számunkra szükséges szemantikai információ.

Az ontológiaépítés problémáiból, nehézségeiből tanulva felmerül a kérdés, hogy nem lehetséges-e pusztán a második megfontolás alapján megoldani a problémát?

A kérdés megválaszolásához először tekintsük át, milyen elemekkel dolgozhatunk egy általános (relációs) adatbázis esetében. Az adatbázis adattartalma leírható egy négyes segítségével, azaz legyen $DB = <V, I, S, A>$, ahol V az adatbázisban található konkrét értékek, kitöltések, I a példányok, rekordok, S a sémák vagy osztályok és A az attribútum, leírók halmazai.

Ha megvizsgáljuk a birtokos szerkezetek adatbázisokra vetített viselkedését, akkor végeredményben hat típust különböztethetünk meg: a birtokos minden esetben séma (osztály) vagy rekord (példány, individuum), míg a birtok szerepű kifejezés egyaránt lehet séma, rekord vagy attribútum (6db aleset).

A ma ismert implementációk (pl. [16; 12; 19; 1; 22; 18; 15]) nem használnak fel ontológiai tudást, kizárólag az adatbázis-tartalomból építkeznek. Rögtön hozzátesszük, hogy az implementációkhoz tartozó adatmodellek nagyobb hányada az `<objektum, tulajdonság, érték>` hármasra illeszthető [13; 14], amely csak a példány és a tulajdonság, valamint a séma (objektumosztály) és az individuum közötti birtokos szerkezettel leírható viszonyokat ($S \Rightarrow A$ és $I \Rightarrow A$ típus) képes megragadni.

Az általánosabb megoldáshoz további fogalmakat kell bevezetnünk.

1. Definíció (Természetes kulcs). *Legyen adva egy $DB = <V, I, S, A>$ adatbázis, ahol rendre V az értékek, I az individuumok, S a sémák és A az attribútumok nem üres halmaza. Legyen $\kappa: S \rightarrow A$ egy függvény, amely minden $s \in S$ elemet valamely $\alpha \in A$ elemre képez le úgy, hogy s minden i individuumát a természetes nyelvben éppen az i α attribútumának a $v \in V$ értékével nevezzük meg. Ebben az esetben azt mondjuk, hogy $\kappa(s)$ az s séma természetes kulcsa.*

Más szavakkal, a természetes kulcsok a mondatban szereplő (esetleg összetett) névszók, amelyek a nyelvben a való világ egy-egy entitását leíró, konkrét, adott nyelvű, és az adott környezetben egyértelmű megnevezései. Például $\kappa(könyv)=cím$, vagy $\kappa(személy)=személynév$. Vegyük észre, hogy a természetes kulcsok a közismert adatbáziskulcsoktól abban tér el, hogy a természetes kulcsokra nincs rögzítve, hogy szükségszerűen egyedinek kellene lennie. Azaz, a természetes kulcs lehetővé teszi a adatbázisban a nyelvi többértelműség megjelenítését.

2. Definíció (Hivatkozási függvény). *Legyen adva egy $DB = <V, I, S, A>$ adatbázis, és legyen $\alpha \in A$ az $s \in S$ séma egy attribútuma. Tegyük fel továbbá, hogy egy attribútumnév legfeljebb egy sémában fordul elő, és DB-n értelmezett a természetes kulcs fogalma. A $\varphi: A \rightarrow A$ függvényt hivatkozási függvénynek nevezzük, ha az adatbázisban*

1.) minden attribútumra $\varphi(\alpha) \in A$,

2.) $\varphi(\kappa(s)) = \kappa(s)$,

3.) $\varphi(\alpha) = \kappa(s')$, ahol $s \neq s'$.

Szintén definiáljuk az inverz függvényt, amelyet $\varphi^{-1}(\alpha)$-val jelölünk.

Könnyen igazolható, hogy a hivatkozási függvény az attribútumokat ekvivalencia-osztályokba sorolja. A továbbiakban jelöljük α attribútum ekvivalencia-osztályát $\|\alpha\|$-val, amelyet az α lezárásának nevezünk. A hivatkozási függvény az adatok használatát leíró egyszerű matematikai konstrukció, hiszen pontosan meghatározza, hogy szemantikailag helyesen mely értékeket milyen másik értékekkel lehet illeszteni.

Vegyük észre, hogy a hivatkozási függvénnyel rendelkező adatbázisban az attribútumértékek és az individuumok halmaza nem válik el egymástól, azaz $V = I$. Az ilyen adatbázisokat a továbbiakban *(V)ISA-modell*nek fogjuk nevezni. Könnyen belátható, hogy minden adatbázishoz létezik egy *(V)ISA-modell*, amely pontosan az adatbázis jellemző használatát írja le. Az egyedi jellemzőket a transzformáció során külön sémába érdemes felvenni, hiszen ennek és csak ennek természetes kulcsa a megfelelő azonosítókód.

3. Definíció (Általános birtokos szerkezet). *Legyen adva egy DB = <V, I, S, A> (V)ISA-modell egy φ hivatkozási függvénnyel. Az $X \Rightarrow Y$ birtokos szerkezetet általánosnak nevezzük akkor és csak akkor, ha az alábbi állítások egyikét igazzá teszi:*

1.) $X \subseteq I$ és $Y \subseteq I$ (pl. Bizet Carmenje)
2.) $X \subseteq I$ és $Y \subseteq S$ (pl. Shakespeare drámái)
3.) $X \subseteq I$ és $Y \subseteq A$ (pl. Edit címe)
4.) $X \subseteq S$ és $Y \subseteq I$ (pl. kazánok Rolls Royce-a)
5.) $X \subseteq S$ és $Y \subseteq S$ (pl. könyvek szereplői)
6.) $X \subseteq S$ és $Y \subseteq A$ (pl. vállalatok címei)

Az általános birtokos szerkezetek a leggyakoribbak a természetes nyelvekben. Általában elmondható, hogy attribútum, jellemzőleírás nem szerepel birtokosként a mondatokban – és éppen emiatt ki is merítettük az összes kombinációs lehetőséget, már ami a relációs adatbázisok kifejezőerejét illeti.

Vizsgáljuk meg alaposabban a 3. definíció 4-6. kritériumait! A birtokos szerepben levő séma megnevezése az adott fogalom általános értelemére utal. Bár a séma lényegében a benne található individuumok, példányok összességével matematikailag azonosan kezelhető, a két „leírás” azonban nem feltétlenül azonos. Amikor például könyvek szereplőiről beszélünk, nem konkrétan egy, de nem is az összes könyvről jelentünk ki valamit. Ráadásul, hogy mi legyen séma és mi individuum azt sokszor a szemlélődő, a modellező nézőpontja határozza meg, nem lehet minden esetben éles különbséget tenni e kettő között. Az azonban elmondható, hogy a séma bizonyos individuumok valamilyen absztrakcióját jelenti, de alapvetően maga is egy individuum egy másik nézőpontból, más viszonyokat tekintve, azaz $S \subseteq I$. Tehát az absztrakciós szinteknek megfelelően hierarchiát definiálhatunk individuumok között, amelyben bár jellemzően csak a hierarchia alján levő elemeket (leveleket) szokás individuumnak, a felettes rétegek elemeit sémának nevezni, de a hierarchiát lényegében ugyanolyan elemek alkotják.

4. Definíció (Speciális birtokos szerkezet). *Legyen adva egy DB = <V, I, S, A> (V)ISA-modell egy φ hivatkozási függvénnyel. Az $X \Rightarrow Y$ birtokos szerkezetet speciálisnak nevezzük akkor és csak akkor, ha $X \subseteq I$.*

3 Birtokos szerkezetek szemantikájáról

Az általános és a speciális birtokos szerkezetek csak a kifejezésben szereplő elemekre vonatkozóan tesznek bizonyos fokú megszorítást, ugyanakkor nem jelennek meg bennük a szemantikára vonatkozó leírások. A szemantika meghatározásához azonban

először szükségünk van az érvényesség kritériumára is, hiszen nem minden birtokos szerkezet mondható helyesnek [5; 17; 21].

5. Definíció (Birtokos szerkezet érvényessége). *Legyen $\Pi : A \times A$ egy reláció a $DB = <V, I, S, A>$ (V)ISA-modell felett. Ha $\alpha, \beta \in A$ attribútumokhoz létezik olyan $s \in S$ séma, amely tartalmazza $||a|| \cap ||b||$ valamely nem üres részhalmazát, akkor $\Pi(\alpha, \beta)$ akkor és csak igaz, ha α minden kitöltésére értelmes speciális birtokos kifejezést kapunk.*

Például a *Π(személynév, születési idő)* érvényes kifejezés, de az ellentéte nem az. Meg kell jegyeznünk azt is, hogy az érvényesség nyelvfüggő reláció, hiszen vannak olyan kapcsolatok, amelyek egyes nyelvekben, birtokos szerkezetben előfordulhatnak, más nyelvekben viszont nem [6; 21]. Gondoljunk csak arra, hogy az angolban használatos anyagleírások (pl. *ring of gold*) a magyarban birtokos szerkezetben nem léteznek (lásd még [7]).

Megvalósítást tekintve az érvényességi relációnak az adatbázisban egy két attribútummal rendelkező sémával ábrázolhatjuk, és azt mondhatjuk, hogy a reláció akkor és csak akkor érvényes, ha az érvényességi relációnak megfelelő sémában megtalálható az adott attribútumpáros.

6. Definíció (Birtokos szerkezet szemantikája). *Legyen $DB = <V, I, S, A>$ egy (V)ISA-modell, amelynek a hivatkozási függvénye φ, és legyen $X \Rightarrow Y$ egy tetszőleges speciális birtokos kifejezés. Vezessük be továbbá az alábbi jelöléseket:*

- *Az $\alpha \mid s$ jelölje azt, hogy $s \in S$ séma rendelkezik az $\alpha \in A$ attribútummal.*
- *$A\ \Sigma : 2^A \rightarrow 2^S$ egy olyan leképezés, amely az attribútumhalmazokhoz hozzárendeli azt a maximális sémahalmazt, amelynek elemeire igaz az, hogy az attribútumhalmaz legalább egy elemét tartalmazza.*
- *$A\ \sigma : I \rightarrow 2^A$ egy olyan leképezés, amely minden individuumhoz meghatározza azt a legnagyobb attribútumhalmazt, amelynek elemeire igaz az, hogy ott az individuum értékként szerepelhet. Azaz σ meghatározza azon attribútumok lezártjainak unióját, ahol egy adott individuum előfordulhat.*
- *Jelölje γ a $\sigma(Y)$, a $\varphi(\kappa(Y))$, illetve a $||Y||$ kifejezéseket, ha Y rendre individuum, séma, illetve attribútum.*

Az adott jelölések és feltételek mellett, ha léteznek $\alpha \in \sigma(X)$, $\beta \in \gamma$ attribútumok úgy, hogy $\exists s \in \Sigma(X) \cap \Sigma(\gamma)$ és $\alpha, \beta \mid s$, továbbá $\Pi(\alpha, \beta)$ igaz, akkor $X \Rightarrow Y$ nem más, mint a β attribútum értékei (individuumai), feltéve, hogy X egyetlen elemből álló halmaz. Máskülönben $X \Rightarrow Y$ az összes $\chi \Rightarrow Y$, $\chi \in X$, egyelemű birtokos szerkezet uniója.

A definíció alapján megállapítható, hogy a szemantika nem függ közvetlenül az adatbázis szerkezetétől, azaz az α, β és s paraméterek szabad változók, azok kizárólag a konkrét helyi adatbázis tartalmának függvényében lehet, illetve kell meghatározni. A megoldás tehát elég hatékony abban az értelemben, hogy a szemantikai feloldáshoz elegendő a megfelelő hármast megkeresni. Egy ilyen kényszerfeloldás könnyen megvalósítható például korlátos logikai programozással. Másfelől azonban szükség van az univerzális szerkezetű *(V)ISA-modell* leképezésére a helyi, keresni kívánt adatbázisra. Magát a leképezést azonban mindig a keresendő adatbázis adminisztrátorai valósítják meg és felügyelik.

A megközelítés újdonságereje éppen ezen alapszik: figyelembe veszi, hogy az Interneten található adatbázisok heterogén szerkezetűek, és csak arra van szükségünk a szemantikai információk kinyerésekor, hogy a megfelelő, értelmes, lekérdezésekben használt navigációkat, útkifejezéseket (path expressions) ábrázoljuk a lehető legtermészetesebb formában. Ez a leírás nem elhagyható, ugyanis ha az adatok összekapcsolásának mikéntje, módja nem adott vagy alulhatározott, akkor még ember számára sem egyértelmű, hogy hogyan kellene használni őket.

Hozzátesszük azt is, hogy a természetes kulcsok bármikor helyettesíthetőek a hagyományos, adatbázisokban használt (elsődleges) kulcsokkal, hiszen a definíciók és a szemantikai leírás nem használta a természetes kulcsnak egyetlen speciális tulajdonságát sem. Azonban azt is látni kell, hogy ebben az esetben a nyelvi többértelműségek idejekorán egyértelműsödhetnek, és nincs ilyenkor garancia a helyes döntésre. Ráadásul az olyan kifejezések esetében, mint a számok vagy a dátumok sokkal bonyolultabb szerkezetet kellene bevezetni.

1. Algoritmus ((V)ISA-algoritmus). *Legyen adva egy DB = <V, I, S, A> (V)ISA-modell egy φ hivatkozási függvénnyel. Az alábbi algoritmus, amelyet (V)ISA-algoritmusnak nevezünk, a speciális $X \Rightarrow Y$ birtokos szerkezetet dolgozza fel.*

```
1. function VISA( X, Y ) returns SQL
2. begin
   if ( isAttrib(Y) ) $gamma = ||Y||
3.   else if ( isSchema(Y) ) $gamma = φ(κ(Y))
     else $gamma = σ(Y);
4. find( α, β, s );
5. $add = isAttrib(Y) ) ? 'β = Y AND ' : '';
6. if ( isSet(X) )
     return 'SELECT β FROM s WHERE $add α = X'
   else {
7.   $Z = ( X is a V ⇒ W ) ? VISA(V, W) : X;
8.   if ( isSchema(X) )
       return 'SELECT β FROM s
               WHERE $add α IN ( SELECT κ(X) FROM X )'
     else
     return 'SELECT β FROM s WHERE $add α IN ( $Z )';
   }
9.  end;
```

Az algoritmusban a 4. sor azt jelenti, hogy a 6. definíciónak megfelelő hármasokat kell megkeresni, azaz rejtve itt jelenik meg a birtokos szerkezet érvényességi relációja.

4 A (V)ISA-algoritmus működés közben

Nézzük meg, hogy az algoritmus hogyan működik 1. táblázat elemeire! Vegyük például a „*Bizet Carmenje*” kifejezést. Tételezzük fel, hogy van egy α = `szerző`, β = `műcím` attribútum az `opera` sémában, azaz `opera` $\in \Sigma(\sigma(\|\texttt{Bizet}\|)) \cap$

Σ(σ(||Carmen||)). Mivel Carmen egy individuum, így az `$add` változó értéke `mű-cím = 'Carmen'`. Vagyis az algoritmus az alábbi kimenetet állítja elő:

```
SELECT műcím FROM opera
WHERE műcím = 'Carmen' AND szerző = 'Bizet'
```

Az eredmény megegyezik az 1. táblázatban látottakkal. Hasonlóan, a „*vállalat vezetői*”-t illetően:

```
SELECT vezető FROM vállalat
WHERE vezető IN ( SELECT vezető FROM vállalat )
```

Az algoritmus nem a legegyszerűbb választ találja meg, mindazonáltal ekvivalens az 1. táblázattal szereplő lekérdezéssel.

Összetett vagy láncolt birtokos szerkezeteket, mint amilyen a „*Petőfi anyjának a neve*” kifejezés volt, az algoritmus az alábbi módon fejti ki. Először is megpróbálja feldolgozni a teljes *(Petőfi ⇒ anya) ⇒ személynév* láncolatot. A fentebb már bemutatott lépések után az algoritmus a *(Petőfi ⇒ anya)* kifejezést dolgozza fel (7. sor). Ennek eredményeképpen létrejön a

```
SELECT anya FROM leszármazás
WHERE gyermek = 'Petőfi'
```

amelyet a teljes kifejezés megoldásába illeszt:

```
SELECT nőszemélynév FROM nőszemély
WHERE nőszemélynév IN ( SELECT anya FROM leszármazás
                        WHERE gyermek IN ( 'Petőfi' ))
```

Felmerülhet a kérdés, hogy az algoritmus hogyan következtet ki, hogy a leszármazás sémában Petőfinek gyermeknek kell lennie és nem anyának? Az algoritmus leírásában látszólag nincs akadálya annak, hogy ilyen alternatív megoldások is szülessenek, azonban a 4. sorban elrejtett *Π*-relációnak köszönhetően ez nem fordulhat elő, hiszen az kizárja a *Π(anya, anya)* lehetőségét.

5 Összefoglalás

A birtokos szerkezetek feldolgozását lehetővé tevő algoritmusok és matematikai modellek hiányoznak a számítógépes nyelvészeti irodalomból – külön tekintettel a magyar nyelvet illetően. A cikkben az alapvető problémákat és egy lehetséges megoldást mutattunk be, amely a természetes nyelvi birtokos szerkezeteket – ha adva vannak a szintaktikus szerepek – formális nyelvre, SQL-re képes fordítani.

Algoritmusunk univerzális abban az értelemben, hogy a magyar birtokos szerkezettípusokra jól működik, sőt, még a láncolásokat, az összetett birtokos szerkezeteket is képes feldolgozni, ugyanakkor vannak korlátai is. Mivel a megoldás során mindig adatbázisból, rögzített sémaszerkezet és relációk mentén nyerjük ki az információkat, így a (V)ISA-algoritmus nem tudja kezelni a metaforikus (pl. filmek Mekkája), az idiomatikus (pl. a jövő embere), valamint az olyan komplex, jelentéstömörítő kijelentéseket, amelyeket csak mélyebb emberi tudással dolgozhatóak fel (pl. mérkőzés győztese).

Beláttuk, hogy a megoldáshoz nem feltétlenül szükséges külön ontológiai tudás, a sémaszerkezetből, az elnevezési konvenciókból, valamint a nyelvfüggő érvényességi relációból a legfontosabb tudáselemek kinyerhetőek. Az algoritmus működését példákkal is illusztráltuk.

Bibliográfia

1. Androutsopoulos, I., Ritchie, G.D., Thanisch, P.: Masque/SQL – an efficient and portable natural language query interface for relational databases, in: P. Chung, G. Lovegrove, M. Ali (Eds.), Proc. of IEA/AIE 93 Conference, IEA/AIE Conferences, Edinburgh, Gordon and Breach Publishing (1993) 327–330.
2. Androutsopoulos, I., Ritchie, G. D., Thanisch P.: Natural language interfaces to databases – an introduction, Journal of Natural Language Engineering **1**(1) (1995) 29–81.
3. Answerbus. http://www.answerbus.com/
4. Askjeeves. http://www.ask.com/
5. Barker, C.: Possessive descriptions, PhD thesis, University of Carolina, Santa Cruz, Department of Linguistics (1995)
6. Barker, C., Dowty, D.R.: Non-verbal thematic proto-roles. In Schafer, A., ed.: Proc. of NELS 23 Conference. North-Eastern Linguistics Conferences, Amherst, Massachusetts, GLSA Publications (1992) 49–62
7. Chisarik, E., Payne, J.: Modelling possessor constructions in lfg: English and hungarian. In Butt, M., King, T., eds.: Proc. of the LFG01 Conference. International Lexical-Functional Grammar Conferences, Hongkong, Stanford CSLI Publications (2001) 49–62
8. Google. http://www.google.com/
9. Ionaut. http://www.ionaut.com:8400/
10. Jensen, P.A., Vikner, C.: The english prenominal genitive and lexical semantics. Workshop on the Semantics/Syntax of Possessive Constructions, Amherst, Massachusetts (2002) Invited paper.
11. Kardkovács, Zs.T.: On the Transformation of Sentences with Genitive Relations to SQL Queries. In Montoyo, A., ed.: Proc. of 10th NLDB Conference. Alicante, Spain, Springer-Verlag Publishing (2005) 10–20
12. Katz, B.: Using English for Indexing and Retrieving, Vol. 1 of Artificial Intelligence at MIT: Expanding Frontiers, MIT Press (1990) 134–165
13. Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J.J., Marton, G., McFarland, A.J., Temelkuran, B.: Omnibase: A uniform access to heterogeneous data for question answering, in: B. Andersson, M. Bergholtz, P. Johannesson (Eds.), Proc. of NLDB 2002, Vol. 2553 of Lecture Notes in Computer Science, Stockholm, Springer-Verlag (2002) 230–234
14. Katz, B., Lin, J.L.: Start and beyond, in: Proc. of SCI 2002, Vol. XVI of World Multiconference on Systemics, Cybernetics and Informatics (2002)
15. Meng, X., Wang, S.: Overview of a chinese natural language interface to databases: Nchiql, International Journal of Computer Processing of Oriental Languages 14 (3) (2001) 213–232
16. Popescu, A.-M., Etzioni, O., Kautz, H.: Towards a theory of natural language interfaces to databases, in: W. Johnson, E. Andre, J. Domingue (Eds.), Proc. of the 8th IUI 2003, Internation Conferences on Intelligent User Interfaces, Miami, ACM Press (2003) 149–157
17. Rappaport, G. C.: The syntax of possessors in the nominal phrase: Drawing the lines and deriving the forms, Possessives and Beyond: Semantics and Syntax, Amherst, Massachusetts, GLSA Publications (2005) 223–262

18. Reis, P., Matias, J., Mamede, N.: Edite: A natural language interface to databases – a new perspective for an old approach, in: Proc. of ENTER'97, Information and Communication Technologies in Tourism, Edinburgh, Springer-Verlag (1997) 317–326
19. Start. http://www.ai.mit.edu/projects/infolab/
20. Storto, G.: Possessives in context. Possessives and Beyond: Semantics and Syntax, Amherst, Massachusetts, GLSA Publications (2005) 59–86
21. Storto, G.: Possessives in context – issues in the semantics of possessive constructions, PhD thesis, University of California, Los Angeles, Linguistics (2003)
22. Stratica, N., Kosseim, L., Desai, B. C.: A natural language processor for querying cindi, in: V. Milutinovic (Ed.), Proc. of SSGRR 2002, International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet, L'Aquila, Italy (2002)

Szintaktikai elemzők eredményeinek összehasonlítása

Hócza András[1], Kovács Kornél[2], Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatika Tanszék
6720 Szeged, Árpád tér 2.
hocza@inf.u-szeged.hu

[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi védtanuk tere 1.
{kkornel, kocsor}@inf.u-szeged.hu

Kivonat: A mondatok szintaktikai elemzése alapvető részét képezi további természetesnyelvi feladatoknak. Ezért fontos feladat, hogy a magyar nyelvre készült szintaktikai elemzők részére is kidolgozzunk egy olyan hátteret, amelyet az angol nyelv esetén a Penn Treebank biztosít. A dolgozat beszámol a Szeged Treebank adattárára épülő adatbázis kialakításáról, amely magyar nyelvű szintaktikai elemzőkhöz készült modellépítés és az egységes összehasonlíthatóság érdekében. Ezen az adatbázison alkalmaztunk néhány rendelkezésre álló módszert, hogy összehasonlítási alapot teremtsünk a későbbi megközelítésekhez, valamint az egyik módszert a Penn Treebank angol szövegeire is alkalmaztuk, hogy képet kapjunk szintaxiselemzési szempontból a két nyelv összetettségéről.

Kulcsszavak: teljes szintaxis, gépi tanulás, szabály alapú módszerek

1 Bevezetés

Egy mondat teljes szintaxisának felismerése olyan folyamat, amely során meg kell határozni, hogy milyen egymás után következő szavak csoportosíthatók egybe, mint például főnévi, melléknévi, igei szerkezetek. Ezek a szócsoportok egymásba ágyazottak, fastruktúrát alkotnak. Egy mondat teljes szintaxisa olyan összefüggő fa, melynek levelei a mondat szavai, illetve írásjelei. A mondat szintaxisának feltárása számos természetesnyelvi feladathoz szolgáltathat alapvetően fontos információkat. Ilyen terület a feltárt mondatszintaxis felhasználására például az adatbányászat, információkinyerés, gépi fordítás. Ezért fontos kifejleszteni egy tetszőleges magyar szövegen jó hatásfokkal működő automatikus szintaktikai elemzőt.

Az angol nyelv szintaktikai elemzésének meglehetősen nagy szakirodalma van, mivel lényegesen korábban kialakítottak szintaktikailag annotált szöveges adatbázisokat. Ilyen például a Penn Treebank (Marcus et al., 1993), amelyen a tudományos cikkekben mérni szokták az elemzési módszerek pontosságát. Magyar nyelvre a Szeged Treebank (Csendes et al., 2005) munkálatai nemrég fejeződtek be, ami után megnyílt a lehetőség gépi tanulási technikákon alapuló magyar szintaktikai elemzők kifejlesztésére és tesztelésére.

A korábbi években már készültek magyar nyelvre szintaktikai elemzők és most, hogy lehetőség adódott rá, fontos feladat ezek megbízhatóságának a feltérképezése és összehasonlítása egymással, továbbá más, a szakirodalomban leírt módszerekkel. Azonban két módszer összehasonlítása csak teljesen azonos feltételek mellett, azonos szövegekre alkalmazva ad pontos képet a módszerek hatékonyságáról. Ezért dolgozatunkban beszámolunk arról, hogyan próbáljuk megteremteni az egységes összehasonlíthatóság feltételeit egy Szeged Treebank adattárára épülő adatbázis kialakításával. Ezen az adatbázison alkalmaztunk néhány rendelkezésre álló módszert, hogy meglegyen a kiinduló összehasonlítási alap a jövőben készülő hatékonyabb magyar szintaktikai elemzők számára. Az is érdekes kérdés, hogy az angol és a magyar nyelv szintaktikai elemzése mennyire összetett feladat, mik a különbségek, és mik a hasonlóságok. Ezért egy magyarra kifejlesztett elemzőt kipróbáltunk a Penn Treebank adataiból vett angol szövegeken is.

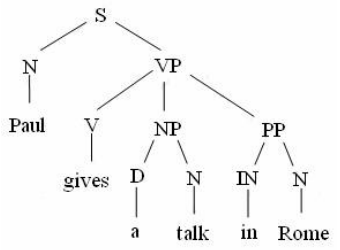
A dolgozat a következő módon épül fel: a 2. rész általánosan mutatja be a mondatszintaxis kutatását angol és magyar nyelvre, a 3. részben a magyar nyelvű szintaktikai elemzők egységes összehasonlításához készült adatbázisról lesz szó, a 4. rész a Szeged Treebank magyar szövegeinek adatbázisán elért eredményeket írja le, valamint az a Penn Treebank angol szövegin végzett próbát, és végül a 5. rész összefoglalja az elért eredményeket.

2 A mondatszintaxis felismerése

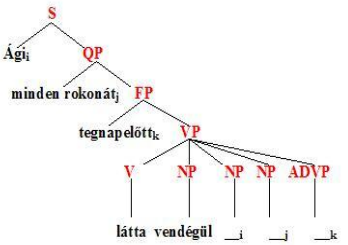
A mondatszintaxis felismerésének célja, hogy egy mondat különféle szócsoportokból álló elemzési fája automatikus módszerrel, minél pontosabban előálljon. A mondatok szintaxisának felismerése, valamint az eredmények mérése és az alkalmazott módszerek összehasonlítása számos problémát vet fel. A feladat nehézsége és az alkalmazott módszerek megvalósítása nagyban függ attól, hogy milyen nyelvről van szó.

2.1 A magyar nyelv szintaktikai elemzésének nehézségei

A magyar nyelv számos olyan nyelvi sajátossággal rendelkezik, ami megnehezíti a szintaxisfelismerést az indoeurópai nyelvekhez (pl. angolhoz) képest. Az egyik jelentős különbség a viszonylag szabad szórend (1. ábra), azaz egy mondat szintaktikai egységei többféleképpen átrendezhetők úgy, hogy a kapott mondatok nyelvtanilag szintén szabályosak lesznek. Azonban az így kapott mondatok jelentései módosulhatnak a kiinduló mondatéhoz képest. A mondatrészi szerepet a magyar nyelv ragozással és névutók alkalmazásával oldja meg. Ebből adódik a másik probléma, a nagyfokú morfológiai változatosság. Az említett sajátosságok összességében jelentősen megnövelik a lehetséges minták, nyelvi sémák számát, melyek rontják a statisztikai alapú gépi tanulás hatékonyságát.



Paul gives a talk in Rome.
– (nincs más lehetséges szórend)



Ági minden rokonát vendégül látta tegnapelőtt.
Minden rokonát vendégül látta Ági tegnapelőtt.

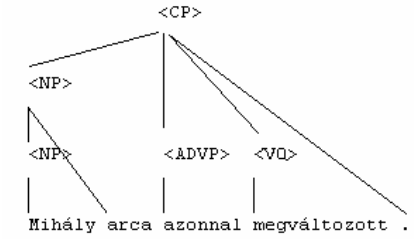
1. ábra: A magyarban a viszonylag szabad szórend miatt egy adott VP sokféle elrendezésben előfordulhat, sőt nem is mindig alkot összefüggő szerkezetet.

2.2 A szintaktikai elemzők pontosságának mérése

Az eredményeket az összehasonlíthatóság érdekében közös mérőszámokkal kell jellemezni. Az elemzett mondatokhoz rendelkezésre kell állnia egy nyelvész szakértő által készített kézi elemzésnek, ami etalonként szolgál. Minél jobban hasonlít az automatikusan előállított elemzési fa az etalonhoz, annál precízebb eredményről beszélhetünk. Fák hasonlóságának a vizsgálata úgy történik, hogy az összehasonlító algoritmus kigyűjti az összes előforduló szócsoportot (2. ábra) mindkét fából és ezeket veti össze a következő, az irodalomban gyakran alkalmazott képletek szerint:

- ***Pontosság:*** a helyesen felismert szócsoportok száma / az összes felismert szócsoportok száma.
- ***Fedés:*** a helyesen felismert szócsoportok száma / az etalonban ténylegesen szereplő szócsoportok száma.
- ***Középarány*** ($F_{\beta=1}$): 2* Pontosság* Fedés/(Pontosság+ Fedés)

Szintaxisfa:



XML leírás:

```
        <CP>
                <NP>
                        <NP>
1:                              Mihály
                        </NP>
2:                      arca
                </NP>
                <ADVP>
3:                      azonnal
                </ADVP>
                <V0>
4:                      megváltozott
                </V0>
5:              .
        </CP>
```

A szintaxisfából képzett szócsoport lista:
NP(1-1), NP(1-2), ADVP(3-3), V0(4-4), CP(1-5)

2. ábra: Példa arra, hogyan lehet egy egyszerű szintaxisfából szócsoportlistát képezni. A listaelemek egyértelműen azonosítják a szócsoportokat, ezért lehet a szócsoportlistákat fák hasonlóságának vizsgálatához felhasználni.

2.3 Az angol nyelv szintaktikai elemzésének szakirodalma

Az angol nyelvre számos eredmény létezik a mondatszintaxis felismerésének témakörében. A Penn Treebank (Marcus et al., 1993) annotált szövegeiben elkülönítettek egy részt, amely a megjelenése óta összehasonlítási alapot képez a témához kapcsolódó publikációk eredményeihez. Ezután többféle módszert alkalmaztak, hogy még jobb eredményt érjenek el, ezekből néhányat az 1. táblázat foglal össze.

Az első publikációban (Abney, 1991) nyelvtani kódok alapján osztályozta a szavakat, hogy azok kezdő, vég- vagy belső elemei-e egy adott típusú frázisnak. (Ramshaw és Marcus, 1995) transzformáción alapuló tanulást valósított meg. (Argamon, 1998) egyszerre végezte főnévi és igei szerkezetek felismerését. (Tjong Kim Sang és Veenstra, 1999) bevezette a több fokozatban (kaszkád) alkalmazott felismerést. A legújabb módszerek úgy érnek el javulást az eredményekben, hogy több módszert összekombinálva, szavazással hozzák meg a döntéseket, (Tjong Kim Sang, 2000) öt különféle módszert kombinált össze.

Hivatkozás	$F_{\beta=1}$	Teszt adatbázis	Módszer
Abney, 1991	-	-	Chunking
Ramshaw el al., 1995	92.0	Penn Treebank	Transzformációs tanulás
Argamon, 1998	91.6	Penn Treebank	NP- és VP-szerkezetek
Tjong K. S. et al., 1999	92.37	Penn Treebank	Többszintű nyelvtan
Tjong K. S., 2000	93.26	Penn Treebank	Több módszer kombinációja

1. táblázat: Néhány fontosabb angol nyelvre elért eredmény, 1993 óta az összehasonlíthatóság érdekében a Penn Treebank adatain történik a tesztelés.

2.4 A magyar nyelv szintaktikai elemzésének szakirodalma

A magyar nyelvre ez idáig lényegesen kevesebb szintaxiselemző készült. Az előzőekben vázolt nehézségek miatt szinte lehetetlen olyan nyelvész szakértők által kézzel készített szabályrendszert megalkotni, ami megfelelő hatékonyságú, és minden lehetséges esetre kiterjed. A másik probléma, hogy idáig nem volt elegendő mennyiségű annotált magyar szöveget tartalmazó korpusz, ami a gépi módszerek alkalmazását lehetővé tette volna.

A MorphoLogic Kft. által kifejlesztett HumorESK mondatelemző (Kis, 2003) 1995 óta folyamatosan fejlődik. Ez idő alatt különféle nyelvészeti területeken alkalmazták. Fő jellemzője, hogy a szimbólumokhoz jegyszerkezeteket (*feature structure*) kapcsol, és elemzési erdőt épít az egyes jegyek örököltetésével. Az elemzőben használt nyelvtan nyelvész szakértők közreműködésével állt elő. A Nyelvtudományi Intézet beszámol egy készülő szintaktikai elemzőről (Váradi, 2003), ami főnévi szerkezeteket ismer fel reguláris kifejezésekkel leírt, többfokozatú (kaszkád) szakértői szabályrendszer alkalmazásával. A Szegedi Egyetemen a nyelvtan előállítása gépi tanulási módszerekkel történt. A főnévi szerkezetek (Hócza, 2004), valamint a teljes szintaxis (Hócza, 2005) felismerésére készült módszerek modelljének forrását a Szeged Treebank annotált szövegei adták. Az eredmények összefoglalása a 2. táblázatban található:

Hivatkozás	$F_{\beta=1}$	Teszt adatbázis	Módszer
HumorESK (Kis, 2003)	-	-	Szakértői szabályok
Váradi, 2003	58.78	100 annotált mondat	Szakértői szabályok
Hócza, 2004	83.11	Üzleti hírek	NP-tanulás
Hócza, 2005	78.59	Szeged Treebank	Rövid faminták tanulása

2. táblázat: A magyar nyelvre elért eddigi eredmények

3 Szintaktikai elemzők összehasonlítása

Két módszer hiteles összehasonlítása megkívánja, hogy ugyanazokon az adatokon alkalmazzuk őket. A magyar nyelvre eddig még nem volt kialakítva olyan nyilvános adatbázis, mint az angol esetén a Penn Treebank. Cikkünk fő célja beszámolni arról, hogy hogyan próbáljuk megteremteni a hiteles összehasonlíthatóság lehetőségét a magyar szintaktikai elemzőkre. Ennek érdekében a következő módszert dolgoztuk ki:

- Kifejlesztettünk magyar nyelvre egy teljes szintaktikai elemzőt (Hócza, 2005), és alkalmaztuk a Szeged Treebank adatain.
- A Szeged Treebank adataiból kialakítottunk egy mintaadatbázist, amely a jövőben lehetőséget teremt minden érdekelt számára, hogy a treebank adatait felhasználva kipróbálja a szintaktikai elemzőjét, és összevesse annak hatékonyságát a magyar nyelvre alkalmazott más módszerekével.
- Néhány további módszert is kipróbáltunk a mintaadatbázison (3. táblázat).
- A rendelkezésre álló módszereket alkalmaztuk a Penn Treebank adatain is a hatékonyság lemérésére, valamint, hogy összehasonlítsuk az angol és a magyar nyelv összetettségét a szintaxis elemzés feladatának szempontjából.

3.1 A Szeged Treebank szövegeiből kialakított adatbázis

Példák véletlenszerű kiválasztása még nem garantálja azt, hogy egy módszer tesztelésénél ne kapjunk torz eredményeket. Például ha véletlenül egymáshoz nagyon hasonló példák kerülnek a tesztadatokba, a kapott végeredmény pontossága több százalékkal is eltérhet egy másik felosztással kapott értéktől. A nemzetközi szakirodalomban az eredmények közlésénél ennek a problémának az elkerülésére alkalmazzák a *tenfold cross-validation* módszert, melynek lépései a következők:

1. A példákat véletlenszerűen szétosztjuk 10 egyforma méretű csoportba.
2. Előállítjuk a tréningállományokat úgy, hogy mindegyik 9 különböző csoportból adódjon össze (mindig 1 kimarad). A különböző lehetőségek száma 10.
3. A tréningállományokhoz hozzárendeljük azok tesztpárját; ez az a csoport lesz, ami kimaradt a tréningből.
4. Lefuttatjuk a tanulást a 10 tréningcsoportra, majd elvégezzük a tesztelést a tréningállomány megfelelő teszt párján.
5. A teszteredményeket összesítjük és átlagoljuk; ez az átlag lesz a példákon végzett tanulás végső eredménye.

A Szeged Treebank adataiból ezt a 10 részre történő felosztást végeztük el ajánlásként a jövőben kipróbálandó módszerek egységes felkészítéséhez és összehasonlításához.

4 Eredmények

Ebben a részben arról számolunk be, hogy a számunkra rendelkezésre álló módszerekkel milyen eredményeket értünk el magyar és angol szövegeken.

4.1 Néhány módszer alkalmazása a magyar nyelvű adatbázison

A Szeged Treebank mintaadatbázisán 4 módszert alkalmaztunk a teljes szintaxis felismerésére. Az első, alapnak (baseline) tekinthető módszer a treebank adataiból kigyűjtött környezetfüggetlen valószínűségi nyelvtant (PCFG) használ, amely chart parsing segítségével építi fel a teljes szintaxisfát. A második módszer esetén a chart parser nyelvtanát kb. 22 ezer szabály képezte, amelyet nyelvész szakértők állítottak elő. A harmadik egy memória alapú módszer volt, amely a treebank adataiból kigyűjtött teljes mondatok szintaxisfáját illesztette a nyelvtani kódok figyelembevételével. A negyedik módszerben a chart parser gépi tanulással előállított famintákat alkalmazott.

Módszer	$F_{\beta=1}$	Teszt adatbázis	Megjegyzés
Baseline	56.01	Szeged Treebank	PCFG, chart parsing
Szakértői szabályok	55.58	Szeged Treebank	~22k szabály, chart parsing
Mondat memória	7.04	Szeged Treebank	Hasonló mondat keresése
Faminták	75.47	Szeged Treebank	Részfák illesztése, chart p.

3. táblázat: A magyar nyelv teljes szintaxiselemzésében a különféle módszerekkel elért eredmények a Szeged Treebank mintaadatbázisán modellt építve és tesztelve.

4.2 A famintákon alapuló módszer alkalmazása angol nyelvre

A magyar nyelvre alkalmazott módszerek közül a faminták tanulásán alapuló módszer érte el a legjobb eredményt, ezért ezt a módszert választottuk arra, hogy az angol nyelvre is kipróbáljuk. Így összehasonlítási alapunk lehet a szakirodalomban jó eredményeket elért további módszerekkel. A körülbelül 20 millió szót tartalmazó Penn Treebank szintaktikailag annotált szövegei szekciókra tagolódnak. Ebből a nagy adatbázisból elkülönítettek egy kisebb, körülbelül 1 millió szót tartalmazó részt, a 2-24-es szekciókat, melyen az összehasonlítani kívánt módszerek egységesen építhetnek modellt, és a pontosság mérése is egységes adatokon történik. Az általunk alkalmazott módszer esetén is ezeknek a feltételeknek megfelelően jártunk el.

A famintákon alapuló módszer alkalmazása során kipróbáltuk azt a speciális esetet is, amikor a faminták egymélységű fák. Ez az eset lényegében a környezetfüggetlen valószínűségi nyelvtant (PCFG-t) alkalmazó módszer, amit a magyar nyelv elemzése során is alapnak tekintettünk. Ezen a modellen az elemző 81.31%-os pontosságot ért el. A többmélységű famintákat alkalmazó elemzővel 85.73% volt az eredmény, ami 4,42%-os javulást jelent az alapmódszerhez képest. Az így elért pontosság megközelíti az angol nyelvre publikált eredményeket.

5 Összefoglalás és fejlesztési lehetőségek

A dolgozatban bemutatásra került egy a Szeged Treebank adattárára épülő adatbázis kialakítása. Ez a magyar nyelvű szintaktikai elemzőkhöz készül, hogy az adatbázis segítségével előállított elemzőket egységesen össze lehessen hasonlítani. Ezen az adatbázison kipróbáltunk néhány rendelkezésre álló módszert, melyek eredményei összehasonlítási alapot adhatnak a jövőben ezen az adatbázison alkalmazandó további algoritmusokhoz. A famintákon alapuló módszert – amely a legjobb eredményt érte el a magyar szövegeken – alkalmaztuk a Penn Treebank angol szövegeire is, és így képet kaptunk a két nyelv összetettségéről szintaxiselemzési szempontból.

A közeljövőben szeretnénk alkalmazni olyan szintaktikai elemzőket magyar nyelvre, amelyek jó eredményeket értek el angol szövegeken.

Irodalom

Abney S. (1991) Parsing by chunks, in *Principle-Based Parsing*. Kluwer Academic Publishers.

Argamon, S., Dagan, I., and Krymolowski, Y. (1998) A memory-based approach to learning shallow natural language patterns, in *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Montreal, pp. 67-73.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A., (2005) The Szeged Treebank, in *Proceedings of the TSD*, Karlovy Vary, pp. 123–131.

Hócza, A (2004) Noun Phrase Recognition with Tree Patterns, in the *Acta Cybernetica*, vol. 16, pp. 611–623.

Hócza, A., Felföldi, L., Kocsor, A., (2005) Learning Syntactic Patterns Using Boosting and Other Classifier Combination Schemas, in *Proceedings of the TSD*, Karlovy Vary, pp. 69–76.

Kis, B., Naszódy, M., Prószéky, G. (2003) Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer, *MSZNY 2003 konferenciakiadványa*, Szeged, 145-151 oldal.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of English: the Penn Treebank, Association for Computational Linguistics.

Ramshaw, L. A., and Marcus, M. P. (1995) Text Chunking Using Transformational-Based Learning, in *Proceedings of the Third ACL Workshop on Very Large Corpora,* Association for Computational Linguistics.

Simov K. (2001) CLaRK – an XML-based System for Corpora Development, in *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster, pp. 553-560.

Tjong Kim Sang, E. F., and Veenstra, J. (1999) Representing text chunks, in *Proceedings of EACL '99*, Association for Computational Linguistics.

Tjong Kim Sang, E. F. (2000) Noun Phrase Recognition by System Combination, in *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, Seattle, pp. 50-55.

Váradi T. (2003) Shallow Parsing of Hungarian Business News, in *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, pp. 845-851.

VI. Pszichológiai szempontú szövegfeldolgozás

Kézzel annotált adatbázis számítógépes feldolgozása a szövegnyelvészet, a szociolingvisztika és a neveléstudomány határterületén

Huszár Zuzsanna[1]

[1] PTE BTK Neveléstudományi Intézet, 7624 Pécs, Ifjúság útja 6.
zsizsu@t-online.hu

Kivonat: A társadalmi időkezelés iskolai fogalmazásokban is tükröződő feltevéséből kiinduló előadás a szövegelemzés alkalmazott módszereiről, tartalmi szempontjairól és eredményeiről, egyes tartalmak előfordulásának korosztályi jellemzőiről szól. A tíz- és tizennégy éves korosztály időszemléletére vonatkozó kutatási eredményeinket e fórumon tesszük közzé első ízben.
Kulcsfogalmak: szövegelemzés, tartalomelemzés, strukturális elemzés, Galois-gráf, időszemlélet, időszerkezet.

1 Módszertani bevezetés

Az előadás egy alkalmazott nyelvészeti PhD-kutatás újszerű eredményeinek összefoglalása a módszertani apparátus számítógépes elemeinek bemutatásával.

A kutatás témája idői struktúrák feltárása, elemzése és értelmezése a 10-16 éves korosztály szövegeiben a tanulói fogalmazások egy 2170 elemű reprezentatív mintáján.

A teljes mintára kiterjedő szóstatisztika [1] korábbi előadásainkban [2, 3] hivatkozott eredményeinek felhasználásán túl kvalitatív metodikával közelítettük meg a negyedik évfolyam és a nyolcadik évfolyam érvelő fogalmazásainak teljes rögzített anyagát; tartalomelemzéssel vizsgáltunk minden egyes fogalmazást, összesen több mint ezret. A kutatás a tartalomelemzés kvalitatív megközelítéséből és hagyományos lehetőségeiből kiindulva az elemzés mélységi kiterjesztésének igényével fordul a számítógépes feldolgozás lehetőségei felé.

Az érvelő fogalmazások két választható témája volt: *„Miért (nem) szeretek iskolába járni?”* és *„Milyen felnőtt szeretnék lenni?”* Az egyes fogalmazásokat a fogalmazások témája szerint 30, illetve 45 kategória előfordulása alapján kétértékű kódokkal láttuk el. A bináris kódok rögzítésével kialakított adatbázisunkban első lépésben az egyes szövegjellemzők előfordulását, illetve a szövegjellemzők előfordulásának települési, életkori és nemi meghatározottságát, valamint a tanulmányi eredményekkel való kapcsolatát vizsgáltuk. Az érvelési pozíció kódolásával azt is vizsgálhatóvá tettük, hogy mennyiben különböznek egymástól az iskolába járni szerető és nem szerető diákok fogalmazásai.

Az egyes szövegjellemzők függetlenségét, illetve egymással való kapcsolatát kereszttáblázatok elemzésével és a Pearson-féle $\chi 2$ próba alkalmazásával végeztük. A kapcsolatok erősségének megállapításakor a szimmetrikus mérés eredményeit is figyelembe vettük.

A nyolcadikosok érvelő fogalmazásainak szűkített, 40 elemű részmintáján tagmondatonként részletes elemzést is végeztünk a szövegalkotás mint folyamat előtérbe állításával. Az időbeliség, a térbeliség és az érzelmi értékelés szempontjainak egymásra vetítésével kísérletet tettünk a fizikai és időbeli távolságokhoz feltehetően kapcsolódó érzelmi attitűdök adott mintára vonatkozó azonosítására. A sajátos jellemzők mentén elkülöníthető fogalmazások kisebb csoportjait, s a tartalomelemzés alapján meghatározható típusait a strukturális elemzés Galois-gráfos módszerével ábrázoltuk és elemeztük. Az alkalmazott metodikában újdonságértékű a Galois-gráfok nyelvészeti, illetve számítógépes alkalmazása. Tartalomelemzési megoldásunk kétes értékű újdonsága, hogy az egyes fogalmazásokban adott tartalomnak nem az összes előfordulását rögzítettük, hanem csak azt regisztráltuk, hogy adott elem adott szövegben előfordul-e vagy sem. Ennek legfőbb indoka, hogy így tudtuk legjobban megoldani, hogy kutatásunk egy sokváltozós modell finom elemzése lehessen egy reprezentatív mintán a számítógépes feldolgozás lehetőségeinek kiaknázásával.

1.1 Tartalmi szempontok

Tartalomelemzési szempontjaink közül a következő néhányat emeljük ki:
- az idő különféle léptékeinek megjelenése a perctől a generációs léptékig;
- az életidőn, az egyén személyes életén túlmutató perspektíva megjelenése;
- az öregség, a saját öregkor előrevetítése a jövőtervekben;
- „holtomiglan-holtodiglan” idő;
- örökidő;
- időhiány;
- a jelenhez tapadó "időtlenség" képzete;
- kitüntetett események és fordulópontok ideje;
- linearitás és ciklikusság a szövegekben.

1.2 Hipotézisek

Alapvető hipotéziseink, hogy az iskolai fogalmazásokban megmutatkozó temporalitás nem homogén és nem statikus, hanem a tanulók bizonyos háttérváltozói mentén differenciált, és az életkorral változik. Feltesszük továbbá, hogy a fogalmazások időre utaló elemei és a mondanivaló időbeli rendezettsége a fogalmazási témától nem független.

2 Eredmények

2.1 Tartalomelemzési kategóriák előfordulása

Az előadás keretében érintett tartalomelemzési kategóriáinkat a következő csoportosításban mutatjuk be:

Az *1. sz. táblázat* az idő, mint lépték jelzésének előfordulásait összesíti. Ennek a táblázatnak a kategóriái az időtartamokat természetes és mesterséges metrumokként, illetve naptári kategóriákként jelenítik meg a percnyi egységtől a generációs léptékig, így ezeket kvázi-ritmikai elemeknek is tekinthetjük.

1. táblázat: Egyes tartalmi kategóriák aránya évfolyamonként és témakörönként az érvelő fogalmazásokban I. rész (%)

Tartalomelemzési kategóriák	Az iskolába járás témája		A felnőttség témája	
	4. o. (N=252)	8. o. (N=263)	4. o. (N=274)	8. o. (N=275)
Percnyi időegység	6,35	7,60	1,82	2,55
Óra és tanóra mint időmérték	61,11	61,98	7,30	4,36
Naptári nap	54,76	57,79	20,07	21,45
Naptári hét	16,67	16,35	6,93	10,55
Naptári hónap	2,78	9,51	2,19	4,36
Évszak	6,75	5,70	7,66	6,91
Naptári év és tanév	26,59	43,73	21,17	26,18
Iskolaszerkezeti váltások léptéke	18,65	44,11	29,20	52,73
Generációs lépték	8,33	11,03	76,64	61,45

Elsőként azokat a lépték-jellegű kategóriákat emeljük ki, amelyek megjelenése tartalom-függetlennek és korosztály-függetlennek mutatkozott a fogalmazásokban. Egyértelműen ilyen az évszak, és a hónap előfordulásában mutatkozó eltéréseknek sem érdemes jelentőséget tulajdonítani. A perc és a hét az iskola témájú fogalmazásoknak érthető módon nagyobb arányban eleme, mint a felnőttség témájúaknak. Kirívó mértékű, de nem meglepő tartalomfüggő eltérés mutatkozik az óra/tanóra használatának gyakoriságában és a naptári naphoz kötött tartalmak megjelenésében is. Míg az iskolába járás mint téma kifejtésének alapjait dominánsan a (tan)óra ritmusa és a napi ritmus adja, addig a felnőttség téma kidolgozása generációs képlet alapján történik. A felnőttség témája kétséget kizáróan távlatos lépték használatát hívja elő mindkét vizsgált korosztályban. Témafüggő, ugyanakkor nem korosztály-függő a naptári nap mint lépték használata. Az év, mint lépték használatában azonban már korosztályi szempontok is megmutatkoznak, amennyiben az iskola téma kifejtése kapcsán a (tan)év a nyolcadik évfolyamon jóval nagyobb hangsúlyt kap. Az iskolaszerkezeti váltások megjelenítése a szövegekben szintén inkább az idősebb korosztályra jellemző. Említésre méltó korosztályi különbség adott témán belül csak az iskolaszerkezeti váltások említésében van, és az iskola téma kifejtésekor a tanév említésében is. A nyolcadikos évfolyam fogalmazásai mindenestre távlatosabbak a negyedikesekéinél.

A *2. sz. táblázat* kategóriái hangsúlyosan jelenítik meg a távlat, a távlatosság kérdését. A távlat kategóriáját a napi rutin kifejezésénél, kifejtésénél túlmutató szövegekben értelmezzük először, majd tovább finomítjuk néhány olyan kategória használatával, amely gyakorlatilag az életidőn innen és túl szempontját vezeti be a kutatásba.

2. táblázat: Egyes tartalmi kategóriák aránya évfolyamonként és témakörönként az érvelő fogalmazásokban II. rész (%)

Tartalomelemzési kategóriák	**Az iskolába járás témája**		**A felnőttség témája**	
	4. o. (N=252)	**8. o. (N=263)**	**4. o. (N=274)**	**8. o. (N=275)**
Az idői lépték a napnál távlatosabb	67,66	62,36	100	100
Utalás a saját felnőtt korra	19,44	34,22	100	100
Életen át tartó kapcsolat	–	–	2,92	6,55
A hosszú élet kívánsága	–	–	4,01	4,00
A saját öregség kérdése	–	–	3,28	14,18
A személyes életidőn túl	5,95	12,93	7,66	19,64
Örökre és mindörökké	1,98	1,90	1,82	2,55
Időhiány irect jelzése	15,08	24,71	8,73	2,19
Szabadidő igénye	–	–	16,42	20,73

Tartalmilag új elem ennek a táblázatnak az utolsó két kategóriája: az időhiány és a szabadidő igényének direkt jelzése. Az időhiány jelzését az értelmezésben a szubjektív időélmény kifejezéséhez kötjük. Az időhiányt az iskola témájú fogalmazásokban a fogalmazó saját iskolás életformájára vonatkoztattuk, a felnőttség témájú szövegekben pedig a saját majdani felnőtt élet velejárójaként kódoltuk. A szabadidő igénye (szabadidős tevékenységek életformaként való beépítése a jövőtervekbe) csak a felnőttség témájú fogalmazásokban szerepel, mivel a szövegek olvasásakor csak ezekben kínálta magát önálló kategóriaként. Bár az időhiány és a szabadidő ugyanannak a jelenségnek a két oldala, a hiány negatív, az igény pozitív megfogalmazást jelent.

Azon túlmenően, hogy a távlatosság megjelenését adott téma eleve előhívja, a felnőtt életre utalás és az öregkorra való előretekintés mégis inkább nyolcadikban mutatkozik meg. A saját életidőn túlra tekintés ugyancsak inkább a 14 éves korosztályra jellemző mindkét fogalmazási témában. A vizsgált két korosztály közül a nyolcadikosok nagyobb arányban, ill. valószínűbben építenek be a szövegeikbe távlatos elemeket. Az egész életen át tartó, „holtomiglan-holtodiglan" idő negyedikben 3%, nyolcadikban pedig közel 7%. Az örökidő kategóriája tartalom- és korosztály-független, 2% körül stabil. Az időhiány iskolai életformával összefüggő jelzése növekvő arányt mutat az életkor előrehaladtával, amennyiben ez a nyolcadikosok mintegy negyede számára problémát jelentő kérdés a negyedikesek mintegy hatodának problémajelzéséhez képest. A „gyorsuló idő" szubjektív tapasztalata a tanulók szempontjából, úgy tűnik, jellemzően intézményhez, iskolához kötött, és elképzelhető, hogy a felnőtt élettervekben bizonyos értelemben korrigálódik a szabadidő (zömmel a családra fordítható, a családdal együtt tölthető idő) igényének jelzése által.

A *3. táblázat* a szövegekben megjelenített eseményekre koncentrál: az események határpontjaira, kitüntetett jellegére, lineáris vagy ciklikus természetére. Ugyanez a táblázat tartalmazza azokat a kategóriákat is, amelyek a fogalmazó saját minősítését

jelzik a fogalmazási téma, mint alapesemény kapcsán. Az iskola témájú fogalmazások alapeseménye az iskolába járás maga, és ezzel kapcsolatban a menekülési késztetés direkt jelzéseit regisztráltuk. Ugyancsak regisztráltuk az iskolára, szűkebben az iskolai tanulásra, annak értelmére, szükségességére, hasznára, élvezetes voltára utaló megnyilvánulásokat, s ha csak egyetlen egy effélét tartalmazott is a fogalmazás, a pozitív értékelést tartalmazó szövegek kategóriájába soroltuk. A felnőttség témájú szövegek alapeseménye a felnőtté válás maga, és ebben a témakörben a pozitív értékelés értelemszerűen nem az iskolára, hanem a fogalmazó saját majdani felnőtt életére, az azzal kapcsolatos pozitív várakozás megjelenésére utal. Az időhiány kategóriájának adata pedig azt jelzi itt, hogy a fogalmazók hány százaléka gondolja eleve úgy, s fejezi is ki, hogy saját felnőtt életének velejárója lesz az időhiány.

3. táblázat: Egyes tartalmi kategóriák aránya évfolyamonként és témakörönként az érvelő fogalmazásokban III. rész (%)

Tartalomelemzési kategóriák	**Az iskolába járás témája**		**A felnőttség témája**	
	4. o. (N=252)	**8. o. (N=263)**	**4. o. (N=274)**	**8. o. (N=275)**
Kitüntetett események ideje	40,08	32,32	42,70	37,82
A tárgyalt esemény kezdete jelzett	14,29	32,32	88,32	69,62
A tárgyalt esemény vége jelzett	38,49	51,71	6,20	16,73
Lineáris események	49,21	62,74	75,91	96,73
Kronológia	1,59	3,80	2,92	22,18
Ciklikusan ismétlődő események	74,60	70,72	37,59	17,45
Jelenhez tapadó szöveg	26,98	13,69	1,09	1,82
Pozitív értékelés	92,86	68,44	98,54	94,91
Menekülési késztetés	12,30	14,45	15,33	22,91
Konkrét cél	11,90	3,42	55,47	38,91
A fogalmazás terjedelme elegendő	78,97	61,22	64,23	68,73

Az itt bemutatott adatok arra engednek következtetni, hogy a kitüntetett események, illetve a kitüntetett események (például az ünnepnapok az emlékezetes egyedi élmények) kitüntetett ideje témától és korosztálytól függetlenül fontos elemei a fogalmazásoknak, és átlagosan a szövegek 38%-ában előfordulnak. A témaválasztásból adódó alapesemények (iskolába járás, felnőtté válás) azonban különböző mértékben igénylik a kezdet, illetve a befejezés jelzését. Míg a felnőttség a tanulói fogalmazásokban a kezdet felől meghatározott, az iskolába járás valószínűbben fogalmazódik meg a lezárás, a vég felől, különösen nyolcadik osztályban. Szembetűnő kimutatott eltérés az is, hogy míg az iskola téma kifejtése a ciklikus elemek túlsúlyát hozza, a felnőttség tárgyalás inkább lineáris keretet kap. A téma mellett az életkor is szerepet játszik a ciklikus és lineáris elemek megjelenésében. A fiatalabb évfolyamon a ciklikusság nagyobb szerepet kap, mint később, s nyolcadikban a linearitás még kifejezettebbé válik. A lineáris kereten belül szövegek vagy szövegrészek kronologikus kifejtése csak nyolcadikban, s ott is csak a felnőttség témájában válik számottevővé.

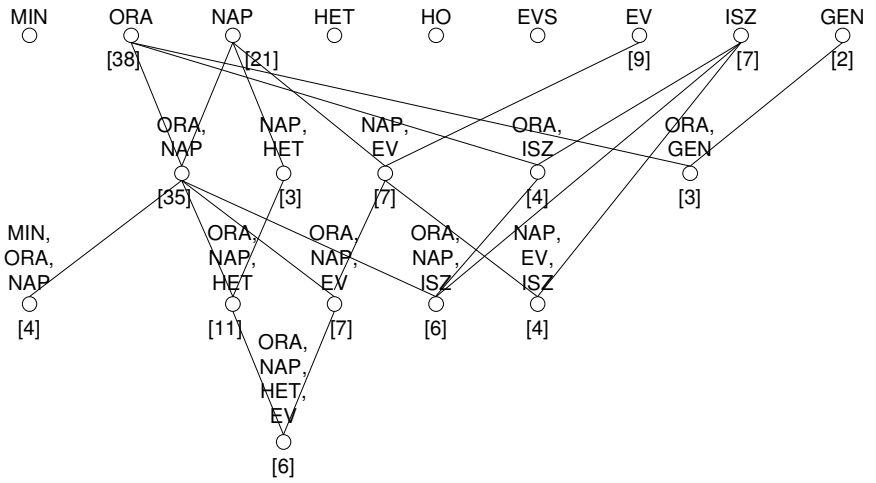
A negyedikesek szövegei konkrétumokban dúsabbak, mint a nyolcadikosokéi, bár meg kell jegyeznünk, hogy ez a konkrétság, mint konkrétan megnevezett pályaelképzelés a felnőttséggel kapcsolatos fogalmazásokban sokszor irreális élettervekben,

irreális foglalkozási elképzelésekben ölt testet. A tanulók jelentős hányada számára ezek a pályaelképzelések is beszűkülnek néhány tipikus választásra (orvos, fodrász, focista), a legkevésbé sem tükrözve a lehetséges foglalkozások akár csak megközelítően széles skáláját sem. A differenciálatlanság abból a szempontból is problémának tűnik, hogy a fogalmazások egy részében egyetlen kategóriaként a jelen fordul elő; a gondolatok kifejtése a jelenhez mint állapotszerű időtlenséghez tapad. Célképzet és perspektíva híján látszik lenni a negyedik évfolyam iskola témájú szövegeinek mintegy negyede. Ennek a kategóriának a visszaszorulásával nyolcadikra az időszemlélet differenciálódni látszik.

Az iskolai tanulás értéke a tanulás fontosságának, hasznának említésében kifejezésre jutó pozitív értékítélet a negyedikesek fogalmazásainak 93 %-ában legalább egyszer megfogalmazódik. Jelentősen visszaesik, 63 %-os az iskolai tanulást pozitív módon (is) megítélő és tanulást pozitívan (is) megélő diákok aránya nyolcadikban. A felnőtt normák elutasítása, az esélytelenség érzése, a tanulás valódi leértékelése egyaránt magyarázat lehet a jelenségre. Lényeges ugyanakkor, hogy a felnőtt életre vonatkozó elképzeléseiben mindkét korosztályra pozitív várakozás jellemző. Ennek alapján meglepő lehet, hogy a felnőttségtől mégis többen félnek, mint az iskolától. Az eszképizmus, a menekülési késztetés az iskolából mindkét évfolyamon hasonló arányú (12-14%-os), viszont a felnőttségtől való félelem az életkorral nő, s a nyolcadikos diákok negyedét-ötödét érintheti.

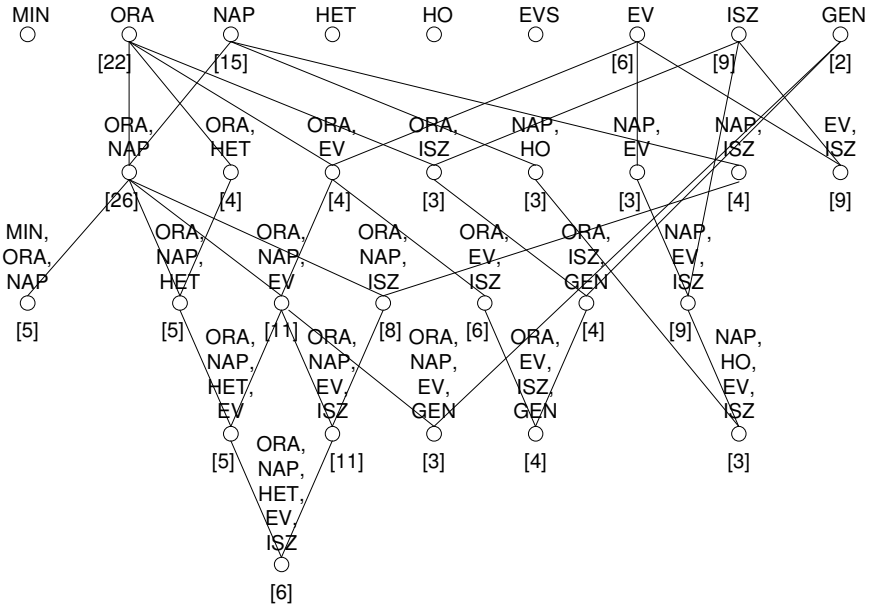
2.2 Lépték és ritmus – strukturális megközelítés

Galois-gráfokkal azt ábrázoljuk, hogy az iskola témájú érvelő fogalmazások teljes negyedikes (N:252) és nyolcadikos (N:263) mintáiban milyen időbeli lépték-jellegű kategóriák milyen kombinációkban, s mekkora elemszámmal fordulnak elő. (A Galois-gráfok leírásáról lásd [4].) A gráfokon csak a többszöri előfordulásokat jelenítjük meg, a számos egyedi variációt figyelmen kívül hagyjuk. A gráfokat mint objektumok és tulajdonságok rendszerét tekintjük, ahol az objektumok az egyes fogalmazások, a tulajdonságok pedig az idő tartamjelző kategóriái a perctől a generációs léptékig. Az ábrázolás adott tulajdonságaiban megegyező objektumok legnagyobb közös halmazát jeleníti meg. Az egyetlen tulajdonságukban közös objektumok jelölését követi a két azonos tulajdonságban közös objektumok és jellemzőik azonosítása, és így tovább. Az ábrázolásban külön-külön szintre kerülnek az egy, két, három stb. tulajdonságukban közös objektumok. Ezek a gráfok elvileg annyi szinten épül(het)nek fel, ahány tulajdonság-kategóriát vettük fel, tehát elvileg kilenc „emeletesek". Ezúttal úgy választottuk meg az ábrázolási módot, hogy fentről lefelé haladva először az egyetlen ritmikai elemet tartalmazó fogalmazások halmazait jelenítettük meg, a legalsó szinten a tulajdonságok legnagyobb számában közös fogalmazások halmazainak jelölései jelennek meg. A tulajdonságokat rövidítéseikkel, az adott halmazhoz tartozó fogalmazások számát pedig kapcsos zárójelben adtuk meg.



1. ábra: Iskolai fogalmazások ritmikai struktúrája – a negyedik évfolyam iskola témájú érvelő fogalmazásainak Galois-gráfja

A vizsgált két évfolyam gráfjainak összevetése első ránézésre is azt mutatja, hogy a nyolcadik évfolyam szövegeinek ritmikai elmei nagyobb komplexitású struktúrába rendeződnek. A 14 évesek fogalmazásai a 10 évesekéhez képest ritmikailag telítettebbek, és a társítási és kombinációs lehetőségeket is jobban kihasználják. Azonban az elvileg lehetséges összes szint egyik évfolyamon sem jeleníthető meg, mivel nincs a fogalmazásoknak olyan halmaza, amely az összes előforduló időbeli léptéket, illetve ritmikai elemet a perctől a generációváltásokig magában foglalná. (A negyedik évfolyamon létezik a négy tulajdonságukban azonos fogalmazásoknak egy hatelemű halmaza, a nyolcadik évfolyamon pedig a közös tulajdonságok legnagyobb realizálódott száma öt.) A gráfok útvonalait követő ábraelemzés az eltérő idői léptékek egymásra épülésére, struktúrába ágyazódására is utal. Az óra és a nap mindkét évfolyamon domináns struktúraképző elem, mind önálló, mind pedig együttes előfordulásuk meghatározó. Negyedikben az óra és a nap együttes előfordulásából „származik” a harmadik és negyedik szintű kombinációk többsége, vagyis az évfolyam szövegeiben az időbeli lépték tartalomfüggő differenciálódása az órához és a naphoz társított újabb ritmikai elemek bevonásával történik. A struktúra legkevésbé integrált eleme a negyedik évfolyamon a generációs lépték. Az ábra mutatja, hogy legfeljebb egyetlen másik ritmikai elemmel társítva fordul elő, ott is alacsony elemszámmal. Amint a táblázatok elemzésekor korábban láttuk, a generációs léptéknek felnőtt témájú fogalmazásokban lesz kitüntetett jelentősége.

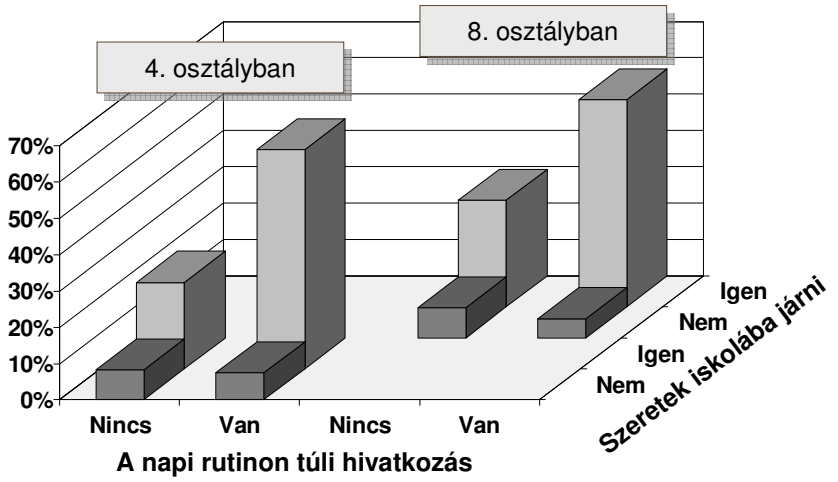


2. ábra: Iskolai fogalmazások ritmikai struktúrája – a nyolcadik évfolyam iskola témájú érvelő fogalmazásainak Galois-gráfja

A nyolcadik évfolyam iskola témájú szöveginek ritmikai struktúráját a negyedikesekéinél intenzívebben alakítják a távlatot kifejező elemek: mind az év, mind az iskolaszerkezeti váltást kifejező, mind a generációs lépték. A nyolcadik évfolyamon ezen elemek mindegyike eljut a struktúra negyedik szintjére, vagyis arra a szintre, ahol három másik ritmikai elemmel együtt fordul elő. Nyolcadikra természetesen nem cserélődnek ki a ritmikai struktúra elemei, hanem a kiindulópontként rögzített, nem távlatos elemek kapcsolatba kerülnek a távlatosságot kifejező elemekkel is. Elhamarkodott dolog volna azonban ennek alapján ezt mint törvényszerű fejlődésképzetet interpretálni.

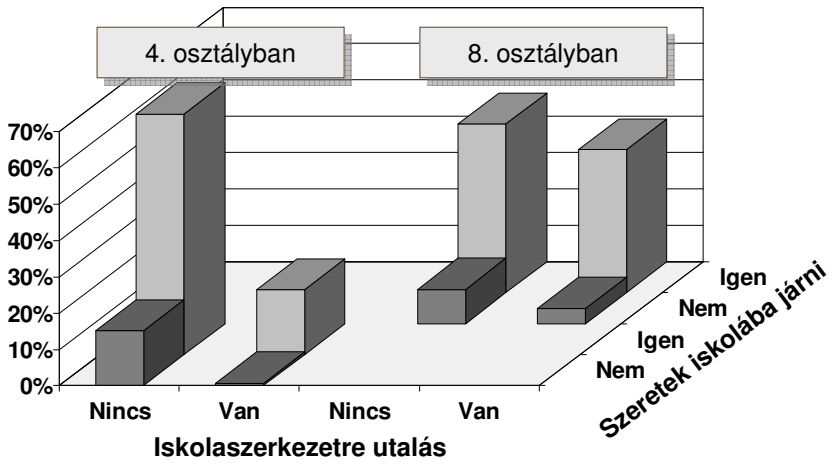
2.3 Az érvelési pozíció mint az iskolai attitűd eleme

A távlatosság tekintetében a következő kérdés, hogy mivel függ össze a gondolkodás, szövegalkotás távlatossága. A 3. ábrán látható diagram azt jeleníti meg, hogy a távlatosság, mint a napi rutinon túlmutató időbeli hivatkozás előfordulása, megoszlása a vizsgált két évfolyamon hasonló az érvelési pozíció mentén bontott részmintákban.. A távlatosság megjelenése ezek körében valószínűbb, akik az iskolába járás mellett érvelnek, vagy kétoldalú bemutatásra törekszenek.



3. ábra: Távlatosság és iskolai attitűd I.

Az is kérdés lehet, hogy a távlatosság tényezőin belül az iskolaszerkezeti váltások említési gyakorisága miként függ össze az életkorral és az érvelési pozícióval. A 4. ábra diagramja alapján általánosságban azt mondhatjuk, hogy negyedikben inkább az említések hiánya jellemző, ugyanakkor a kisebb arányban mégis előforduló említések nem a negatív, hanem, a pozitív érvelést választó tanulók körében fogalmazódnak meg. Nyolcadikban az iskolaszerkezeti váltások említési aránya megnövekszik.

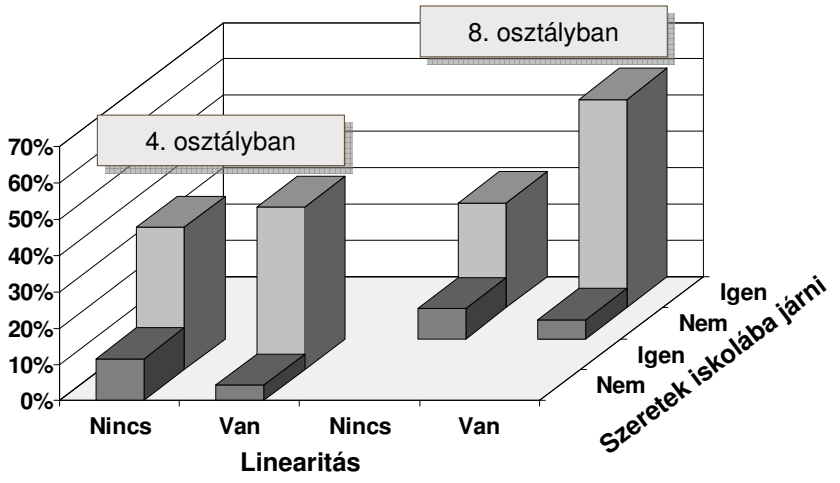


4. ábra: Távlatosság és iskolai attitűd II.

A tanulók iskolai attitűdje, választott pozitív vagy negatív érvelési pozíciója a szövegekben megmutatkozó időkezelés szignifikáns eltéréseit hozza magával. Egyrészt az iskolát pozitív oldalról (is) megközelítő szövegek valószínűbben lépnek túl a napi rutinon, mint a negatív érvelésűek, másrészt az iskolába járás előnyeit (is) taglaló fogalmazásokban ritkán fogalmazódik meg időhiány, míg a negatív érvelésmódot választó írások több mint felében igen. Pozitív érveléssel valószínűbben jár együtt távlatos lépték, konkrétan az iskolaszerkezeti váltások jelzésének használata. A távlatosság valószínűbb az iskola mellett érvelők és a kétoldalú megközelítést adók csoportjában, mint a kontra érvelési pozíciót választók körében, és a negatív érvelésnek valószínűbben lesz argumentuma az időhiány jelzése.

2.4 Linearitás és ciklikusság

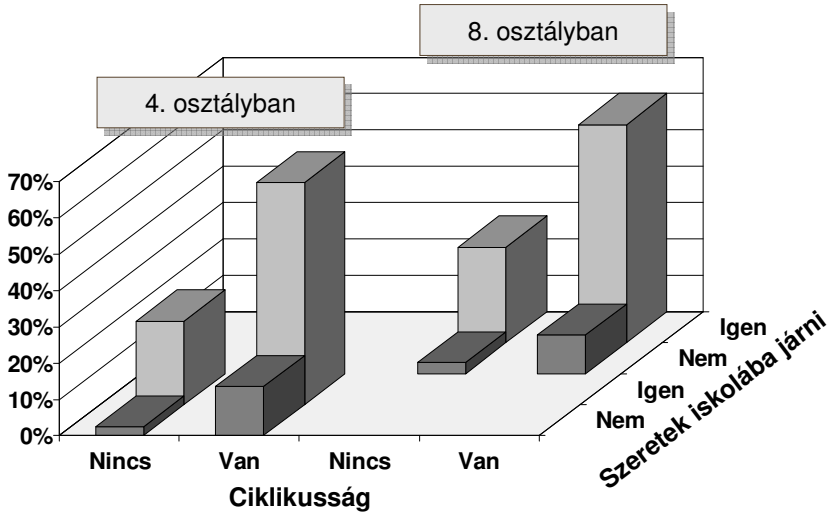
Feltételezve, hogy linearitás és ciklikusság két markánsan különböző időszemlélet kifejeződései, további kérdésünk, hogy vannak-e életkorfüggő elemei a szövegekben megmutatkozó időszemléletnek. Az 5. ábra alapján úgy találjuk, hogy a linearitás tekintetében van különbség a két évfolyam között. Az idő lineáris szerveződésére utaló elemek megjelenése nyolcadikban gyakoribb, emellett az iskolai tanulás fontossága mellett érvelők körében valószínűbb.



5. ábra: Linearitás és iskolai attitűd

A ciklikusság megjelenése azonban nem ilyen képet mutat (6. ábra). A ciklikusság mintázata megdöbbentően azonos az évfolyamok és az érvelési pozíció mentén. Linearitás és ciklikusság kérdésében úgy találtuk, hogy azok körében, akik a fogalmazási helyzetben arról írtak inkább, hogy miért nem szeretnek iskolába járni, szignifikánsan alacsonyabb a linearitást beépítő megoldások aránya. Megjegyezzük, hogy míg a pozitív érvelésű fogalmazások 63 %-ában mutatkozik meg linearitás, a negatív érvelésű szövegek esetében ez az arány csak 27 %. Ennek a ténynek az ismeretében

talán azt gondolhatjuk, hogy a „nem szeretek iskolába járni” típusú szövegek bizonyára hajlamosabbak ciklikus szemléleti elemeket tükrözni, azonban eredményeink szerint ez nem így van.



6. ábra: Ciklikusság és iskolai attitűd

Az idő ciklikus természetével, jelenségek vagy események rendszeres ismétlődésével a negatív érvelésű szövegek 85 %-ában, a pozitív érvelésű szövegek 73 %-ában találkozunk. A különbség nem szignifikáns. Ha tehát azt firtatjuk, hogy van-e összefüggés a pozitív vagy negatív iskolai attitűd mint a fogalmazási helyzetben az érvelési pozíció elfoglalásában megmutatkozó viselkedéses válasz és a szövegekben megjelenő időszemlélet között, azt mondhatjuk, hogy igen, de a csak a linearitás megjelenése tekintetében. Másként úgy is fogalmazhatunk, hogy univerzális szövegelemnek csak a ciklikusság tekinthető.

3 Összefoglalás

Előadásunkban a társadalmi időkezelés és időszemlélet vizsgálatára bevezetett tartalomelemzési kategóriáinkat, az egyes kategóriák előfordulási adatait, valamint az ezekből levonható néhány következtetést ismertettük. Strukturális elemzéssel igyekeztünk bemutatni az időbeli viszonyok szövegbeli megjelenésének differenciálódását, és kiemeltük az érvelési pozíció elfoglalásában megmutatkozó viselkedéses válasz jelentőségét a téma szempontjából. Linearitás és ciklikusság elemzésekor arra a feltevésre jutottunk, hogy míg a ciklikusság stabil minőségnek tűnik, a linearitás tekintetében jelentős szemléleti változás történik az életkor előrehaladtával az iskolai szocializáció folyamatában.

Bibliográfia

1. Cs. Czachesz Erzsébet – Csirik János: 10-16 éves tanulók írásbeli szókincsének gyakorisági szótára. BIP, 2002.
2. Huszár Zsuzsanna – Sramó András: Idői struktúrák feltárása kvalitatív és kvantitatív szövegelemzéssel. In. I. Magyar Számítógépes Nyelvészeti Konferencia. Szerk.: Alexin Zoltán és Csendes Dóra, SZTE Informatikai Tanszékcsoport. Szeged, (2003) 225-230. o.
3. Huszár Zsuzsanna – Sramó András: Az iskolai idő értékelése nyolcadik osztályosok érvelő fogalmazásainak tartalomelemzése alapján. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szerk: Alexin Zoltán és Csendes Dóra, SZTE Informatikai Tanszékcsoport. Szeged, (2004) 227-229, 351. o.
4. Takács Viola: Galois-gráfok pedagógiai alkalmazása. Iskolakultúra- könyvek 6. Iskolakultúra, Pécs, (2000)

Élettörténeti traumákról szóló rövid beszámolók idői szerveződésének vizsgálata az INTEX tartalomelemző szoftverrel

Ehmann Bea[1]

[1] MTA Pszichológiai Kutatóintézet, 1132 Budapest, Victor Hugó u. 18-22
ehmannb@mtapi.hu

Kivonat: A tanulmány a számítógépes pszichológiai tartalomelemzés egy új módszerét mutatja be. A szerző az INTEX nyelvi fejlesztőprogramot alkalmazta angol nyelvű elbeszélt traumatikus élmények idői szerkezetének feltárására. A feladat keretében készített gráf által automatikusan nyerhető konkordanciák alapján az idői szerkezet mintázata SPSS grafikonok formájában vizuálisan ábrázolható. A grafikonok elemzéséből számos pszichológiai következtetés vonható le. Egyebek közt az, hogy annál traumatikusabb hatású egy-egy elbeszélt esemény, minél nagyobb amplitúdójú és minél töredezettebb a grafikon, illetve, ha az elbeszélt történet nem a múltban, hanem a jelenben vagy present perfect idősíkban végződik.

1 Előzmények és háttér

Az elbeszélt idő jellegzetességeit hagyományosan az önéletrajzi emlékezetkutatás, újabban pedig a narratív pszichológia paradigmájának keretében is vizsgálják. A narratív pszichológiai kutatás módszere a pszichológiai tartalomelemzés egy speciális formája, a narratív pszichológiai tartalomelemzés [9].

A két székhelyű – Pécsi Tudományegyetem és MTA Pszichológiai Kutatóintézet – narratív pszichológiai kutatócsoport jelen számítógépes nyelvészeti konferencián történő részvételének legitimitását az adja, hogy a pszichológiai, s ezen belül a narratív pszichológiai tartalomelemzés számítógépes szoftvereket alkalmaz.

A Morphologic Kft-vel közösen kifejlesztett LINTAG nevű program segítségével nyert eredményeinkről az MSZNy 2004 című konferencián számoltunk be [4, 7, 8]. A narratív pszichológiai munkacsoport eddigi tartalomelemző arzenáljának (LIWC + ATLAS.TI + LINTAG) negyedik eleme – negyedik történeti ugrása – az INTEX világába történő belépés. Jelen tanulmány az INTEX szoftver pszichológiai felhasználójaként kapott eredményeket ismerteti.

2 Az INTEX alkalmazása az elbeszélt idő szerveződésének vizsgálatában

2.1 A vizsgálat célkitűzései

Korábbi szövegvizsgálatokban empirikusan igazoltam, hogy az élettörténeti (ezen belül a traumatikus eseményeket elbeszélő) beszámolók nem lineáris-naptári, hanem ún. narratív kronológiát követnek, azaz az elbeszélés nem az egyszerű múlt idő síkján, hanem mintegy hurkokat leírva, a régmúltba, a félmúltba, a jelenbe és a jövőbe kanyarogva halad [1, 2, 3, 5, 6]. Hipotézisem szerint az elbeszélés módjának idői szerveződéséből következtetések vonhatók le arra, hogy az elbeszélt trauma milyen pszichológiai jelentőséggel bír az elbeszélő számára. A tágabb keretbe illeszkedő jelen vizsgálat célja kettős: (1) az INTEX program segítségével gráfokat létrehozni angol nyelvű, rövid terjedelmű traumatikus elbeszélések idői szerkezetének elemzésére; (2) az így elvégzett elemzések alapján feltárni a traumatikus elbeszélések idői szerkezeti típusait, megalkotni ezek taxonómiáját.

2.2 A minta

Az jelenleg is folyó elemzéseket a James W. Pennebakertől (University of Texas at Austin, USA) kapott, kontrollált körülmények között felvett, több száz, angol nyelvű traumatikus elbeszélés által alkotott szövegkorpuszon végzem.

2.3 Az INTEX

A Max Silberztein által Franciaországban 1993-ban megalkotott INTEX számítógépes nyelvészeti fejlesztő rendszer alkalmazóinak köre, az ún. INTEX Community tíznél több országban száznál is több fejlesztőből és felhasználóból áll (Silberztein, 2001). Az INTEX, illetve ennek újabb változata, a NooJ magyar változatának kifejlesztését célzó munkálatokat akadémiai társintézetünkben, az MTA Nyelvtudományi Intézetében Váradi Tamás vezetésével a Korpusznyelvészeti Osztály munkatársai végzik [10, 11, 12].

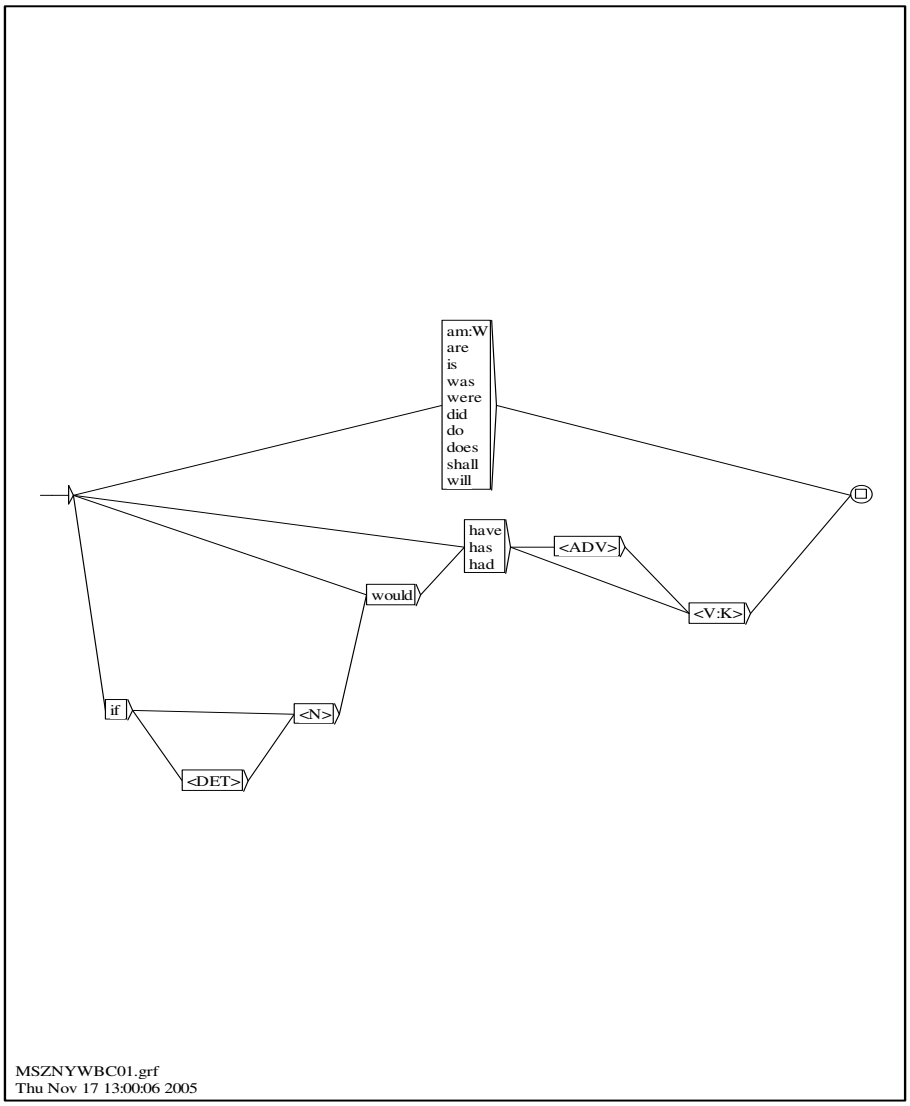
A szoftvert Váradi Tamás bocsátotta rendelkezésünkre, akinek a program használatához nélkülözhetetlen szakmai segítségét ezúton is köszönöm.

2.4 A módszer

Az általam kifejlesztett módszer két összetett lépésből áll: (1) INTEX gráf/ok/ készítése az elbeszélések idői szerveződésének vizsgálatára; (2) a gráf révén kapott konkordanciák mintázatának grafikus megjelenítése az SPSS segítségével.

Az INTEX gráf arra szolgál, hogy a különféle igeidőkre utaló nyelvi markerek konkordanciája a szöveg szerinti előfordulás sorrendjében kilistázhatóvá váljon.

A konkordancia lista úgy válik az SPSS-ben grafikusan ábrázolhatóvá, hogy az egyes idősíkokra utaló nyelvi jelek számértékeket kapnak (pl. simple past = 0, past perfect = -1, present perfect = 1, simple present = 2, stb.). A grafikon alapvonala a simple past idősík.



1. ábra. Az INTEX szoftverrel készített gráf

Ha a mindennapi életben a velünk történt eseményeket lineáris módon beszélnénk el, akkor az elbeszélésben egymás után kibontakozó mondataink előrehaladó irányban – mintegy öltésről öltésre haladva – rávarródnának a naptári vagy óraidő tengelyére. A narratív kronológia útvonala azonban fel-le irányban, változó amplitúdóval eltér ettől az alapvonaltól. Ekképp megkapjuk az idői szerkezet vizuális mintázatát.

Ez az eljárás megteremti az alapját annak, hogy az egyes elbeszélések egyedi idői szerveződési mintázatai a későbbiekben objektíven összehasonlíthatóak és típusokba sorolhatóak legyenek.

2.5 A Gráf

Az idősíkok kijelzésére szolgáló fenti gráf ez idáig az első működésképes prototípus, a későbbiekben nyilván változni fog. A gyakorlat azt mutatja, hogy nem szükséges a simple past alapvonalon futó múlt idejű igék összességét megkeresnünk a konkordancia számára – ez ugyanis a mintázaton nem változtat, ám kezelhetetlenül szélessé teszi a grafikont. Illusztrációképpen beletettem a feltételes módra vonatkozó egyik keresést is, ennek indoklását lásd a megvitatásban.

3 Eredmények

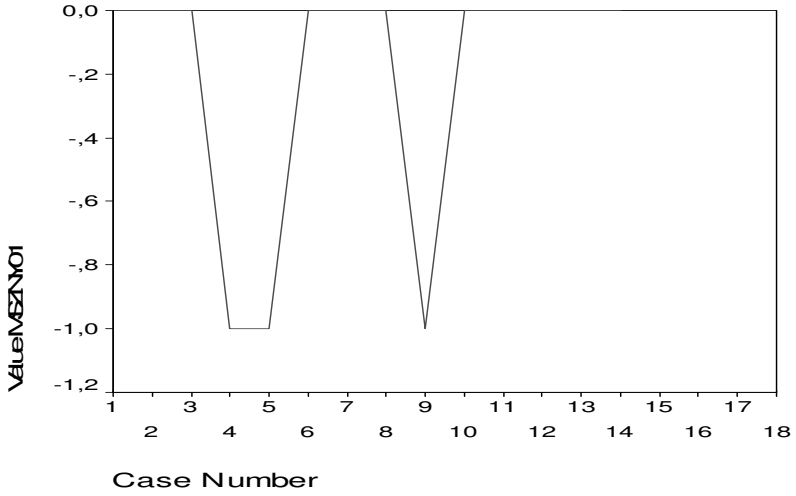
Az alábbiakban három tipikusnak tekinthető példát mutatok be. Egyenként sorra következnek az eredeti szövegek, az INTEX által generált konkordanciák és az SPSS grafikonok.

3.1 Feldolgozott veszteségtörténet – A nagypapa halála

A trauma feldolgozottságára utaló szerkezeti jegynek azt tekintettem, hogy a történet a múlt egy konkrét pontján kezdődik, és a múlt idősíkján is fejeződik be. A történetekben nem tematizálódik a jelen idősíkja, ám jelen vannak a past perfect-re, valamint a third conditional-ra történő utalások.

> « In 1978 my Grandfather was diagnosed to have diabetes. In April 1979 my mother and her sisters and brother decided to have Grandpa checked out by another doctor. Grandpa was found to have cancer. In the hospital, he did not know and he trusted that he had diabetes. The cancer had spread throughout his pancreas, liver, colon, and esophagus. Only if his doctor would have been more careful.
> Grandpa fought like a soldier. He would stay happy so we would. My parents told my sister, brother and me even though the other grandchildren were not told.
> In the summer my family returned to Florida to see him. The hearty man I knew was now 100 pounds. I returned again to find a 90 pounds man who could not walk on his own. He could not lift our spirits like he did before.
> On July 5, 1979 I knew his time had come. I walked to his house even though I had orders to stay at my cousin's house. After falling asleep my mother woke me to tell me Grandpa was gone. I could not grasp it. They took him away and I just stood outside looking down the road. He was gone. He was gone forever.
> Grandpa was only about 67 years old. It seem things happen to the best people. I was the only grandchild out of 12 that stayed with Grandpa towards the end. It was an honor. »

In 1978 my Grandfather was diagnosed to have diabetes. In another doctor. Grandpa was found to have cancer. In the h er. In the hospital, he did not know and he trusted that he ad diabetes. The cancer had spread throughout his pancreas, on, and esophagus. Only if his doctor would have been more the other grandchildren were not told. In the summer my m. The hearty man I knew was now 100 pounds. I returned aga lift our spirits like he did before. On July 5, 1979 I kn 5, 1979 I knew his time had come. I walked to his house ev ke me to tell me Grandpa was gone. I could not grasp it. T oking down the road. He was gone. He was gone forever. road. He was gone. He was gone forever. Grandpa was o one forever. Grandpa was only about 67 years old. It se to the best people. I was the only grandchild out of 12 t dpa towards the end. It was an honor. 1



2. ábra. A feldolgozott veszteségtörténet idői szerveződési mintázata

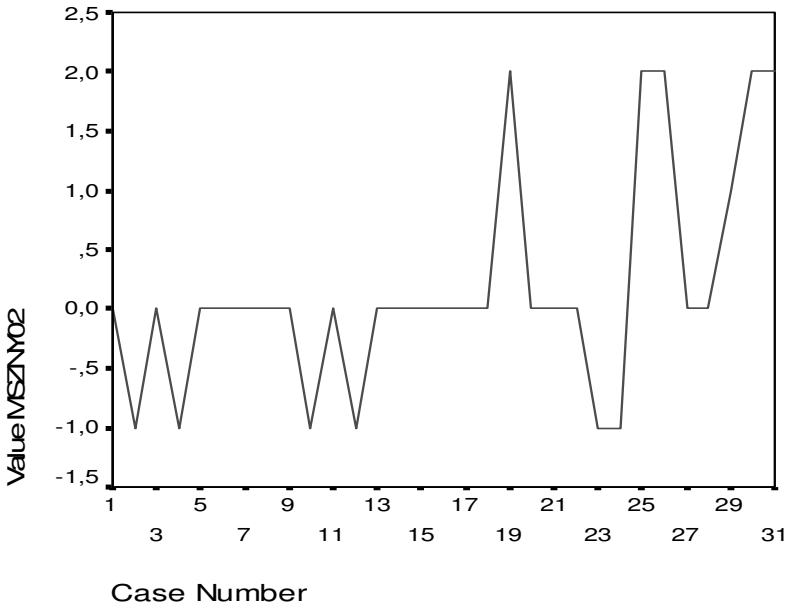
3.2 Jelenben is ható múltbéli trauma – Autóbaleset látványa

A történet egy lényeges jegyben különbözik az előzőtől: a múlt egy konkrét pontján kezdődik, ám a jelen idősíkján fejeződik be. Az esemény sokszoros átgondolásának jeleként itt is, csakúgy, mint az előző történetben, jelen vannak a past perfect-re, valamint a third conditional-ra történő utalások.

« I was only 9 years old. I had left my home in California to go to my Uncle's house in Mound Minnesota. A neat place right on the lake. I was going there for Uncle Bills camp. Something my siblings had done and it was now my turn. A time to learn some things I did not get a chance to at home. A wonderful trip but with an experience I can describe. My cousin lived there as well and worked at the local fast food chain. Uncle Bill, Aunt Anne and myself jumped into Rosey, a 69 Dodge Dart in shining red. We got up to the restaurant and went in and saw

my cousin Sally. She was working and we ordered a big dinner. Then I was a bit hungry again. As we left, Uncle Bill went back and we got an apple pie. Now we got in Rosey and headed home. About 100 yards up the four lane road we stopped to witness the very first moments of the victims of a car accident. We were there first, the only civilians helping. We pulled over and Uncle Bill jumped out with Aunt Anne. They told me to lie down in the back. I had seen the accident already. Blood was dripping out of the car door. High pitched screams of pain from the passengers side in the car. I had seen these bodies. A 4 seater 2 door hatchback with 4 people inside. The two in back were crushed into the back of the car. The ones in front were lying on the dash trying to move. They could not get out. They could not move. The car was a ball of steel. I looked up to see Uncle Bill pulling one of the front seat, I think the driver, out of the car. He was covered in blood screaming at the top of his lungs, talking about his friends in the back. I was terrified. I lay back down and grabbed the small pillow that sat in the back seat and hugged it for security. The sound of screams and cars was terrifying. I do not remember sirens. Suddenly the door opened and Uncle Bill told me he needed the pillow. I gave it to him, my last security blanket, and jumped to the seat again. I was crying, terrified, and nervous. I did not know what to do. I kept thinking that if I had not gotten the apple pie we would have seen the accident - the compression of bodies - the splattering of blood. I do not remember hearing sirens or seeing rescue people. I do not remember driving home. I just know that I sat awake for 3 nights terrified. I remember repeatedly seeing the crash behind the driver. There was glass everywhere. I remember more people somewhere and the car was white. I have seen that scene again in my mind when I am very tired. I do not know what happened to any of them. I do not want to know. »

```
.....................    I was only  9 years old.  I had left
as only  9 years old.  I had left my home in California to g
ace right on the lake. I was going there for Uncle Bills cam
p. Something my siblings had done and it was now my turn.  A
siblings had done and it was now my turn.  A time to learn s
e to learn some things I did not get a chance to at home. A
saw my cousin Sally. She was working and we ordered a big di
ed a big dinner.  Then I was a bit hungry again.  As we left
ms of a car accident. We were there first, the only civilian
 lie down in the back. I had seen the accident already.  Blo
accident already.  Blood was dripping out of the car door.
gers side in the car.  I had seen these bodies.  A 4 seater
inside.  The two in back were crushed into the back of the c
 car.  The ones in front were lying on the dash trying to mo
 could not move. The car was a ball of steel.  I looked up t
ver, out of the car.  He was covered in blood screaming at t
 friends in the back.  I was terrified.  I lay back down and
ound of screams and cars was terrifying.  I do not remember
 cars was terrifying.  I do not remember sirens. Suddenly th
ed to the seat again.  I was crying, terrified, and nervous.
errified, and nervous. I did not know what to do.  I kept th
. I did not know what to do.  I kept thinking that if I had
 kept thinking that if I had not gotten the apple pie we wou
 gotten the apple pie we would have seen the accident - the
 splattering of blood. I do not remember hearing sirens or s
seeing rescue people.  I do not remember driving home. I jus
ehind the driver.  There was glass everywhere.  I remember m
le somewhere and the car was white. I have seen that scene a
and the car was white. I have seen that scene again in my mi
hen I am very tired.   I do not know what happened to any of
pened to any of them.  I do not want to know.  1
```



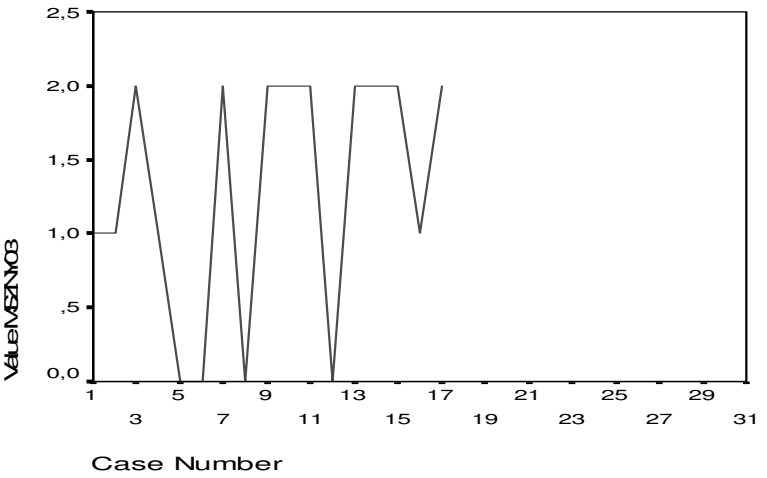
3. ábra. A jelenben is ható trauma idői szerveződési mintázata

3.3 Jelenben is zajló trauma – Iszákos apa

A régebben kezdődött, és a jelenben is zajló trauma idői tartománya ez esetben több, mint tíz évet ölel fel. Az elbeszélés mintázata a „félmúltban", a „befejezett jelenben" kezdődik, és a jelenben végződik. Az asszociatív kapcsolatok a múlt egy-egy konkrét pontja és a jelen között ívelnek át. A történet nem epizodikus, hanem generikus emlék: az összkép sok apró eseményből áll össze.

« Since we moved to our new house over 10 years ago there has been a change in my father. He has always been a good and generous man and still is. But since we moved he has started to drink more and more. He was even told to stop or he would kill himself and he did stop for a few months.
When he drinks he does not get happy or "crazy" like a child would say have a great time. He gets spiteful and mean. He never hit any of us he just got mean.
Sometimes I thought it was me. When my brother would come home he would not drink a thing. But I came home from school one summer he would get wasted and spiteful. He will pick anything I say apart. I can say something nice and if he is drunk will turn it around. Sometimes I figured he would rather be at work because once he got home he would pour a drink.
His drinks were not nice tasting because he would have so much liquor in them they would have no taste. He does not always drink now but when he does it is hard on all. Now that the kids are gone my mother must catch the brunt of it. She does not deserve that. »

```
ver 10 years  ago there has been a change in my father.
ange in my  father.  He has always been a good and gene
generous man  and still is. But since we moved he has s
But  since  we moved he has started to drink more and m
drink more and more. He was even told to stop or he wou
uld kill himself and he did stop for a few months.    W
ths.  When he drinks he does not get happy or "crazy" l
Sometimes  I thought it was me. When my brother would c
asted and  spiteful. He will pick anything I say art. I
mething  nice and if he is drunk will turn it n say som
nice and if he is drunk will turn it around.
 a drink.    His drinks were not nice tasting or in the
would have no taste. He does not always drink now e doe
drink   now but when he does it is hard on all. Now tha
but  when  he   does it is hard on all. Now that the ki
all. Now that the  kids are gone my mother must catch t
  the brunt of it. She does not deserve that.  1
```



4. ábra.A jelenben zajló trauma idői szerveződési mintázata

4 Megvitatás és kitekintés

A tanulmány a számítógépes pszichológiai tartalomelemzés egy új módszerét mutatja be. Az INTEX nyelvi fejlesztőprogramot alkalmaztam angol nyelvű elbeszélt traumatikus élmények idői szerkezetének feltárására. A feladat keretében készített gráf által automatikusan nyerhető konkordanciák alapján az idői szerkezet mintázata SPSS grafikonok formájában vizuálisan ábrázolható.

A grafikonok elemzéséből számos pszichológiai következtetés levonható. Egyebek közt az, hogy annál traumatikusabb hatású egy-egy elbeszélt esemény, minél nagyobb amplitúdójú és minél töredezettebb a grafikon, illetve, ha a történet a jelenben vagy present perfect idősíkban végződik.

Az idői szerkezet ábrázolásában számos nyitott kérdés létezik. Az egyik, hogy a különböző típusú feltételes mondatok lehorgonyozhatóak-e a régmúlt, a múlt, a jelen vagy a jövő idősíkjára, avagy inkább megszakítják a gondolatáramlás idői menetét. A második, hogy elválasztható-e egymásról az elbeszélt esemény főszereplőinek illetve az elbeszélőnek az idői világa. Ez a kérdés az önreferencia kérdéskörébe tartozik. A harmadik kérdés a traumák tényleges megtörténte és a jelen közötti távolság ábrázolhatósága – a klinikai pszichológia álláspontja szerint ugyanis a trauma a hatását tekintve lényegében időtlen, azaz – egy metaforával élve – mintegy betokozódva lebeg a naptári vagy óraidő tengelyén kívül.

Az eljárás segítségével a közeljövőben több száz rövid traumatikus epizódot tervezek feldolgozni. E kutatás révén árnyaltabb felbontású képet kaphatunk arról, hogyan befolyásolják a múltbeli események az elbeszélő jelenét.

Az INTEX-szel kapcsolatos további terveket illetően: jelenleg folyó projektünk keretében az MTA Nyelvtudományi Osztályának Korpusznyelvészeti Osztálya által fejlesztett INTEX/NooJ segítségével magyar nyelvű szövegeken is a fentihez hasonló elemzéseket szeretnék végezni.

Bibliográfia

1. Ehmann B. (2004): A pszichológiai idő kutatásáról. *Pszichológia, 2004, 24, 4 : 317-324*
2. Ehmann B. (2004): A szubjektív időélmény mintázatainak pszichoanalitikus és narratív pszichológiai párhuzamai. *Pszichológia, 24, 4, 403-425*
3. Ehmann B. (2004): Laikus történetek időstruktúrája. In: László, J. Kállai J. és Bereczkey T. (Szerk.): *A reprezentáció szintjei.* Gondolat Kiadó, Budapest. pp. 356-371.
4. Ehmann Bea (2004): A LAS VERTIKUM időmodulja. In: *II. Magyar számítógépes nyelvészeti konferencia, Szeged 2004. Cikkgyűjtemény.* 257-261.
5. Ehmann Bea (2004): Tartalomelemzési módszerek a szubjektív időélmény vizsgálatára laikus beszélők szövegeiben. In: Szerk.: Erős Ferenc: *Az elbeszélés az élmények kulturális és klinikai elemzésében*. Pszichológiai Szemle Könyvtár 8. Akadémiai Kiadó, Budapest, 57-73.
6. Ehmann Bea (2004): Time related word categories and narrative patterns in life story recalls. In: *7th International Conference on Philosophy, Psychiatry and Psychology, „Time, Memory and History". Heidelberg, September 23-26, 2004. Absztrakt kötet*, 31-32.
7. Ehmann Bea, Kiss Balázs, Naszódi Mátyás, László János (2005): A szubjektív időélmény tartalomelemzéses vizsgálata. A LAS Vertikum időmodulja. *Pszichológia*, 2005/2. 133-142.
8. László J., Ehmann B. (2004): A narratív pszichológiai tartalomelemzés új eljárása: A LAS Verticum. In Erős F. (szerk.): *Az elbeszélés az élmények kulturális és klinikai elemzésében.* Budapest: Akadémiai Kiadó. 75-87. (Pszichológiai Szemle Könyvtár)
9. László János (2005): *A történetek tudománya. Bevezetés a narratív pszichológiába.* Budapest, Új Mandátum Könyvkiadó.
10. Oravecz Cs., Varasdi K. és Nagy V. (2004): Többszavas kifejezések számítógépes kezelése. *MSZNY 2004. II. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 141-154.
11. Vajda P., Nagy V., Dancsecs E. (2004): A Ragozási szótártól a NooJ morfológiai moduljáig. *MSZNY 2004. II. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 183-190.
12. Váradi Tamás és Gábor Kata (2004): A magyar INTEX fejlesztéséről. *MSZNY 2004. II. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 3-10.

Az élettörténeti narratív perspektíva modul angol nyelvű változatának fejlesztése

Pólya Tibor

MTA Pszichológiai Kutatóintézet
1132 Budapest, Victor Hugo u. 18-22.
polya@mtapi.hu

Kivonat: Az előadásban bemutatom a narratív perspektíva fogalmát, és tárgyalom a fikcionális, egyes szám harmadik személyű elbeszélőtől származó narratívum illetve a nem fikcionális, egyes szám első személyű elbeszélőtől származó (élettörténeti) narratívum perspektívája közötti különbséget. Korábbi munkámban az élettörténeti narratív perspektíva három formáját határoztam meg (visszatekintő, átélő és metanarratív). Bemutatom a narratív perspektíva automatikus kódolására korábban kifejlesztett eszközöket, illetve részletesen tárgyalom az angol nyelvű élettörténeti narratívumok perspektíváját kódoló modul fejlesztésének lépéseit és eredményeit.

1 Bevezetés

A narratív perspektíva a narratív szöveg strukturális jellemzője. A fogalom arra a jelenségre utal, hogy a történet elbeszélőjének lehetősége van arra, hogy a narratívum tartalmát adó elemeket (eseményeket, szereplőket és körülményeket) valamely meghatározható nézőpontból mutassa be.
A jelenség legtöbb elemzője a nézőpont két változatát különbözteti meg. A szereplő nézőpontjának érvényesítésekor a történet elbeszélője úgy mutatja be a narratív elemeket, ahogy azokat az adott szereplő a múltbeli események során észlelte, mondta, vagy gondolta. Az elbeszélő nézőpontjának érvényesítésekor pedig úgy mutatja be a narratív elemeket, ahogy azokat az elbeszélés helyzetében értelmezi.
A nézőpont érvényesítésének két összetevője különböztethető meg: a nézőpont egyrészt meghatározható térben-időben, másrészt valamely szereplőhöz kapcsolódóan [például 9, 11, 14]. Mindkét összetevőnek számos nyelvi elemzése van [például 1, 2], ezek az elemzések azonban csak az egyes szám harmadik személyű elbeszélőtől származó fikcionális narratívumok perpspektivizációs jelenségeit tárgyalják.

2 Az élettörténeti narratívum perspektívája

2.1 Az élettörténeti narratív perspektíva meghatározása

A nem fikcionális élettörténeti narratívum perspektivizációs lehetőségei azonban jelentősen különböznek az egyes szám harmadik személyű elbeszélőtől származó fikcionális narratívumétól. Amikor az egyes szám harmadik személyű elbeszélő valamely szereplője nézőpontját érvényesíti mindig egy másik személy perspektívája érvényesül. Az élettörténeti narratívum elbeszélője esetén azonban ez a személybeli különbözés nem szükségszerű, hiszen az elbeszélő személy saját korábbi nézőpontját is érvényesítheti. Ebben az esetben a nézőpont két változata csak időbeli elhelyezésükben különbözik minden esetben egymástól: az elbeszélő nézőpontja az elbeszélés jelenében, az elbeszélővel azonos szereplő nézőpontja pedig az elbeszélt események múltjában rögzített. Természetesen az élettörténeti narratívum elbeszélője is érvényesítheti valamely másik szereplő nézőpontját. Ebben az esetben fennáll a személybeli különbség, de a két nézőpont időbeli elhelyezése is különbözik: az elbeszélő nézőpontja szintén az elbeszélési események jelenében, az elbeszélővel nem azonos szereplő nézőpontja pedig az elbeszélt események múltjában rögzített.

Az élettörténeti narratívum perspektiváján a nézőpont és a narratív elemek időbeli viszonyát értem. Az idői különbség fontossága miatt az élettörténeti narratív perspektíva meghatározásának alapja az idői elhelyezés. A meghatározás másik fontos jellemzője, hogy a narratív perspektíva a nézőpont és a tartalom közötti kapcsolatra vonatkozik, azaz nem korlátozódik csak a nézőpontra.

2.2 Az élettörténeti narratív perspektíva három formája

Azt feltételezve, hogy a nézőpont és a narratív tartalom idői elhelyezése két értéket vehet fel (az elbeszélt események múltja versus az elbeszélési események jelene) az élettörténeti narratív perspektíva három formája írható le (lásd 1. Táblázat) [6]. Visszatekintő perspektíva forma esetén a nézőpont az elbeszélési események jelenében, a tartalom az elbeszélt események múltjában rögzítettek. Átélő perspektíva forma esetén a nézőpont és a tartalom is az elbeszélt események múltjában rögzítettek. Végül metanarratív perspektíva forma esetén a nézőpont és a tartalom is az elbeszélési események jelenében rögzítettek.

1. Táblázat: Az élettörténeti narratív perspektíva három formája

Élettörténeti narratív perspektíva	Nézőpont idői elhelyezése	Narratív elemek idői elhelyezése
Visszatekintő forma	Jelen	Múlt
Átélő forma	Múlt	Múlt
Metanarratív forma	Jelen	Jelen

2.3 Az élettörténeti narratív perspektíva formák nyelvi meghatározása

Számos elemzés mutatott rá arra, hogy az elbeszélő és a szereplő nézőpontjának érvényesítéséhez eltérő nyelvi jegyek kapcsolódnak [például 2]. Jelen megközelítés az élettörténeti narratív perspektíva formák nyelvi meghatározásában a nézőpont és a tartalom idői elhelyezését biztosító, illetve az ezzel szisztematikus összefüggést mutató nyelvi jegyeket azonosítja. Mint minden narratívumban, az élettörténeti narratívumban is két módja van az idői elhelyezésnek. Az elbeszélő személy egyrészt megadhatja dátum kifejezésekkel a tartalom vagy a nézőpont idői elhelyezkedését, például *2001 szeptember 11-én történt…* (tartalom idői elhelyezése), vagy például *2005 decemberében azt gondolom…* (nézőpont idői elhelyezése). Az elbeszélő személy azonban idő deiktikus nyelvi jegyekkel is megadhatja a tartalom és a nézőpont idői elhelyezkedését, például *Akkor történt…* (tartalom idői elhelyezése), vagy például *Most azt gondolom…* (nézőpont idői elhelyezése). Az idői elhelyezés mindkét módja képes valamely egyedi időpont azonosítására az idő intervallum skáláján. A dátum kifejezéseket használó idői elhelyezés idő skálák (például évszámok, hónapok) alapján azonosítja az egyedi időpontot, és minden egyedi időponthoz az idő skálák más-más értékeit rendeli. Az idő deiktikus nyelvi jegyekkel történő idői elhelyezéshez nem szükséges az idő skálák használata, ezek hiányában ez a mód csak két értéket használhat valamely egyedi időpont azonosításában: a nézőpont a tartalomhoz képest közelikét (például *most*) vagy távoliként (például *akkor*) lokalizált. Jóllehet az idő skálákra építő elhelyezési mód kidolgozottabb, az idő deiktikus nyelvi jegyek jóval gyakrabban fordulnak elő az élettörténeti narratívumban, ezért a narratív perspektíva formák nyelvi meghatározásában is nagyobb szerepük van.

Az élettörténeti narratív perspektíva formák meghatározása a nyelvi jegyek négy csoportját foglalja magában (lásd 2. Táblázat). A legfontosabb csoport az idő deiktikus nyelvi jegyek (igeidő és idői határozószavak) elemzése. A leggyakrabban előforduló idő deiktikus nyelvi jegy az igeidő. A múlt idő távoliként, a jelen idő közeliként lokalizálja a nézőpontot a narratív tartalomhoz képest. Emellett bizonyos idői határozószavak is képesek közeliként (például *most, ma*) vagy távoliként (például *akkor, tegnap*) lokalizálni a nézőpontot és a narratív tartalmat.

Számos olyan nyelvi jegy van, amely szisztematikusan kapcsolódik a nézőpont és a narratív tartalom időben közeli vagy távoli elhelyezéséhez. Ennek részben az az oka, hogy a nézőpont és a narratív tartalom nemcsak időben, de térben is lokalizált. Mivel feltehető, hogy a nézőpont és a tartalom lokalizációjának két dimenziója – idő és tér – összefügg egymással (lásd például [12]), a hely deiktikus nyelvi jegyek is része a narratív perspektíva formák nyelvi meghatározásának. Bizonyos helyhatározó szavak és mutató névmások is képesek egymáshoz közeliként (például *itt, ez*) vagy távoliként (például *ott, az*) lokalizálni a nézőpontot és a narratív tartalmat.

A nyelvi meghatározás harmadik csoportját azok a nyelvi kifejezések adják, amelyek az egyes narratív perspektíva formákhoz kapcsolódnak. A dátum kifejezések (például *2001, szeptember*) a visszatekintő narratív perspektíva formához kapcsolódnak, mivel a dátum kifejezések jelentése a múltra utal. Az indulatszavak (például *hopp, hűha*) az átélő narratív perspektíva formához kapcsolódnak, mivel az ilyen megnyilatkozások nézőpontja és a megnyilatkozás eseménye egymáshoz közeliként lokalizált [15]. A meta-narratív perspektíva formához a szubjektív modalitást kifejező módosítószók (például *esetleg, talán*), módosító határozószók (például *valószínűleg,*

feltehetőleg) és módosító mondatrészletek (például *úgy tudom, azt hiszem*) kapcsolódnak, amelyek az elbeszélő személy aktuális értékelését hangsúlyozzák [3].

A nyelvi jegyek negyedik csoportja a mondat módja. Visszatekintő perspektíva formát érvényesítő mondat módja csak kijelentő lehet, az átélő és a metanarratív perspektíva formák esetében bármilyen lehet a mondat módja.

2. Táblázat: Az élettörténeti narratív perspektíva formák nyelvi jegyei magyar nyelvű szövegben

Nyelvi jegyek	Visszatekintő forma	Átélő forma	Metanarratív forma
1. Idő deixis			
Igeidő	Múlt	Jelen	Jelen
Deiktikus kifejezések	Pl. *Akkor*	Pl. *Most, ma*	Pl. *Most, ma*
2. Hely deixis			
Deiktikus kifejezések	Pl. *Ott*	Pl. *Itt*	Pl. *Itt*
Mutató névmások	Pl. *Az*	Pl. *Ez*	Pl. *Ez*
3. Specifikus kifejezések	Dátum kifejezések Pl. *Szeptember*	Indulatszavak Pl. *Jaj*	Módosítószók Pl. *Esetleg* Módosító határozószók Pl. *Valószínűleg* Mentális igék Pl. *Tudom*
4. Mondat módja	Kijelentő	Nincs megkötés	Nincs megkötés

3 Az élettörténeti narratív perspektíva modul fejlesztése

3.1 Előzmények

A narratív perspektíva automatikus kódolására elsőként Wiebe [14] dolgozott ki algoritmust. Az algoritmus a nézőpont szereplőhöz kapcsolódó összetevőjére korlátozódik és egyes szám harmadik személyű narrátortól származó narratív szöveg nézőpontjának azonosítását végzi. Az algoritmus két fő összetevőből áll. Az első lépés az elbeszélő nézőpontját érvényesítő (Wiebe kifejezésével objektív) mondatok és a szereplő nézőpontját érvényesítő (szubjektív) mondatok megkülönböztetése. A második lépés annak a szereplőnek az azonosítása, akinek a nézőpontját a szubjektív mondat érvényesíti. A szereplő azonosításához az adott szubjektív mondatot és a narratív szövegben korábban előforduló szubjektív mondatokhoz való kapcsolatot is elemzi az algoritmus, mivel minden szubjektív mondat elkezdi, folytatja, vagy befejezi valamely szereplő nézőpontjának érvényesítését.

A II. MSZNY konferencián mutattam be azt a MorphoLogic Kft-vel közösen kifejlesztett modult [7, 8], amely a nézőpont téri-idői összetevője alapján képes azonosítani az élettörténeti narratívumban érvényesülő perspektívát magyar nyelvű szövegben.

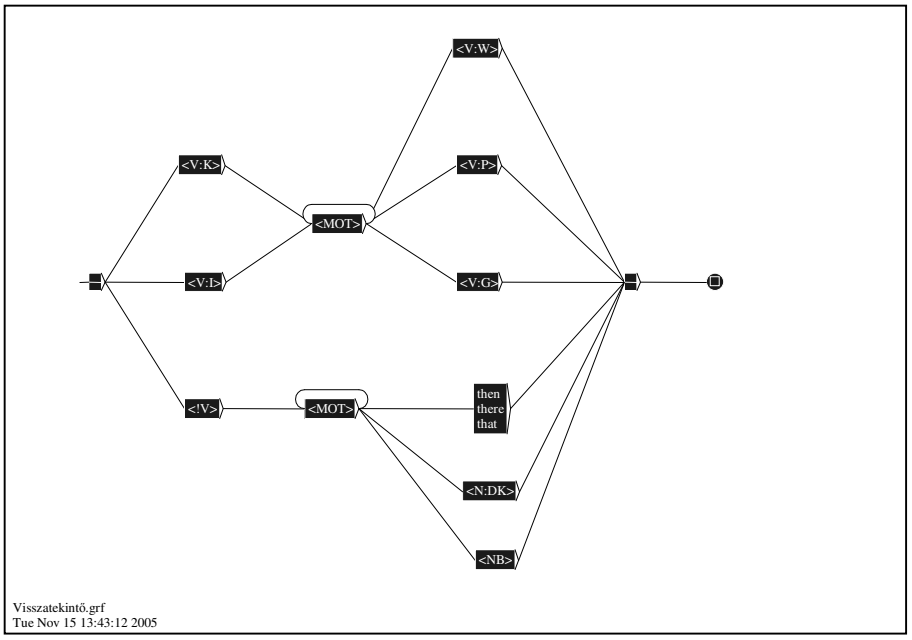
3.2 Az angol nyelvű modul fejlesztése

Narratív perspektíva nyelvtől független jelenség, ugyanakkor eltérés lehet azok között a nyelvi jegyek között, amelyek egyes narratív perspektíva formák érvényesítésében részt vesznek. Az angol változat kidolgozásának kiindulópontja így a magyar változatban szereplő nyelvi jegyek angol megfelelőinek meghatározása volt, amelyet angol nyelvű élettörténeti narratív szöveg elemzése során pontosítottam. Az elemzett narratívumot a 'Szeptember 11. Digitális Archívumból választottam' [5]. Az elemzéshez az INTEX számítógépes nyelvészeti fejlesztő rendszert használtam [10, 13]. A perspektíva formák elemzésének egysége a narratív tagmondat volt. Mivel az INTEX rendszerben a mondat az elemzési egység, az élettörténeti narratívum szövegét előzetesen úgy módosítottam, hogy minden tagmondatot önálló mondattá alakítottam át. Az elemzés során pontosított nyelvi jegyek listáját a 3. Táblázat foglalja össze. Az igeidő kategóriájában az igék harmadik alakja (past participle) és a preterit igealakok visszatekintő, a határozói igenevek (gerundive), jelen idejű illetve infinit igealakok pedig átélő perspektíva formát jeleznek. Mivel a metanarratív formát érvényesítő tagmondatok gyakran nem tartalmaznak igét, az igeidőt ennél a formánál nem vettem figyelembe. Az idő és hely deiktikus kifejezések illetve mutató névmások listáját korlátoztam, mindegyik kategória csak egy-egy elemet tartalmaz. A specifikus kifejezések kategóriáit illetően az indulatszavakat és évszámokat az INTEX rendszerben rendelkezésre álló kódok alapján, a dátum kifejezéseket és a módosítószavak kategóriáját újonnan létrehozott kódok alapján azonosítottam. A mondat módját a mondatvégi írásjelek alapján kódoltam.

3. Táblázat: Az élettörténeti narratív perspektíva formák nyelvi jegyei angol nyelvű szövegben
(Zárójelben az INTEX-ben használt kódok jelei
* Újonnan létrehozott kódok)

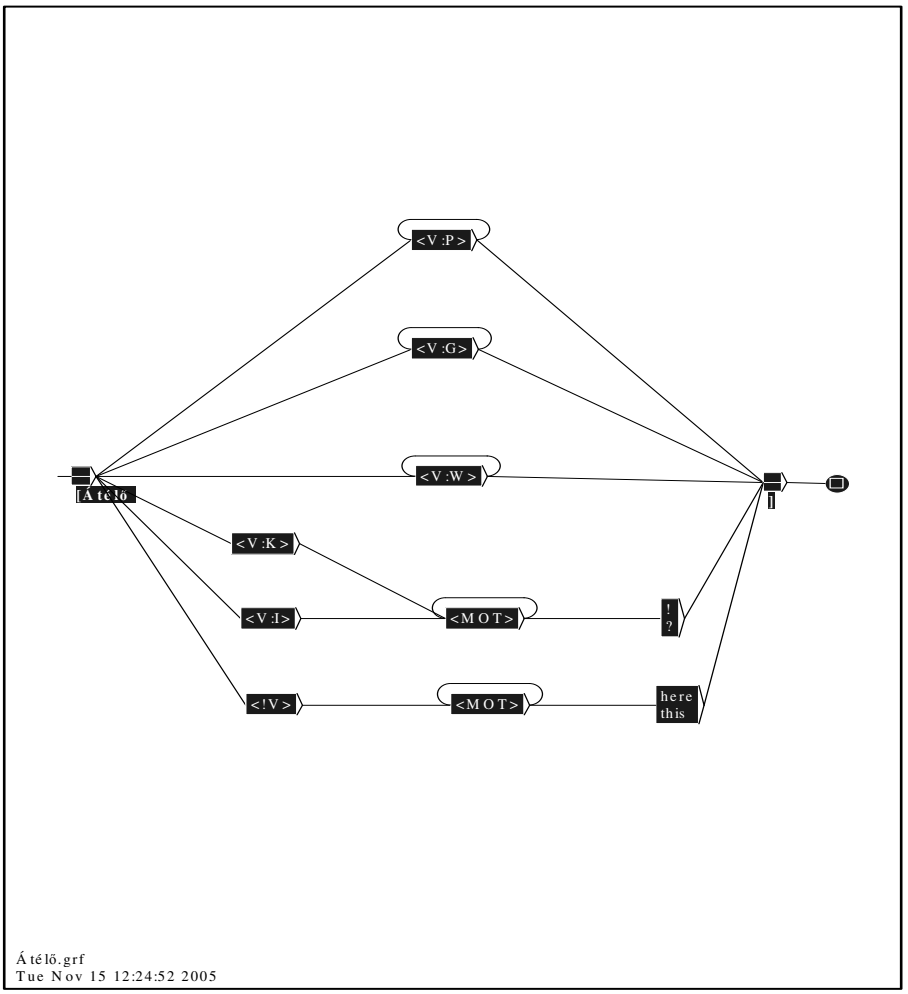
Nyelvi jegyek	Visszatekintő Forma	Átélő Forma	Metanarratív forma
1. Idő deixis			
Igeidő	Past participle (K) Preterit (I)	Present tense (P) Gerundive (G) Infinitive (W)	--
Deiktikus kifejezések	*Then*	*Now*	*Now*
2. Hely deixis			
Deiktikus kifejezések	*There*	*Here*	*Here*
Mutató névmások	*That*	*This*	*This*
3. Specifikus kifejezések	Dátum kifejezések (DK)* Pl. *September* Évszámok (NB) Pl. 2001	Indulatszavak (INTJ) Pl. *Damn*	Módosítószók (MOD)* Pl. *Perhaps* Módosító határozószók (MOD)* Pl. *Apparently* Mentális igék *Think, remember*
4. Mondat módja	Kijelentő	Nincs megkötés	Nincs megkötés

Az élettörténeti narratív perspektíva három formáját egy-egy gráf azonosítja. A visszatekintő formát azonosító gráf (lásd 1. Ábra) az igét tartalmazó tagmondatok igeideje alapján kiválasztja a múlt idejű igealakot tartalmazó tagmondatokat függetlenül attól, hogy jelen idejű igealakot is tartalmaz-e a tagmondat. Az igét nem tartalmazó tagmondatokban a távoli idő és hely deiktikus kifejezések, illetve mutató névmások, dátum kifejezések és szám karakterekből álló sorozatok alapján azonosítja a visszatekintő narratív perspektíva formát.



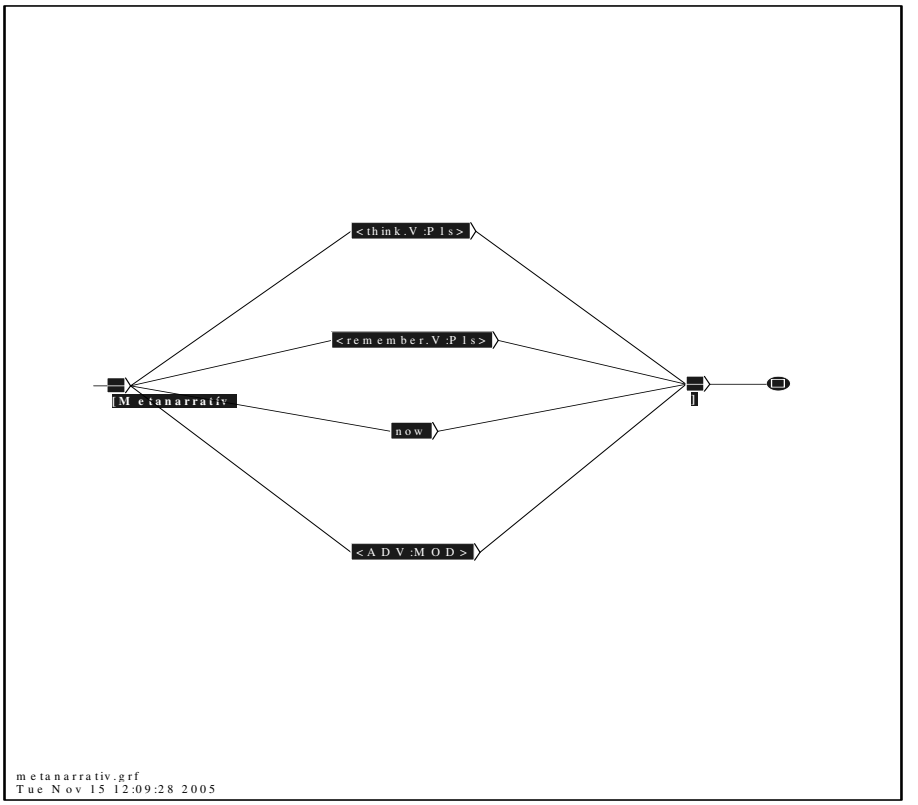
1. Ábra A visszatekintő perspektíva formát azonosító gráf

Az átélő formát azonosító gráf (lásd 2. Ábra) a jelen idejű, infinit vagy határozói igeneveket tartalmazó tagmondatokat választja ki. A gráf átélőként azonosítja azokat a tagmondatokat is, amelyekben múlt idejű igealak és mondatvégi kérdő- vagy felkiáltó jel együtt fordul elő. Az igét nem tartalmazó tagmondatokban a közeli idő és hely deiktikus kifejezések illetve mutató névmások alapján jelez találatot a gráf.



2. Ábra Az átélő perspektíva formát azonosító gráf

Végül a metanarratív formát azonosító gráf (lásd 3. Ábra) a *think* és *remember* igék jelen idejű, egyes szám első személyű alakjának előfordulásakor, a *now* időhatározószó illetve a módosítószavak kategóriájába sorolt szavak előfordulása alapján azonosítja a metanarratív formát.



3. Ábra A metanarratív perspektíva formát azonosító gráf

Az élettörténeti narratív perspektíva formák automatikus azonosításának megbízhatóságát a 4. Táblázat eredményei mutatják. A visszatekintő formát azonosító gráf a legmegbízhatóbb. Az átélő forma esetében a hatékonysági, a metanarratív forma esetében pedig a pontossági mutató értéke marad el a pszichológiai elemzésekben megkövetelt 80 %-os kritériumtól.

4. Táblázat: Az élettörténeti narratív perspektíva azonosításának megbízhatósága (%)

	Visszatekintő forma	Átélő Forma	Metanarratív forma
Hatékonyság	91.7	66.7	81.3
Pontosság	94.8	88.9	62.1

Bibliográfia

1. Banfield, A.: Unspeakable sentences: Narration and representation in the language of fiction. Routledge & Kegan Paul, Boston (1982)
2. Ehrlich, S.: Point of view. A linguistic analysis of literary style. Routledge, London (1990)
3. Kiefer F.: Modalitás. MTA Nyelvtudományi Intézete, Budapest. (1990)
5. Kuyon, M. G.: Story #11463, The September 11 Digital Archive, 13 June 2005, <http://911digitalarchive.org/stories/details/11463>
6. Pólya T.: Az élettörténet narratív perspektívája és az elbeszélő személy identitás állapotának minősége. Ph.D. disszertáció. PTE, Pécs (2003)
7. Pólya T.: Élettörténeti narratív perspektíva és érzelemszabályozás. II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged. (2004) 278-281
8. Pólya T., Kiss B., Naszódi M., & László J.: Az érzelmi tapasztalat minősége az élettörténeti elbeszélésben. A LAS-verticum perspektíva modulja. Pszichológia. (2005) 25(2), 143-155
9. Prince, G.: Dictionary of narratology. University of Nebraska Press, Lincoln. (1987)
10. Silberztein, M.: INTEX. Université de Franche-Comté, Paris. (1997-2004)
11. Uszpenszkij, B.: A kompozíció poetikája. Európa, Budapest. (1984)
12. Zubin, D.A., & Hewitt, E.L.: The deictic center: A theory of deixis in narrative. In J.F. Duchan, G.A. Bruder, & L.E. Hewitt (eds), Deixis in narrative. A cognitive science perspective. Lawrence Erlbaum, Hillsdale, N.J. (1995) 129-155
13. Váradi, T. & Gábor K.: A magyar INTEX fejlesztésről. II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged. (2004) 3-10
14. Wiebe, J.M.: Tracking point of view in narrative. Computational Linguistics. (1991) 20(2), 233-287.
15. Wilkins, D.P.: Expanding the traditional category of deictic elements: Interjections as deictics. In J.F. Duchan, G.A. Bruder, & L.E. Hewitt (eds), Deixis in narrative. A cognitive science perspective. Lawrence Erlbaum, Hillsdale, N.J. (1995) 359-386

Oksági viszonyok azonosítása önéletrajzi narratívumokban

Papp Orsolya[1]
Mészáros Ágnes[2]

[1] PTE BTK Pszichológia Doktori Iskola
H-7624 Pécs, Ifjúság útja 6.
papporsi@lycos.com
[2] MTA Nyelvtudományi Intézet, Korpusznyelvészeti Osztály
H-1399 Budapest, Benczúr u. 33. Pf. 701/518
magnes@nytud.hu

Kivonat: Az előadás a narratív pszichológia és a narratív pszichológiai tartalomelemzés szemléleti és módszertani keretein belül a *koherencia* fogalmát és operacionalizálási lehetőségeit járja körül. A pszichológiai kérdésfeltevés oldaláról az elbeszélő világról és önmagáról alkotott mentális reprezentációinak szerveződésével identikusnak feltételezett személyes élettörténetek összefüggésrendszere a számítógépes nyelvi elemzés során a szöveg strukturális tulajdonságaként közelíthető meg. Adott lexémákat és grammatikai mintázatokat azonosítva így olyan tematikus modulokat képezhetünk, melyek együttes statisztikai alkalmazása kirajzolhatja az önéletrajzi narratívumok szerveződésének különbségeit. Az előadás ezen belül egy oksági kapcsolatokat mérő modul kidolgozásának kezdeti lépéseit mutatja be, melynek alapját Trabasso és van den Broek [19] rekurzív tranzíciós hálózat modellje nyújtja.

1 Bevezetés

Az identitást aktuális célok szerint szerkesztett személyes történetek közös fókuszaként, „gravitációs pontjaként" kezelő konstruktivista pszichológiai elméletek [1] egy jelentős része az önéletrajzi interjúk során kapott szövegek elemezhetőségének gyakorlati problémájával nézett szembe [4]. A kutatási eredmények megbízhatóságának alapját képező korpusz-méret emellett egyre inkább automatizált feldolgozási lehetőségek bevonására ösztönzött.

Az elméleti konstruktumok és az operacionalizálásuk során kifejlesztett mérési eljárások, eszközök között levő rés a narratív koherencia és a szövegkohézió fogalmának elkülönítésében is tetten érhető [5]. Az előbbi egy konstrukciós folyamat eredményeképpen létrejött mentális reprezentáció rendezettségére vonatkozik, melyet a narratív pszichológiai metaelmélet a kultúra által közvetített elbeszélésmintákból kiindulva képzel el [9]. A kohézió pedig az élet jelentős eseményeiről szóló verbális beszámolók mint szövegek összefüggésrendszerét, kapcsoltságát megteremtő nyelvi

jegyek jelenlétére utal. Az elemzésben használt számítógépes eszközök ez utóbbi azonosításában nyújthatnak segítséget[75].

Az élettörténeti interjúkból származó természetes, magyar nyelvű szövegekben a szövegkohézió automatizált, számítógépes elemzéséhez megfelelő kiindulópontot nyújt a Memphisi Egyetem Pszichológiai Kutatócsoportja által angol nyelvre kifejlesztett COH-METRIX szövegelemző program felépítése és általános működési elvei [részletesen lásd 5]. Ebben a kohéziót a szöveg strukturális tulajdonságaként kezelik, melyet hat különböző, a klasszikus narratívum alapvető dimenzióival (idő, tér, ágencia, perspektíva, kauzalitás, intencionalitás) párhuzamba állítható modullal mérnek: kauzális, intencionális, temporális, referenciális, téri és strukturális. Ez a szemléleti keret arra nyújt lehetőséget, hogy az MTA Pszichológiai Intézet Narratológiai Kutatócsoportja által magyar nyelvre kifejlesztett számítógépes modulok [2], [6], [12], [14], [15] eredményeinek együttes statisztikai kezelése jelentse a kohézió–elemzés alapját.

Az előadás további részében egy eddig hiányzó, az oksági kapcsoltságot mérő modul fejlesztésének első lépéseit szeretnénk felvázolni.

A kauzalitás fogalma és mérési lehetőségei

A kauzalitás fogalma az ’50-es évek végétől kezdve a humán viselkedés megértésének magyarázó elemeként [7], a világreprezentáció szervező elveként [16], illetve a szövegértés során végbemenő következtetési folyamatok egyik típusaként [például 10] is kitüntetett figyelmet kapott a pszichológiai kutatásokban. Az oktulajdonítással foglalkozó attribúcióelmélet a ’70-es években kapcsolódott össze a diszkurzuselemzés azon törekvésével, mely a kommunikáció folyamatában próbálta azonosítani a kauzalitásra utaló nyelvi elemeket [ennek áttekintését lásd 3]. A világról alkotott fogalomrendszerünk vizsgálata során az okság formalizált megragadásában pedig Schank és Abelson [16] munkája jelentett mérföldkövet. Az utóbbi oksági tipológiát felhasználva dolgozta ki Trabasso és van den Broek [19] a narratívumokat véges esemény–kategóriák (szetting, cél, elérési kísérlet, reakció, kimenetel) és oksági viszonyok láncolataként leíró modelljét, melyre elemzésünk során támaszkodunk.

Az alábbi táblázat tartalmazza a Trabasso és van den Broek által átvett oksági típusokat, az azonosításukhoz szükséges nyelvi kritériumokat, illetve az ezeknek megfelelő szavak és mintázatok lehetséges szövegbeli megjelenésének példáit. Az eredeti modellhez képest kutatásunkban változtatást jelent, hogy van den Broek és Trabasso [18] későbbi eredményei alapján nem soroljuk az oksági viszonyok közé a Lehetővé tevés kategóriáját, melyben egy adott esemény vagy fennálló körülmény szükséges, de nem elégséges feltétele egy másik történésnek.

[75] Ez a kísérlet abba a tágabb célkitűzésbe illeszkedik, hogy az interjúszövegekben megjelenő explicit nyelvi jegyekhez és az ezek együttjárásából kirajzolódó mintázatokhoz elsősorban – validált tesztkorpuszokkal mért – vonás jellegű személyiségjegyeket kapcsoljanak.

1. Táblázat: Oksági kapcsolatot jelölő nyelvi elemek önéletrajzi narratívumokban

Okság típusa[76]	Nyelvi kritérium[77]	Szövegbeli példa[78]
Motivációs (Okozás)	Célt kifejező nyelvi elemek az első tagmondatban	"Talán így, jó, utólag kibékültünk az anyukámmal, **megpróbálunk** normális kapcsolatot kialakítani, talán ilyen jóban még nem is voltunk…"
Pszichológiai (Kiváltás)	Belső állapotra vagy reakcióra utaló szavak (emóció, kognitív folyamat) a második tagmondatban	"A legnagyobb **veszteség** számomra, de valahol nem is veszteség, de ez a legnegatívabb, hogy elköltöztem otthonról, és hogy ez így állandóan **vívódok**, hogy veszteség vagy nem. Tehát valami **kihalt bennem."**
Fizikai (Eredmény)	Egy esemény szükséges és elégséges feltétele egy másik esemény létrejöttének (két cselekvést vagy egy cselekvést és egy állapotváltozást kifejező ige egymáshoz közeli jelenléte)	„Hát engem először nem **vettek fel** az orvosira, se másodjára. Így **elmentem dolgozni** nővérként…"

Mindenképp kiemelendő a modell hierarchikus felépítése és működése, mely jelen esetben azt jelenti, hogy amennyiben egy adott mondatpár első tagjában célra vonatkozó információt tartalmazó nyelvi elem található, akkor első típusú oksági kapcsolatról beszélhetünk. Ha ez nincs jelen a vizsgált szövegrészben, azonban a második tagmondatban található valamilyen belső állapotra vagy reakcióra utaló kifejezés, akkor a Pszichológiai okság típusa áll fenn. Ennek hiánya esetén érdemes azt meghatározni, hogy a mondatpár első felében szereplő esemény(ek) szükséges és elégséges feltételeket nyújt-e egy utána következő változás létrejöttéhez. Ebben az utolsó esetben a legnyilvánvalóbb az a körülmény, hogy a modell események közötti kapcsolatokat vizsgál. Élőnyelvi szövegek esetében azonban akadályt jelent az a megkötés, mely a célszavak jelenlétét a tagmondatpárok első vagy második elemére korlátozza.

„Hát például, amikor bekerültem középiskolába, akkor így nagyon **nehéz** volt az elején, **mivel** egy alacsonyabb színvonalú általános iskolába jártam,…"

[76] A Trabasso és van den Broek által felállított típusok mellett zárójelben megadjuk az ezeknek megfelelő, ismertebb Schank és Abelson-féle kategóriák elnevezéseit [magyarul lásd 8, 13].

[77] Az egyes kritériumok jelenlétét a szövegben Trabasso és van den Broek-től eltérően, a vizsgált szöveg jellegzetességeinek megfelelően, nem egymás utáni tagmondatok között vizsgáljuk, hanem az egyes szavak közeli környezetében.

[78] A kritériumoknak megfelelő szavakat dőlt betűvel emeltük ki az adott szövegkörnyezetből.

A példában az érzelmi reakcióra utaló kifejezés (*nehéz volt*) a tagmondatpár első elemében található, melyet egy magyarázat követ. Így a Trabasso és mtsai által felállított munkahipotézistől - melyet szoros történetvezetésű, tanulságra kiélezett mesék manuális elemzésével támasztottak alá - , az általunk használt természetes nyelvi szövegek automatikus feldolgozásánál a célszavak tagmondatok közötti sorrendjére vonatkozóan el kellett térnünk.

A modell nem nyújt megfelelően kidolgozott támpontokat abban a tekintetben sem, hogy az egyes típusú oksági kapcsolatok vizsgálatában pontosan milyen nyelvi elemek jelenléte meghatározó; a Motiváció esetében példaként hozza fel, hogy célinformációt bizonyos igék (például **akar**) mellett egyes főnévi igeneves szerkezetek (elment horgász**ni**), illetve egyedi vonzatkeretek (elfáradt a munká**ban**) is hordozhatnak.

Az oksági modell alkalmazása

A fent említett modellekre támaszkodva kutatásunk elsődleges célja, hogy természetes nyelvi szövegekben tudjunk különböző típusú oksági kapcsolatokat számítógépes nyelvészeti eszközök segítségével azonosítani. Ennek megvalósításához az MTA Pszichológiai Kutatóintézet 2004 tavaszán felvett élettörténeti szövegkorpuszát használjuk, melyben a vizsgálati személyek hat (első emlék, teljesítményről, veszteségről, félelemről, jó és rossz kapcsolati élményről szóló) személyes narratívumot meséltek el. Az általunk összeállított próbakorpusz 14 személy történeteit tartalmazza.

Az elemzést az Intex számítógépes eszközzel végezzük [17], melyet természetes nyelvek fomalizált leírására fejlesztettek ki. Az eszköz nagy méretű korpuszok valós idejű feldolgozását teszi elérhetővé: a gyors elemzés lehetőségét az nyújtja, hogy mind a bemeneti szöveget, mind a nyelvtanokat tömörített formában tárolja.

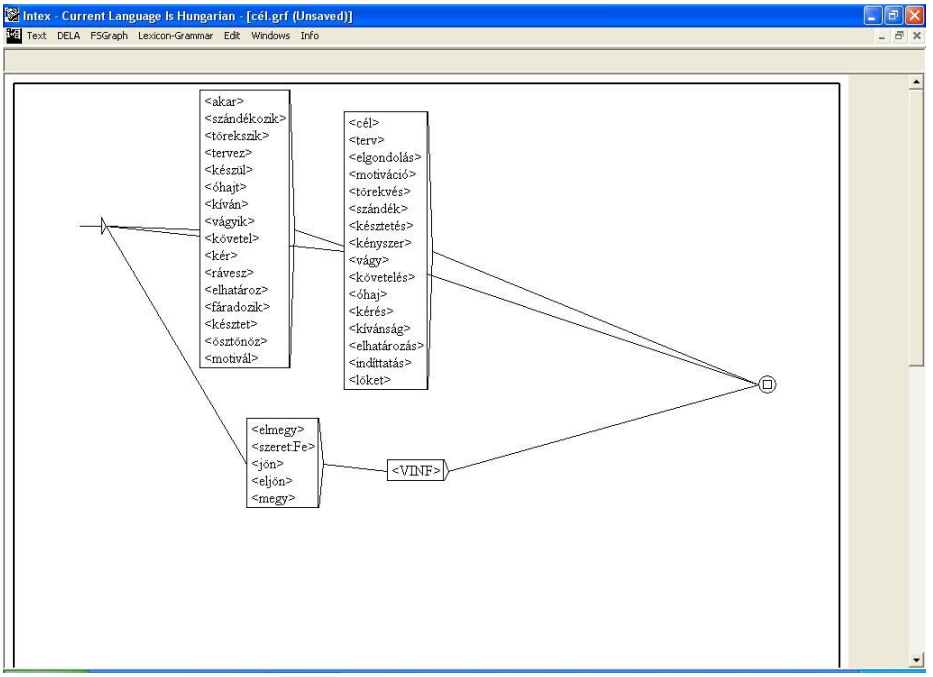
A programot lexikalista nyelvelemzésre fejlesztették ki, melynek alapjai az elektromos szótárak, és véges állapotú gráfokban ábrázolt nyelvtani szabályok. A szótárakban a szavak morfoszintaktikai és szemantikai jegyei egy szinten vannak kódolva, így a gráfokban mindezen információra egyidejűleg lehet hivatkozni. Az eszközt adottságai különösen alkalmassá teszik tartalomelemzési feladatokra, hiszen lehetőség van előre összeállított lexikonok elemeinek – akár szemantikai jegyek alapján történő – együttjárási vizsgálatára.

Az elemzés menete

Az okságra vonatkozó elméleti modell hierarchikus felépítését követve első lépésben a cél-információt tartalmazó szavak automatikus felismerését végző gráfot szerkesztettünk, mely az 1. ábrán látható: itt gyűjtöttük össze (a Magyar Szinonimaszótár [11] segítségével) azokat az igéket és főneveket, amelyek szemantikájuk alapján az ágens intencionális hozzáállását feltételezik. A gráf másik ágán azokat a szövegünkben

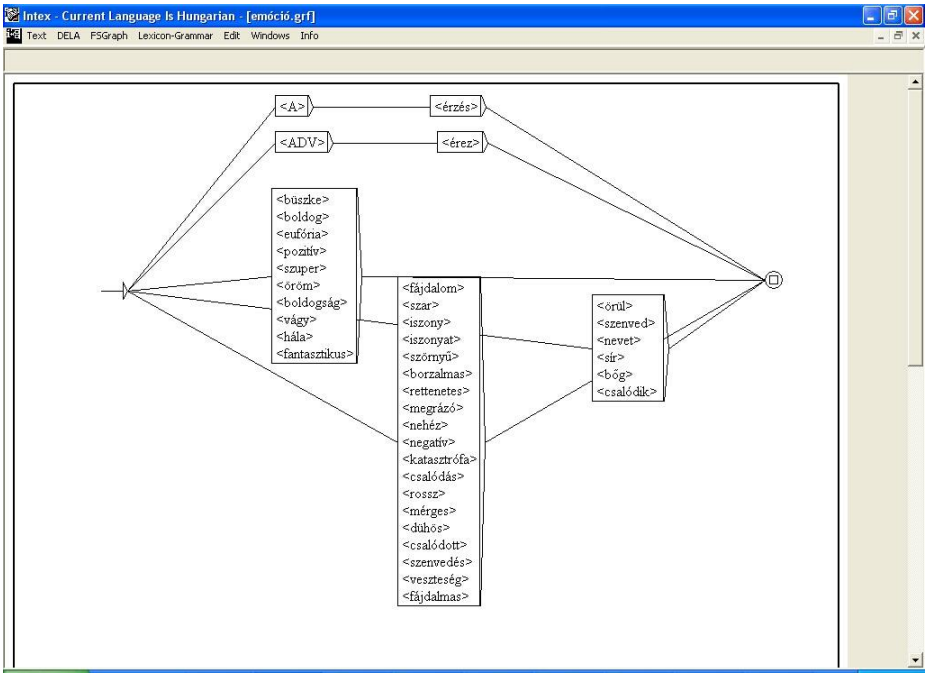
megtalálható igéket soroltuk fel, amelyek csupán főnévi igeneves vonzatukkal együtt utalnak egy cél jelenlétére. Ez utóbbi csoporton belül a *szeret* ige viselkedése annyiban tért el, hogy csupán a feltételes módú alakjai hordoznak intencionális információt (szeretek síelni vs. szeretnék síelni). Az alábbi példa ezen gráf találatainak természetes nyelvi környezetben való előfordulását mutatja:

„…és emlékszem rá, az elsõ történelem dolgozatunkra, dolgozatunkat írtuk, és én arra **készültem** elõtte egész hétvégén, és így, már föladtam, már untam, és mondtam édesanyámnak, hogy én nem megyek el iskolába, mert én nem tudom, meg nagyon meg nagyon nehéz, meg sok, meg satöbbi, és így másnap **elmentem megírni** a dolgozatot,…”



1. ábra Cél-információt tartalmazó igei szerkezetek és főnevek.

A második típusú oksági viszony azonosítását az emocionális állapotokat leíró szavak felől közelítettük., melyek lehetnek önmagukban álló szemantikailag specifikus szavak, illetve az érzés/érez szó előtt álló bármilyen melléknév vagy határozószó. Ez a gráf látható a 2. ábrán, melyet az élettörténeti szövegekből vett példák követnek.



2. ábra Érzelmi állapotokat azonosító gráf

„…és hát ilyen **katasztrófa** volt, a gyerekek mittudomén, basszusgitároztak az órán, meg ki-be mászkáltak, tehát semmi, semmilyen szinten nem voltak hajlandóak együtt mûködni,...”

„...és ha ezt tudom, ezen túl vagyok, utána megiszom két pohár Heinekkent, és én **jól érzem** magamat attól az embertömegtõl, ami ott hömpölyög,...”

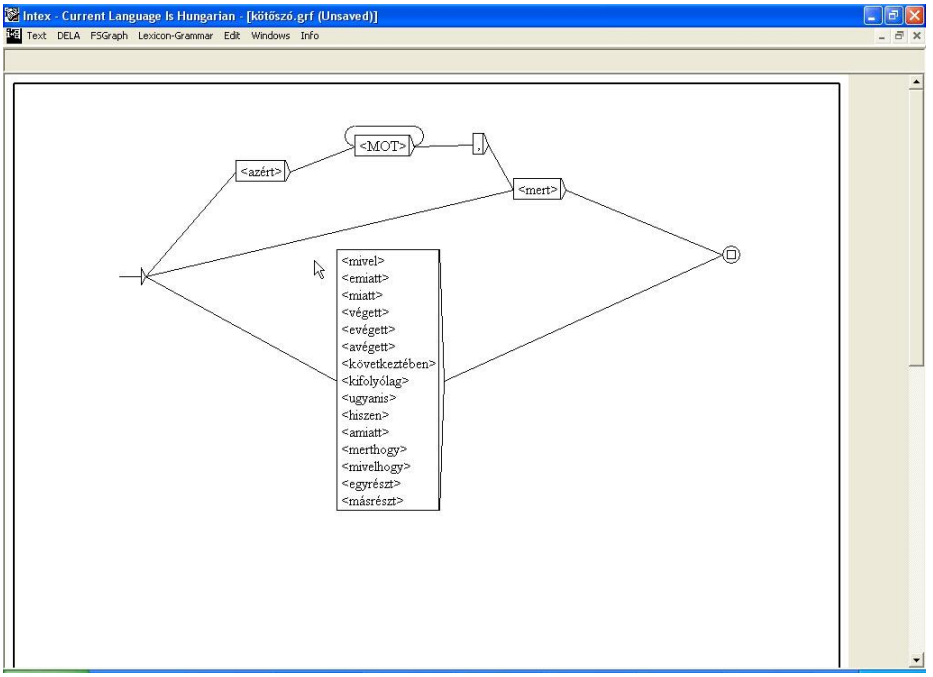
Az oksági kapcsolatok harmadik csoportja olyan események vagy cselekedetek közötti összefüggéseket tartalmaz, melyben az első, a kiváltó történés vezet el egy eredményként értékelhető változáshoz. Kiindulásul olyan változást jelző szavakat gyűjtöttünk össze, amelyek vélhetően egy új cselekedetet vezetnek be:

„…tényleg Amszterdamban is elõfordult, hogy a lánynak beszólt egy fiú az utcán, egy tök ismeretlen férfi, és a lány **elkezdett** futni utána,....”

Ennek az automatikus azonosításnak a nehézsége, hogy ugyanezek a szavak vezetik be a belső állapotváltozásokat is; bizonyos esetekben pedig egyenesen eldönthetetlen hogy egy adott reakciót cselekvésként vagy érzelmi válaszként értelmezzünk:

„...az akkori barátnõmmel, vagy hát ismerõs kislánnyal, azt hiszem Zsuzsinak hívták, kitaláltuk, hogy felmászunk egy mászókára, egy ilyen ház alakú mászóka volt, és **sikerült** a legtetejére felmásznunk, és akkor így lovagló ülésben ültünk a legtetején, és akkor kitaláltuk azt, hogy most csináljunk úgy, mintha sírnánk, és ebbe a helyzetbe annyira beleéltük magunkat, hogy **elkezdtünk** a végén tényleg **sírni**, azért, mert fölmentünk és mi nem tudunk onnan lejönni. Persze valószínûleg le tudtunk volna jönni, de azért odajött hozzánk a bölcsõdei dolgozó, és leszedett bennünket. ...”

Az elsősorban elméleti kiindulópontként alkalmazott Schank és Abelson illetve Trabasso és mtsai nevéhez fűződő modellek árnyalt képet adnak a lehetséges oksági kapcsolatok típusairól, jelen célkitűzésünk szempontjából azonban egy általánosabb kategória felvétele is hasznosnak látszik, mely pusztán a kauzális viszony jelenlétére utaló kötőszavakat, névutókat illetve utalószavakat tartalmazza.



3. ábra Az oksági viszony jelenlétére általánosságban utaló szavak

A kutatás jelen stádiumában ez a robusztus módszer sok olyan oksági találatot tesz lehetővé, amelyet a környezetében előforduló tartalmas szavak mentén nem tudnánk azonosítani. Az alábbi bekezdés ezt az esetet illusztrálja:

„...igazából csak ezért maradok még vele, **merthogy** kollégiumi szobát fizetünk itt ketten, három helyeset, és nem is lenne nagyon értelme igazából három-négy hónapra kilépni ebbõl..."

Értékelés

A tanulmány lezárásaként egy összefüggő emléken mutatjuk be a szövegben jelenlevő oksági kapcsolatokat, illetve ezeknek az általunk szerkesztett gráfok alapján történő azonosítását.

Gólyatábort szerveztünk. Körülbelül húszan voltunk, és vezető is volt, és ebből én voltam az egyik, marketinges részleget vezettem, és ez azt jelentette, hogy a nyár folyamán reggel tíztől este nyolcig itt voltunk szinte minden nap, és például hatalmas lepedőkre festettünk

képeket dekorációként, és nagyon **jól éreztük** magunkat egész nyár folyamán. És jó volt az egyetemen, és nekem ez nagyon nagy pluszt adott, hát én **úgy érzem**, hogy sokat tanultam ebből az egészből, sokkal **bátrabb lettem**, sokkal inkább szókimondóbb, és nagyon élveztem, és a végén az történt, hogy elvesztek a lepedők, és nem jutottak el a gólyatáborba, és hatalmas **csalódás** volt, de így is újra csinálnám, újra, és újra.

2. Táblázat: Egy összefüggő emlék oksági elemzése

Az oksági viszony	Automatikus találat
A gólyatábori élmények miatt jól érezte magát egész nyáron	**Jól éreztük** magunkat
A nyári kalandok eredménye: ➢ tapasztalatszerzés ➢ bátorság ➢ szókimondás	**Úgy érzem**, hogy sokat tanultam ebből az egészből
	Bátrabb lettem
	-nincs találat
Mivel elvesztek a lepedők, nem jutottak el a gólyatáborba, ez csalódást okozott	**Csalódás**

Bibliográfia

1. Denett, D.: Consciousness explained. Boston, Little Brown (1991)
2. Ehmann, B.: A LAS VERTICUM időmodulja. In: Alexin, Z., Csendes, D. (szerk.): MSZNY 2004 Konferenciakötet, Juhász Nyomda, Szeged (2004) 257-261
3. Fiske, S.T., Taylor, S.E.: Attribution Theory: Theoretical Refinements and Empirical Observations = Social Cognition. 2nd edn. McGraw-Hill, Inc., New York Toronto () 57-95
4. Gergen, K.J., Gergen, M.M.: A narratívumok és az én mint viszonyrendszer. Narratívák, 5. Kijárat Kiadó, Budapest (1988/2001) 77-121
5. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments and Computers (2004)
6. Hargitai, R.: A LAS VERTICUM tagadás és self-referencia modulja. In: Alexin, Z., Csendes, D. (szerk.): MSZNY 2004 Konferenciakötet, Juhász Nyomda, Szeged (2004) 261-265
7. Heider, F. (szerk.): The Psychology of Interpersonal Relations. John Wiley and Sons, Inc., New York (1958)
8. László, J.: A kognitív egyensúly elmélettől a személyközi forgatókönyvekig = Szerep, forgatókönyv, narratívum: Szociálpszichológiai tanulmányok. Scientia Humana, Budapest (1998) 83-97.
9. László, J.: Narratív pszichológia: új megközelítés a pszichológiában. Narratívák, 5. Kijárat Kiadó, Budapest (1988/2001) 7-15
10. Linderholm, T., Everson, M.G., van den Broek, P., Mischinski, M., Crittenden, A., Samules, J.: Effects of Causal Text Revisions on More- and Less-Skilled Readers' Comprehension of Easy and Difficult Texts. Cognition and instruction, Vol. 18 (4). Lawrence Erlbaum Associates, Inc. (2000) 525-556
11. O. Nagy, G., Ruzsiczky, É. (szerk.): Magyar Szinonimaszótár. Akadémiai Kiadó, Budapest (1999)
12. Péley, B.: A LAS VERTICUM 'Szereplő-funkció' modulja. In: Alexin, Z., Csendes, D. (szerk.): MSZNY 2004 Konferenciakötet, Juhász Nyomda, Szeged (2004) 265-269
13. Pléh, Cs.: A történetszerkezet és az emlékezeti sémák. Akadémiai, Budapest (1986)
14. Pohárnok, M.: Kapcsolati mozgások számítógépes nyelvészeti vizsgálata élettörténeti narratívumokban. In: Alexin, Z., Csendes, D. (szerk.): MSZNY 2004 Konferenciakötet, Juhász Nyomda, Szeged (2004) 274-278
15. Pólya, T.: Élettörténeti narratív perspektíva és érzelemszabályozás. In: Alexin, Z., Csendes, D. (szerk.): MSZNY 2004 Konferenciakötet, Juhász Nyomda, Szeged (2004) 278-285
16. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and understanding: an Inquiry into human Knowledge Structures. Erlbaum, Hillsdale (1977)
17. Silberztein, M.:Intex Manual. Internetes elérhetőség: www.nyu.edu/pages/linguistics/**intex/**
18. Trabasso, T., van den Broek, P.: Causal Networks versus Goal Hierarchies in Summerizing Text. Discourse Processes, Vol. 9. Alex Publishing Corporation, New Jersey (1986) 1-15
19. Trabasso, T., van den Broek, P., Suh, S.Y.: Logical Necessity and Transitivity of Causal Relations in Stories. Discourse Processes, Vol. 12. Alex Publishing Corporation, New Jersey (1989) 1-25

A külső-belső kontroll nyelvi markerei

Füleki Bettina[1], László János[2]

[1]PTE BTK Doktori képzés
[2] MTA Pszichológiai Kutatóintézet, PTE BTK Pszichológiai Intézet

Kivonat: A cikk egy folyamatban lévő kutatást mutat be. A kutatás célja a külső belső-kontroll nyelvi markereinek meghatározása elbeszélt szövegekben. A nyelvi markerek meghatározásával olyan tartalomelemző programot fejlesztünk ki, amely bármely elbeszélt szövegen lefuttatható, és feltárható vele a szöveg létrehozójának külső vagy belső kontrollos személyiség dimenziója.

1. Bevezetés

Elbeszélt szövegeknek a rendszerváltásról készült 50 db interjút használtunk fel. Első lépésben tartalomelemzéssel meghatároztuk a rendszerváltás és a külső belső kontroll összefüggéseit, majd az így kapott eredményeket tovább lebontva meghatároztuk a külső, illetve belső kontroll nyelvi markereit.

1. 1 Külső-belső kontroll és a rendszerváltás megítélésének összefüggése

A belső kontrollos személyek el tudják választani a saját (vagy a szüleik) esetleges negatív tapasztalataitól a rendszerváltás egészének értékelését.

A rendszerváltás lényegének a nagyobb szabadságot nevezték meg, egyéni és közösségi szinten. A belső kontrollos alanyok többen beszámoltak arról, hogy kialakult életstratégiájuk van. Kiemelték, hogy a rendszerváltás kevésbé hatott az életükre. Igyekeztek a saját értékeik szerint élni, függetlenül a fennálló rendszertől. Itt többféle érték is megjelent: vallásos nevelés, szakmai érdeklődés, család. A rendszerváltás előtt is megkeresték a lehetőségeket a saját életükben.

A külső kontrollos személyek elsősorban gazdasági alapon ítélték meg a rendszerváltást, aminek lényegének tartották, hogy „minden drágább lett”.

Egyértelműen negatívnak ítélték meg, olyan eseménynek, ami magával sodorta őket, önmagukat tehetetlennek és vesztesnek élik meg. A politikai változást nem vagy alig említették, a szabadságot a zűrzavarral azonosították2. A külső-belső kontroll nyelvi markerei

1. 2. A modul kialakítása

A szövegben azonosítható külső-belső kontroll kategóriákból képezzük a külső-belső kontroll modult, ami az atlas.ti tartalomelemző programmal lefuttatható.

Első lépésben a vizsgált elbeszélt szövegekben meghatároztuk tartalomelemzéssel, a kontroll helyét, és olyan szempontok szerint, mint „magának vagy másnak tulajdonítja" „külső eseménynek tulajdonítja. Ezután azonosítottuk a külső-belső kontrollra utaló nyelvi markereket, majd kategorizáltuk.

2. 1. 1. A belső kontroll nyelvi markerei

Olyan kifejezéseket kerestünk, amik a belső kontrolos személyiségdimenzióra utalnak, majd ezeket kategóriákba rendeztük.

Ezek a kategórianevek a belső kontrolos kifejezések között értendők, tehát lehetnek olyan kifejezések, amelyek a kategórianév alapján ide tartozhatnának, de a kontrollhellyel nem állnak kapcsolatban, ezekkel nem foglalkoztunk.

2. 1. 2. Énre vonatkozó markerek

1. Táblázat: énre vonatkozó markerek

maga* módján
magam*
magunk*
Önálló

Az egy kategóriába tartozó nyelvi markereket leszámoltuk a fiatal-idős és alacsonyan képzett-magasan képzett szempontok szerint osztályozott interjúkon.

2. Táblázat: énre vonatkozó markerek száma csoportok szerint.

Idősek	26
Fiatalok	5
Alacsonyan képzettek	0
Magasan képzettek	14

Az énre vonatkozó kifejezések, ahol az alanyok a saját maguk szerepét, és a környezettől való függetlenségüket fejezik ki pl.: „magam módján" az idősebbeknél gyakrabban jelennek meg, ami egybevág azzal a kézzel kódolt eredménnyel, hogy az úgynevezett „kialakult életstratégiák" csak az idősebb korosztályban jelentek meg. Szintén nő a markerek száma az iskolázottsággal, ami egybevág azzal, hogy az iskolázottabb alanyok inkább a belső kontrolos csoportba tartoznak.

2. 1. 3 Fejlődés markerei

3. Táblázat: fejlődés markerek

fejlőd*
halad*
felé vezet

Markerek száma csoportok szerint.

4. Táblázat: fejlődés markerek száma csoportok szerint.

Idősek	13
Fiatalok	17
Alacsonyan képzettek	0
Magasan képzettek	30

A fejlődésre vonatkozó kifejezések, ahol az alanyok a változás irányát pozitívnak értékelik, a rendszerválásra vonatkozóan szintén egy jobb irányba való elmozdulást érzékelnek a magasan képzettek között a legmagasabb. Ők a változásokból képesek a lehetőséget látni, és megragadni a változásokat egy folyamat részeként szemlélni. A fiataloknál szintén magasabb, mint az idősebbeknél, ez a fejlődési sajátosságnak értelmezhető, a készségek és ismeretek elsajátításával együtt járó jelenég.

2. 1. 4. Hitre vonatkozó markerek

5. Táblázat: Hitre vonatkozó markerek

hisz*
hit*
(azt) hisz*

Markerek száma csoportok szerint.

6. Táblázat: Hitre vonatkozó markerek száma csoportok szerint.

Idősek	3
Fiatalok	12
Alacsonyan képzettek	0
Magasan képzettek	15

A magasan képzettek és a fiataloknál többször találjuk meg a hit markereit, ami ahol a jövőbe vetett pozitív elvárásokat mutatja, illetve az „azt hiszem” kifejezés a képzettség szintjével, mint magasabb szintű nyelvi megfogalmazás is összefügghet.

2. 1. 5. Szabadságra vonatkozó markerek

7. Táblázat: szabadságra vonatkozó markerek

lehetőség*
lehetség*
autonóm*
szabad*
választ*
dönt*
tehet*

Markerek száma csoportok szerint.

8. Táblázat: szabadságra vonatkozó kifejezések száma csoportok szerint.

Idősek	0
Fiatalok	97
Alacsonyan képzettek	15
Magasan képzettek	57

A belső szabadság markerei kiugróan magasak a fiatalok és a magasan képzettek csoportjában. A szabadság a belső kontrolt jelző markerek között a legnagyobb számban megjelenő. Ez egybevág azzal, hogy a rendszerváltás értékei között az egyik leghangsúlyosabb az interjúkban a szabadság volt, és a belső kontrollos interjúalanyok a rendszerváltást pozitívan értékelték.

2. 1. 6. Belső motivációra vonatkozó markerek

9. Táblázat: belső motiváció markerek

hajt*
akar*
igyeksz*
próbál*

Markerek száma csoportok szerint.

10. Táblázat: belső motivációra vonatkozó markerek száma csoportok szerint.

Idősek	5
Fiatalok	2
Alacsonyan képzettek	3
Magasan képzettek	4

A belső motiváció markerei az idősebb korosztályban a jellemzőbbek, ez a kitartás és a frusztrációs tolerancia életkorral való növekedésére utal, kialakul egyfajta „bölcsebb” életszemlélet. Kevésbé függ össze a végzettséggel.

2. 1. 7. Érzelmi kontrollra vonatkozó markerek

11. Táblázat: érzelmi kontroll markerek

hajt*
akar*
igyeksz*
próbál*

Markerek száma csoportok szerint.

12. Táblázat: érzelmi kontrollra vonatkozó markerek száma csoportok szerint.

Idősek	14
Fiatalok	0
Alacsonyan képzettek	4
Magasan képzettek	10

Az érzelmi kontroll markerei legnagyobb számban az idősebb csoportban fordulnak elő, ami szintén arra utal, hogy az érzelmi kontroll az érettséggel megszerezhető, szintén magasabb a jobban képzettek körében.

2. 1. 8. Alkotás-monitorozásra vonatkozó markerek

13. Táblázat: érzelmi kontroll markerek

csinál*
megvalósít*
teremt*
odafigyel*
figyel*
kitalál*
megtalál*

Markerek száma csoportok szerint.

14. Táblázat: alkotás-monitorozásra vonatkozó markerek száma csoportok szerint.

Idősek	7
Fiatalok	13
Alacsonyan képzettek	5
Magasan képzettek	15

Az alkotás-monitorozás a legmagasabb a magasan képzettek körében, ami elsősorban egy a tanulás és a magasabb intellektuális erőfeszítést igénylő munka alatt kialakult attitűdöt takar. A fiatalok és idősek között kisebb a különbség, de a fiatalok a nagyobb aktivitási szint miatt jobbak.

2. 2. Összefoglalás a belső kontroll nyelvi markereiről kategóriák szerint

Táblázat: a belső kontroll kategóriái csoportokon belül

	Idős	Fiatal	magasan képzett	alacsonyan képzett
Énre v.	11	3	14	0
Hit	3	12	15	0
Szabadság	0	97	57	40
Fejlődés	13	17	30	0
Belső motiváció	5	2	4	3
Érzelmi kontroll	14	0	10	14
Alkotás-Monitorozás	7	13	15	5
Összesen	57	161	166	52

Az idős csoportban a belső kontroll markerei közül az érzelmi kontroll a legtöbb, majd a fejlődés és harmadik az énre vonatkozó kifejezések. Legkevésbé a szabadság jelent meg. A fiatalok között kiugró számúak a szabadság markerei, majd a fejlődés és az alkotás-monitorozás következik. Legkevesebb az érzelmi kontrollra vonatkozó marker. Jegyezzük meg, hogy az időseknél pedig éppen ez a legtöbb!

Az alacsonyan képzett csoportban összességében a legkevesebb belső kontrollra utaló marker jelenik meg. A megjelenteken belül a legtöbb a szabadságra vonatkozó, majd ezt követi az érzelmi kontroll. Ez arra utalhat, hogy aki stabil énképpel rendelkezik, az el tudja fogadni a helyét, és nem a külső körülményeket okolja, pl. a továbbtanulás hiányáért. A legkevesebb marker az énre, a hitre és a fejlődésre vonatkozik. Ez egybevág azzal a tartalomelemzéssel kapott eredménynek, hogy a rendszerváltás vesztesei a tanult tehetlenség és kiábrándultság állapotában vannak.

3. A külső kontroll nyelvi markerei

3. 1. 1. Tagadásra vonatkozó markerek

16. Táblázat: érzelmi kontroll markerek

nem
rossz*

Markerek száma csoportok szerint.

17. Táblázat: tagadásra vonatkozó markerek száma csoportok szerint.

Idősek	690
Fiatalok	1065
Alacsonyan képzettek	1068
Magasan képzettek	687

A tagadással a legtöbbet az alacsonyan képzettek és a fiatalok éltek. A fiataloknál ez a kritikus magatartás fejlődési fázisnak is értelmezhető, a probléma akkor jelentkezik, a felelősség vállalása a belső kontroll az elhatárolódás helyett később sem jelentkezik.

3. 1. 2. Függésre vonatkozó markerek

18. Táblázat: függés markerek

*ja, ’*ják
*nának
*nének
Mondta
mondták
megszab*
hallottam
Nézik

Markerek száma csoportok szerint.

19. Táblázat: függésre vonatkozó markerek száma csoportok szerint.

Idősek	61
Fiatalok	72
Alacsonyan képzettek	133
Magasan képzettek	0

A függésre utaló markerek, a másokra hivatkozás, harmadik személyben való beszéd az alacsonyabb iskolázottságúak között a leggyakoribb, ezt követik a fiatalok, a magasan képzettek használnak a legkevésbé a másokra hivatkozást jelző markereket, ami egybevág azzal, hogy a belső kontrollos jegyek közül ők használják a legtöbb énre vonatkozó markert.

3. 2. Összefoglalás a külső kontroll nyelvi markereiről kategóriák szerint

20. Táblázat: a külső kontroll kategóriái csoportokon belül

	Idős	Fiatal	magasan képzett	alacsonyan képzett
Tagadás	690	1065	687	1068
Függés	61	15	0	133

A külső kontrollos jegyek közül erősen jelzik a kontrollhelyt a tagadás nyelvi markerei, amelyek legnagyobb arányban az alacsonyan képzetteknél fordulnak elő, a függés nyelvi markerei legtöbbször szintén az alacsonyan képzetteknél jelentkeznek, és gyakorlatilag nem láthatók a magasan képzettek csoportjában.

4. Összefoglalás

A közölt adatok nem statisztikai elemzés céljára szolgálnak, hanem a kutatás további elmélyítéséhez és a szólisták bővítéséhez szolgálnak alapul.

Az eddig azonosított nyelvi markerek alapján is megállapíthatjuk, hogy a külső-belső kontroll további alkategóriákra bontható dimenzió, amely alkategóriák mentén a fiatalok és idősek, alacsony és magasan iskolázottak között jelentős eltérések vannak.

Bibliográfia

1 Ehmann Bea, Kis Balázs, Naszódi Mátyás, László János: A szubjektív időélmény tartalomelemzéses vizsgálata. In: Pszichológia, 2005, (25), 2, 133-142

2. László János, Ehmann Bea: Narratív pszichológia és narratív pszichológiai tartalomelemzés. In: Magyar Pszichológiai Szemle, 2005, 3.

VII. Beszédtechnológia, kommunikáció

Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével

Bánhalmi András[1], Kocsor András[1], Paczolay Dénes[1]

[1] MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport, Aradi vértanúk tere 1, H-6720 Szeged, Hungary
{banhalmi, kocsor, pdenes}@inf.u-szeged.hu

Kivonat: Cikkünkben újszerű megoldásokat javasolunk a valós idejű beszédfelismeréshez szükséges nyelvi modellek területén, a felismerési pontosság és sebesség növelése érdekében. Különböző nyelvi modellek (pl. szabály alapú modellek, fonéma N-gram, szó és szócsoport N-gram modellek) párhuzamos futtatásával, illetve aggregálásával egyrészt a szó N-gram simítása, másrészt a hipotézisek számának hatékonyabb csökkentése érhető el. A szócsoport N-gramok kiértékeléséhez a szavak csoportosítását a szavak mondattani szerepét leíró MSD-kódok (Morpho Syntactic Description) [3] felhasználásával végeztük el. Az N-gram alapú statisztikai modellek hagyományos kiértékelés esetén csak az n. szó teljes felismerése után szolgáltatnak valószínűségi értékeket. Olyan eljárásokat is kidolgoztunk, amelyek használatával már az n. szó felismerésének befejezése előtt rendelkezésre állnak közelítő valószínűségi becslések.

1 Bevezetés

A számítógépek megjelenésével a dokumentumok tárolása, nyilvántartása és visszakeresése nagyságrendekkel gyorsabbá vált, azonban a szövegek és adatok bevitele még mindig túlságosan sok humán erőforrást igényel. Igaz ez az orvosi vizsgálati eredmények rögzítésére is, amely folyamat egy speciális diktáló rendszer segítségével lényegesen felgyorsítható, illetve egyszerűsíthető. Kisebb népcsoportok által beszélt, illetve speciális tulajdonságokkal rendelkező nyelvekre – mint például a magyar – egyelőre nagyon kevés diktáló szoftver látott napvilágot. Az MTA-SZTE Mesterséges Intelligencia Kutatócsoport beszédfelismerési kutatásokkal foglalkozó műhelyében kifejlesztettünk egy, a magyar nyelv automatikus felismerésére alkalmas keretrendszert, amelyre egyedi diktáló rendszerek építhetők.

A beszédfelismerő keretrendszer magját két fő modul, az akusztikai és a nyelvi modul alkotja. Az akusztikai modul egy saját implementálású Rejtett Markov Modell segítségével alkalmas a magyar nyelv beszédhangkészletének hatékony felismerésére. A beszédhangmodellek felépítése egy nagyméretű beszédadatbázis [9] alapján történt. Önmagában az akusztikai modell által szolgáltatott hipotézisek, azaz a feltételezett beszédhangsorozatok száma a bemondás hosszával exponenciálisan növekszik. A nyelvi modul feladata, hogy a lehetséges hipotézisek számát kezelhető

számúra korlátozza. A nyelvi modulba beépített tudásbázisok (nyelvtani szabályok, statisztikák) rendszerint a nagyobb hatékonyság érdekében nem a teljes beszélt nyelvet, hanem csak egy-egy szakterület speciális szóanyagát és nyelvtani szabályait modellezik. Cikkünkben egy, pajzsmirigy szcintigráfiás leletekből álló szövegkorpuszt használtunk fel a nyelvi modellek definiálására.

A nyelvi modellek területén – az ismert módszerek mellett [4] – olyan újszerű megoldásokat dolgoztunk ki és valósítottunk meg, amelyek a felismerési pontosság és sebesség növelése révén valós idejű beszédfelismerést tesznek lehetővé. A kifejlesztett nyelvi modul egyik újszerű eleme, hogy különböző nyelvi modelleket (pl. szabály alapú modellek, fonéma N-gram, szó és szócsoport N-gram modellek) képes párhuzamosan alkalmazni, különböző súlyokkal aggregálva azokat. A szócsoport N-gramok kiértékeléséhez a szavak csoportosítását a szavak mondattani szerepét leíró MSD-kódok (Morpho-Syntactic Description) [3] felhasználásával végeztük el.

Az N-gram alapú statisztikai modellek hagyományos kiértékelés esetén csak az n. szó teljes felismerése után szolgáltatnak valószínűségi értékeket. Olyan eljárásokat is kidolgoztunk, amelyek használatával már az n. szó felismerésének befejezése előtt is rendelkezésre állnak közelítő valószínűségi becslések.

2. A beszédfelismerő modul felépítése

A modern beszédfelismerő rendszerek két fő modult tartalmaznak. Az akusztikus modul a beszédhangok felismerését végzi, a nyelvi modul pedig egyfajta vezérlő szereppel rendelkezik, a nyelvileg és nyelvtanilag valószínű szerkezeteket emelve ki.

2.1 Akusztikai modellek

A közép- és nagyszótáras felismerők mindegyikének gyakorlati megvalósításakor a legkisebb egység, amelyet a rendszernek fel kell ismernie, az a beszédhang. A szavak felismerése ezen alkotóelemek felismerésén keresztül valósul meg. A téma kutatása során több különböző gépi tanuló-osztályozó algoritmus [2] hasznát javasolták a nyelvi alapegységek és az összetettebb struktúrák (szavak, mondatok) felismerésének érdekében[4]. Az ilyen algoritmusokra épülő akusztikai modelleknek a két legfontosabb ága a HMM (Rejtett Markov Modell) alapú [1] illetve a szegmens alapú megközelítés [5].

A két irányvonalban közös, hogy a mikrofonból érkező digitalizált beszédjelből kis időközönként megfelelő méretű mintát veszünk, és minden ilyen kis jeldarabból bizonyos számú jellemzőt vonunk ki [7], amelyekkel az adott jeldarabot jól tudjuk jellemezni. A beszédfelismerésben használt jellemzőkinyerő algoritmusok száma igen nagy, amelyek közül az összetettebbek a hallás és a központi idegrendszer jelfeldolgozásának vizsgálatából származó tudományos eredményeket is figyelembe veszik [7].

A HMM és a szegmens alapú megközelítések abban térnek el leginkább, hogy a HMM egységnyi jeldarabokból (adatkeretekből) építkezik, míg a szegmens alapú a feltételezett (változó hosszúságú) fonetikai szegmenseket egyben modellezi. Cikkünkben, a nyelvi modellek összehasonlításához alkalmazott beszédfelismerő kere-

trendszerként egy saját implementálású, HMM alapú akusztikus modellt használtuk fel.

2.2 Nyelvi modellek

Az akusztikus modul önmagában beszédhangsorozatokat ismer fel. Az, hogy a beszédhangsorozatok közül – egy adott természetes nyelven – melyek felelnek meg értelmes szósorozatok (mondatok) fonetikus átiratainak, azt az alkalmazott nyelvi modell dönti el. A nyelvi modell feladata tehát a lehetséges beszédhangsorozatok halmazának a szűkítése, illetve az egyes beszédhangsorozatok valószínűségének megadása.

Nyelvi modelleket általában nem a teljes természetes nyelvre készítjük, hanem annak egy szűkebb, témaorientált részén. A legalapvetőbb nyelvi modell egyetlen szótárat tartalmaz csupán, és minden szó után minden szó következhet. Ez a felismerési pontosság, és a memóriahasználat szempontjából sem hatékony megoldás. A hipotézisek számának redukálása a felismerés sebességének és pontosságának nagymérvű javulását eredményezi. Azonban a keresési tér, azaz a modell által generált nyelv redukciójának általában az szab határt, hogy a modellnek le kell fednie azokat a – természetes nyelv szerint – helyes mondatoknak a többségét is, amelyek nem álltak rendelkezésre a modell építése közben. Másik lényeges szempont, hogy a modellnek „szűknek" kell lennie, azaz a nem valószínű szósorozatokat megfelelő módon „büntetnie" kell.

A nyelvi modellek többsége 2 fő csoportból, illetve ezek kombinációiból kerül ki. Az egyik csoport a formális nyelveken alapuló szabály alapú modelleket tartalmazza. Itt a szabályok a különböző ún. szócsoportok lehetséges követési sorrendjeit írják le, ahol a szóalakok egy vagy esetleg több szócsoportba vannak besorolva.

A nyelvi modellek másik nagy csoportját a statisztikus nyelvi modellek alkotják, amelyek között a legelterjedtebb modell az ún. szó N-gram. A szó N-gram az előző (n-1) szó ismeretében megadja, hogy a rákövetkező szónak milyen a statisztikai valószínűsége; ezen statisztika alapján közelítjük a $w_1,\ldots,w_n$ szósorozat valószínűségét [4], ahol w_0, …, w_{-N+2} a mondat kezdő szimbólum:

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}, \ldots, w_{i-N+1})$$

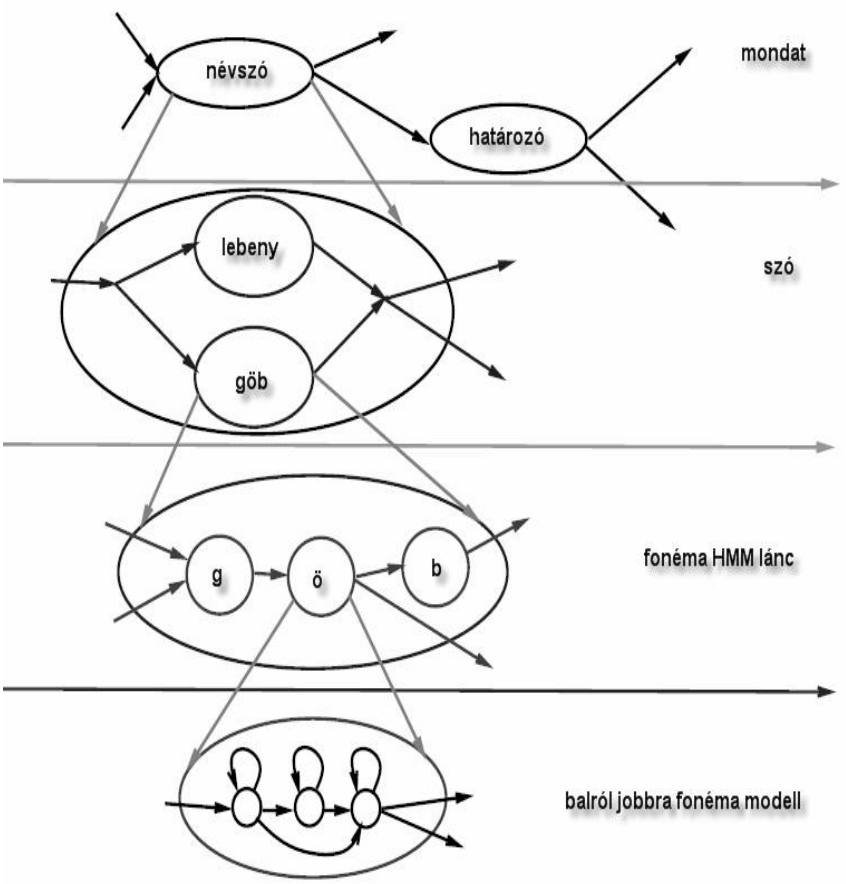
Az n növelésével egy ideig növelhető a nyelvtan megbízhatósága, de ez a memóriahasználat és a nyelvi modell tanítására használt szövegkorpusz exponenciális növekedésével jár. Emiatt a gyakorlatban 2-gram illetve 3-gram modelleket használnak. Az N-gram modellek tanítása a szövegkorpuszban található szó n-esek leszámolásával történik, azonban a rendelkezésre álló szövegkorpusz általában nem elegendő ahhoz, hogy tartalmazza kellő számban az összes bemondható helyes szó n-est, emiatt az elő nem forduló szó n-eseket rosszul modellezi a szó N-gram. A modell javítására különböző ún. N-gramsimító eljárást dolgoztak ki, amely a hiányzó vagy ritka szó n-esekhez is képes megadni valamilyen megfelelő értéket [4].

3 A nyelvi modellek kiértékelésének módszertani részletei

3.1 A kiértékeléshez használt akusztikai modul

A kiértékeléshez használt akusztikai modul egy saját HMM (Hidden Markov Model) [8] implementáció. Minden fonémához egy-egy három állapotú balról-jobbra topológiájú HMM-et rendelünk. A felismeréskor a nyelvtan által szolgáltatott beszédhangsorozat alapján HMM-ek lánca épül fel (1. ábra). Egy HMM-lánc felel meg egy hipotézisnek, azaz minden hipotézis egy-egy értelmű kapcsolatban áll egy adott beszédhangsorozattal.

1. Ábra: HMM láncot felépítő hierarchia.



A folytonos felismerő modul a hipotéziseket prioritási sorban tárolja, minden időponthoz egy-egy prioritási sor tartozik. A prioritási sort alkotó hipotéziseket pontérték szerint rangsoroljuk, amely értékeket az akusztikai modul által szolgáltatott

valószínűség, illetve a nyelvi modul szerinti pontozás határoz meg. A prioritási sor n-best vágása mellett a hipotéziseket Viterbi Beam típusú vágás segítségével is szűrjük.

A nyelvi modul adja meg, hogy az adott hipotézishez tartozó beszédhangsorozat mely fonémákkal folytatható. A nyelvi modelltől elvárjuk, hogy ezek a beszédhang kiterjesztések diszjunktak legyenek, ezáltal egy adott időponthoz tartozó prioritási sorban a hipotézisek nem ismétlődhetnek. A felismerő modul minden prioritási sornak csak az első legfeljebb *k* db hipotézisét terjeszti ki, addig, amíg be nem telik a következő időponthoz tartozó, rögzített méretű prioritási sor.

3.2 A kiértékeléshez használt nyelvi modul alapvető elemei

A saját fejlesztésű nyelvi modul egyszerre több nyelvi modell párhuzamos kiértékelésére képes. Alapvetően három független szinten képes N-gramot kiértékelni (a nyelvi modulunk értelmezi a környezetfüggetlen nyelvtani szabályokat is, de erre a jelen cikkben nem térünk ki). Az N-gramkiértékelés legalsó szintje a beszédhangok szintje (ezt jelen cikkünkben nem vizsgáljuk). A második szint a szóalakokra létrehozott N-gram. A szó N-gram – speciális esetként – tartalmaz ún. beágyazott csoportokat is, mint például a számok csoportja (pl. az 'egy', 'kettő', ..., 'kilenc', ... szavakat tartalmazó szótár szavai nem külön-külön kerülnek bele a szó N-gramba, hanem egyben, külön szabályokkal leírva). Az általunk javasolt és megvalósított 3. szintet – a bizonyos csoportok feletti N-gram modellt – a következő fejezetekben írjuk le részletesen.

3.3 A tanítás és tesztelés során használt adatbázisok

3.3.1 Az akusztikai modul tanításához használt adatbázis

Az akusztikai modul beszédhang szintű Rejtett Markov Modelljeinek tanításához a következő adatbázisokat használtuk fel:

1) MRBA beszédkorpusz: egy nagyméretű, 250 beszélő által bemondott, szegmentált adatbázis, amely összesen nagyságrendben 2000 mondatot tartalmaz. A hangadatbázis 70%-ban tartalmaz férfi, és 30%-ban női bemondást.
2) MBA beszédkorpusz: 250 beszélő által bemondott, szegmentált adatbázis, amely összesen 2500 mondatot tartalmaz. A hangadatbázis egyenlő arányban tartalmaz férfi-, illetve női bemondást. A beszélők között 50 iskoláskorú gyerek is volt.
3) OASIS-Mirigy: Az adatbázis korlátozott szókincsű és nyelvtanú mondatokat, 200 orvosi szcintigráfiás leletet tartalmaz, amely több mint 1100 mondatból, illetve kb. 11000 szóból áll.

3.3.2 A nyelvi modul tanításához és teszteléséhez felhasznált szövegkorpusz

A beszédfelismerő nyelvi modelljének létrehozásához egy pajzsmirigy-szcintigráfiás leletekből álló szövegkorpuszt használtunk. Az írásos vizsgálati anyagokat 1998 és 2004 között rögzítették. A 9231 leletet a különböző formátumokból egy közös

szöveges formátumra kellett konvertálnunk, majd többlépéses javítási folyamat következett. A vizsgálatokról készült minden egyes lelet a következő részeket tartalmazza:

a) fejléc
b) klinikai adatok
c) kérdés
d) előző vizsgálat
e) jelen vizsgálat
f) összefoglaló vélemény
g) aláírás

A szövegkorpuszból töröltük a hiányos leleteket az átvizsgálás során. A szöveges adatbázis létrehozásakor az a) és g) részek nem lettek felhasználva. A végleges adatbázis 8546 szövegből áll. A szövegekben 2500 szóalak fordul elő (számok és dátumok nélkül). Átlagosan 11 mondatot, és mondatonként 6 szót tartalmaz egy-egy lelet. A szavak számának mondatonkénti eloszlása nem normális eloszlást mutat, ami annak tudható be, hogy a közel 95000 mondat közül mindösszesen 12500 különböző mondatott tartalmazott az adatbázis.

A teszteléshez használt beszédkorpusz létrehozásakor az előbb említett szövegkorpusz bizonyos mondatai lettek beolvasva. Emiatt, a teljes átfedés elkerülése érdekében a nyelvi modell tanítását nem a teljes szövegkorpuszon, hanem annak egy részhalmazán végeztük el.

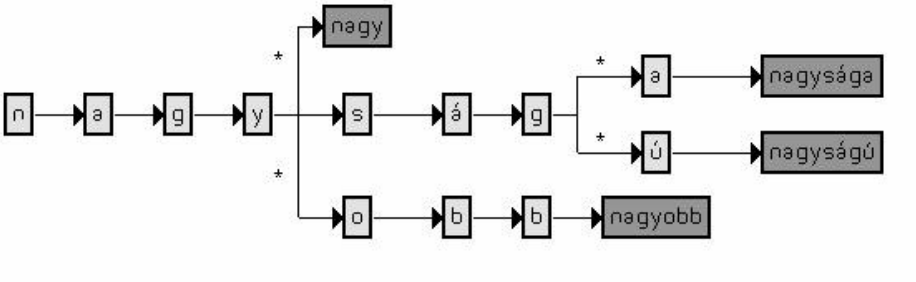
A nyelvi modellek tesztelése a Mirigy-valid nevű adatbázison történt, amely 100 szcintigráfiás orvosi lelet bemondását tartalmazza. Az összesen kb. 1000 mondat 5 beszélőtől származik.

4 Előrehozott N-gram-kiértékelés és -becslés

A szokásos N-gram kiértékelésekor egy szó n-es valószínűsége akkor derül csak ki, amikor már felismertük az n. szót is. Ha ezt a valószínűséget korábban – az n. szó vége előtt – tudnánk, akkor a hipotézisek nyelvi modul szerinti pontozása korábban megtörténne, és így csökkenthető lenne a hipotézisek száma. A megvalósítási lehetőségeket hatékonyság és memóriahasználat szempontjából vizsgáljuk meg a következő bekezdésekben.

4.1. A szótár hatékony reprezentálása

A nyelvi modellben előforduló szavakat egy tömör fastruktúrában tároljuk. A fa levelei maguk a szavak, és addig futnak közös ágon, amíg közös prefixszel rendelkeznek. Ez a tárolási módszer azért fontos, mert így minden csomópontban a lehetséges következő fonémák (többszöri előfordulás nélküli) halmaza könnyen lekérdezhető.



2. Ábra: Szavakat reprezentáló derivációs fa.

4.2 Az N-gram-kiértékelés előrehozása és becslése

A kiértékelés előrehozásának egyik kézenfekvő megoldása az, hogy a nyelvi modul a szavakat megadó derivációs fa utolsó elágazásánál már visszaadja a megfelelő N-gramértéket (lásd a 2. ábrán a csillaggal jelzett éleket). Ezt az egyszerű módszert tovább finomíthatjuk oly módon, hogy becslést adhatunk akár minden elágazásnál. Egy részfa valószínűsége felülről korlátozott, mivel a gyökeréből levezethető szavakhoz tartozó N-gram valószínűségek között van egy maximális. Ennek a meghatározására két egyszerű megoldás kínálkozik. Az egyik esetben, minden csomópontra azt is ráírjuk, hogy milyen szavak vezethetők le belőle. Ebben az esetben a csúcspont valószínűségének kiértékelésekor minden levezethető szóhoz kiszámítjuk az N-gramot, majd a maximumot adjuk vissza. Ennek a módszernek – a keresés miatt – a műveletigénye annál nagyobb, minél több szó vezethető le a csomópontból. Másik lehetőség, hogy előre kiszámítjuk és eltároljuk ezeket az értékeket minden, a becslésben részt vevő csomópontra. Mivel egy szó valószínűsége az előző (n-1)-től függ, így a csomópontokban az eltárolandó adatok száma legrosszabb esetben a szavak számának az (n-1). hatványa. Tehát ez a módszer gyors, de nagy tárigényű.

A valószínűségek előrehozott becslésével sok esetben egész hipotéziságakat tudunk levágni a felismerő modulban, tehát a nyelvi modul segítségével a valószínűtlen hipotézisek számát csökkenthetjük. Szó N-gram esetében a módszer kis-, illetve közepes méretű szótárak esetében lehet eredményes, a viszonylag nagy tárigénye miatt. A később javasolt MSD alapú csoport N-gramok esetében viszont a becsléses módszer jól alkalmazható, tárigénnyel kapcsolatos problémák nem merülnek fel.

4.3 Eredmények

A nyelvi modellek tanítása a 3.3.2 fejezetben leírtaknak megfelelően történt, a teljes szövegkorpusznak három véletlenszerűen kiválasztott részhalmazán (T1, T2, T3). Az 1. táblázat második oszlopában a hagyományos N-gram kiértékelés esetén kapott szófelismerési hibaarányt tüntettük fel. A harmadik oszlop tartalmazza az előre hozott

számítással kapott eredményeket, a 4. oszlop mutatja a relatív hibaarány csökkenését, amely azt mutatja, hogy a számítás előrehozása 12,3-18,4%-kal csökkenti a hibát.

1. Táblázat: A hagyományos (2. oszlop) és az előre hozott (3. oszlop) kiértékeléskor kapott szófelismerési hibaarányok, illetve a hibaarány csökkenése (4. oszlop)

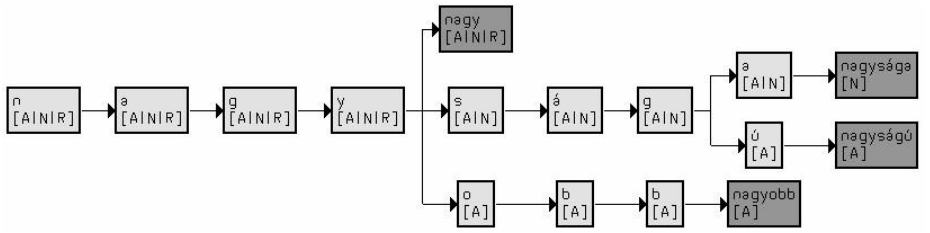
tanító korpusz	hagyományos kiértékelés	előrehozott kiértékelés	a hibaarány relatív csökkenése
T1	26.0%	22.8%	12.3%
T2	30.5%	26.3%	13.8%
T3	24,5%	20.0%	18,4%

5 MSD-kód alapú csoport N-gram

Az MSD kódolás (Morpho-Syntactic Description) minden szóhoz hozzárendel egy vagy több, változó hosszúságú karaktersorozatot, amelyek a szavak lehetséges mondattani szerepét írják le. A szavak MSD-kódjait a szegedi NLP csoport által kifejlesztett TnT alapú POS tagger programcsomag segítségével kaphatjuk meg [6]. Az MSD-kódok segítségével szabályokat hozhatunk létre a következőképpen: a szövegkorpusz mondataiban lévő szavakat lecseréljük azok mondattani szerepét leíró egyértelműsített MSD-kóddal. A cserével párhuzamosan szócsoportokat képezünk az ugyanazon MSD-kóddal rendelkező, azaz a mondatban ugyanabban a nyelvtani szerepben álló (pl. jelző, határozó) szavakból. Mivel egy-egy szónak más-más mondattani szerepe lehet a különböző mondatokban, így a csoportok nem feltétlenül diszjunktak. A folyamat végén minden mondathoz egy kódsorozatot rendelünk, ezek összessége adja meg az MSD-kódos nyelvtani szabályokat (az azonosak összevonása után). Az MSD-kód igen részletes leírást ad, így előfordulhatnak olyan csoportok, amelyekbe nagyon kevés szó esik. A szabályépítés közben, a hasonló mondattani szerepet leíró ritka kódokat érdemes összevonni, ezzel csökkenteni a csoportok számát.

MSD-kód alapú nyelvi modell nem csak konkrét, mondatokat leíró szabályhalmazként alkalmazható. Tesztjeinkben az MSD-kódolást csoport N-gram létrehozására használtuk. Az MSD-kód alapú leírásnál a szavak jelentése eltűnik, csak a szavak mondattani szerepe marad meg, így önmagában nem alkalmas nyelvi modellként. Tesztjeinkben azt vizsgáltuk, hogy alkalmazható-e, és milyen eredménnyel az MSD alapú csoport N-gram és a szó N-gram módszer kombinációja a felismerési pontosság javítására. Cikkünkben a módszerek lehetséges kombinációi közül az egyik legegyszerűbbet vizsgáljuk meg: a két modell által szolgátatott valószínűségi érték szorzata adja az aggregált értéket.

Csoport N-gramok esetében is lehetőség van előrehozott kiértékelésre. Az MSD alapú szócsoportok száma már nagyságrendekkel kisebb a szavak számánál (3. ábra), így az előre hozott becslés (4.2 fejezet) hatékonyan megvalósítható.



3. Ábra: Szavakat reprezentáló derivációs fa MSD-kódokkal kiegészítve.

A becslést a csoportok esetén hasonlóan végezhetjük el, mint szavak esetén. A tesztjeink során azt a megoldást választottunk, hogy a derivációs fa minden csomópontjához egy táblázatot rendelünk hozzá. A táblázat sorai a különböző előzményekhez tartoznak. A táblázatok kitöltéséhez minden egyes csomóponthoz meghatározzuk, hogy milyen csoportba tartozó szavak vezethetők le ebből a részfából. A csomóponthoz hozzárendelt táblázat ezután egyetlen oszlopot tartalmaz, melyben a csoport N-gram által meghatározott maximális levezethető érték szerepel (2. táblázat). A hipotézisek kiterjesztésekor a derivációs során előre haladva a levezethető csoportok száma szűkül, így a maximális érték csökken. Minden olyan csomópontban, ahol a maximális érték csökken, a nyelvi modul megadja a megfelelő arányszámot.

2. Táblázat: A 3. ábrán látható derivációs fához tartozó, MSD alapú csoport N-gram alapján számított lehetséges táblázatok. A 2. oszlopok tartalmazzák a maximummal becsült valószínűségi értékeket a fejlécében megadott csoportokra vonatkoztatva, az első oszlopokban feltüntetett előzményeket véve (- a kezdést jelenti).

	[A]
-, -	0,4
-, A	0
-, C	0
-, V	0
…	
M, N	0,25
N, A	0,20
R, A	0,16
…	

	[A|N]
-, -	0,4
-, A	1,0
-, C	0
-, V	1
…	
M, N	0,25
N, A	0,52
R, A	0,65
…	

	[N]
-, -	0,2
-, A	1,0
-, C	0
-, V	1
…	
M, N	0,005
N, A	0,52
R, A	0,65
…	

	[A|N|R]
-, -	0,4
-, A	1,0
-, C	0
-, V	1
…	
M, N	0,47
N, A	0,52
R, A	0,65
…	

5.1 Eredmények

Az MSD alapú csoport N-grammal kapcsolatos tesztjeinket az előző fejezetben már leírt tanító és tesztkorpuszokon végeztük. A 3. táblázat 2. oszlopa a csak szó N-gram használatakor kapott szófelismerési hibaarányt tartalmazza (előre hozott számítást

alkalmazva). Az MSD alapú csoport N-gram előre hozott számításával kapott teszteredményeket a 3. oszlop mutatja. Az utolsó oszlopban lévő eredményeket a csoport N-gram előrehozott becslésével kaptuk. A táblázatból kiolvasható, hogy minden esetben javított a csoport N-gram modellhez való hozzávétele, illetve a becslés. A 4. táblázatban a szófelismerési hiba relatív csökkenését foglaltuk össze.

3. Táblázat: Szófelismerési hibaarány MSD-kód alapú csoport N-gram használata nélkül (2. oszlop), csoport N-gram előre hozott számításával (3. oszlop), valamint folyamatos becslés esetén (4. oszlop).

tanító korpusz	csoport N-gram nélkül	előre hozott csoport N-grammal	becsült csoport N-grammal
T1	22.8%	21.4%	18.8%
T2	26.3%	21.2%	19.0%
T3	20.0%	17.5%	15.7%

4. Táblázat: Szófelismerési hibaarány relatív csökkenése csoport N-gram előre hozott számításakor (2. oszlop), valamint folyamatos becslés esetén (3. oszlop).

tanító korpusz	előre hozott csoport N-grammal	becsült csoport N-grammal
T1	6.1%	17.5%
T2	19,4%	27.8%
T3	12.5%	21.5%

6. Összefoglalás

Cikkünkben olyan módszereket adtunk meg, amelyek segítségével a hagyományos szó N-gram alapú nyelvtanokkal elérhető szófelismerési hibaarány csökkenthető. Tesztekkel igazoltuk, hogy a szó N-gram kiértékelésének előrehozásával a felismerés pontossága növekszik. A hagyományos szó N-gram rosszul kezeli azokat a bemondásokat, amelyek nem voltak benne a tanításához használt korpuszban. Ennek kiküszöbölésére cikkünkben a szó N-gram és az MSD típusú csoport N-gramértékeinek aggregációját javasoljuk. A teszteredményeink alapján a felismerés hibája nagymértékben csökken az aggregációs technika használatakor. Összességében az aggregációs technikával és a csoport N-gram, valamint a szó N-gram előrehozott számításával akár több mint 30%-os szófelismerési hibaarány-csökkenés érhető el.

Bibliográfia

1. C. Becchetti, L. P. Ricotti: Speech Recognition, John Wiley & Sons LTD, Chichester, England (2000)

2. R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification, Wiley (2001)

3. Erjavec, T., Monachini, M., (ed.): Specification and Notation for Lexicon Encoding. Copernicus Project 106 "MULTEX-EAST", Work Package 1 - Task 1.1, Deliverable D1.1F, (1997)

4. X. Huang, A. Acero, H. Hon: Spoken Language Processing, Prentice Hall, New Jersey (2001)

5. A. Kocsor, A. Kuba, L. Tóth , M. Jelasity, L. Felföldi, T. Gyimóthy, J. Csirik: A Segment-Based Statistical Speech Recognition System for Isolated/Continuous Number Recognition, Proceedings of the FUSST'99, Aug. 19-21, Sagadi, Estonia, 201-211, (1999)

6. Kuba A., Hócza A., Csirik J.: POS Tagging of Hungarian with Combined Statistical and Rule-based Methods in Proc. of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), Brno, Czech Republic 8-11 September, pp. 113-121 (2004)

7. L. R. Rabiner, R. W. Schafer: Digital Processing of Speech Signals, Prentice-Hall, Englewood (1978)

8. V. N. Vapnik: Statistical Learning Theory, Wiley (1998)

9. Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben, II. Magyar Számítógépes Nyelvészeti Konferencia, (2004)

Középszótáras folyamatos beszédfelismerőrendszer fejlesztési tapasztalatai

Vicsi Klára, Velkei Szabolcs, Szaszák György, Borostyán Gábor, Teleki Csaba,
Tóth Szabolcs Levente, Gordos Géza

BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
1117 Budapest, Magyar tudósok körútja 2.
vicsi@tmit.bme.hu

Kivonat: A Beszédakusztikai Laboratóriumban kifejlesztésre került egy Windows XP alatt működő, statisztikai elvi alapokra épülő, folyamatos beszédfelismerő fejlesztői környezet (MKBF 1.0), amely alkalmas különböző középszótáras 1000-10 000 szavas szövegek betanítására és felismerésére. Új megoldásokat dolgoztunk ki az akusztikai előfeldolgozásban, a statisztikai modellépítésben valamint fonetikai, fonológiai és morféma nyelvi szinteket vonunk be a felismerési folyamatba. A felismerő a statisztikai alapon működő HMM akusztikai fonémamodellekkel valamint a statisztikai alapú bigram nyelvi modellekel működik, nem lineáris simítást használva. Vizsgálataink során változtattuk a betanító anyagokat, a szótárkészletet. Kétfajta bigram alappal dolgoztunk: először a hagyományos ragozott szóalakokból építettük fel a bigram mezőket, majd a szóalakokat morfémákra bontottuk, és ezekből a morfémákból építkeztünk. A cikkben a tesztelés eredményeiről, a továbbfejlesztéshez nyert tapasztalatainkról számolunk be. A perplexitási vizsgálatok eredményeinek felhasználásával a felismerési biztonságot 70%-ról 91% fölé tudtuk vinni.

1 Bevezetés

A BME Beszédakusztikai Laboratóriumban kifejlesztett folyamatos beszédfelismerő (MKBF 1.0) optimális működését az akusztikai, fonetikai [4] és nyelvi modellek változtatásával állítottuk be. Természetesen a két szint szétválasztása nem mindig lehetséges, hiszen a tesztfelvételek minősége, zajossága, az artikuláció gondossága, stb. mind befolyásolják a felismerés eredményét, így az nem csak a nyelvi modultól függ.

A felvételek mindegyike – mind a betanításnál, mind a tesztelésnél – 16 kHz-en mintavételezett, 16 biten lineárisan kvantált jel, amely a megfelelő előfeldolgozás után kerül felismerésre.

Az **akusztikai modellek betanítását** az MRBA beszéd adatbázissal végeztük [9].

Végeredményben tehát a fonémaszintű felismerőnk 16 kHz mintavételezésű, 17 Bark frekvenciatérbeli derivált, + 17 időbeni derivált, + 17 időbeni második derivált, + energia bemeneti jelvektor mellett, 4-5 állapotú kvázi-folytonos, 24 lépcsős, rejtett Markov-modellekkel (QCHMM), fonéma alappal dolgozik. Az akusztikai, fonetikai szint optimalizálásáról már korábban beszámoltunk [8] .

A **nyelvi betanításhoz** a budapesti SOTE II. sz. Belgyógyászati Klinikájától (2700 lelet) és a szegedi Orvostudományi Egyetemről (6365 lelet) gyűjtött korábbi leletanyag korpuszt használtuk. Ezen szöveg korpusz alapján készítettük el a teljes szóalakszótárat, amely 14 331 szót tartalmaz, a kiejtés szótárat és ezek téma szerint osztott kisebb szótárait. A valamint a korpusz alapján morfémaszótárat is készítettünk, amelynek nagysága 6 824 morfémaelem.

Teszteléshez az orvosok által bemondott leletanyagot használtuk, ezek a SOTE II. sz Belgyógyászatán készültek, szakorvosok bemondásával, a rendszerhez illeszkedő mintavételi és kvantálási paraméterekkel. Az összes felvételből a férfi orvosok bemondásaiból véletlenszerűen, öt beszélőtől egyenként négy-négy darab, azaz összesen 20 darab gasztroszkópiás felvételt válogattunk ki tesztelési célokra.

Lényegében a nyelvi modellhez bigram modelleket használtunk, de az egyik megoldásban a hagyományos szóalakok (lexémák) az alkotó elemek, a másik megoldásban viszont a morfémák. A morfémabontáshoz a Humor morféma elemzőt használtuk fel [5].

2 Bigram nyelvi modellek

2.1 Az endoszkópiai felismerő nyelvi modelljének leírása

A sokféle szómodell közül az angolszász területeken jól bevált n-gram szómodelleket használtuk a nyelvi szintű felismeréshez. Az n-gram modellek segítségével egy tetszőleges korpuszon minimálisra igyekszünk csökkenteni a perplexítás mértékét, aminek következménye a kevesebb hibázás.

Az n-gram modell szószekvenciáik valószínűségének halmazából áll:

$$\hat{P}\ (\ w_{\ 1}\ ,\ w_{\ 2}\ ,...,\quad w_{\ m}\ ) \tag{1}$$

A szekvencia valószínűsége ekkor:

$$P(w_1, w_2, ... w_m) = P(w_1)\prod_{i=2}^{m} P(w_i | w_{i-1} ... w_1) \tag{2}$$

A kontextust limitálva:

$$P(w_1, w_2, ... w_m) \cong P(w_1)\prod_{i=2}^{m} P(w_i | w_{i-1} ... w_{i-n+1}) \tag{3}$$

ahol n>0 tetszőlegesen választott konstans egész. A nyelv olyan tulajdonságokkal rendelkezik, hogy a folyamat során egy későbbi állapot valószínűsége gyakorlatilag

független a kezdőfeltételektől, így n értékére nem kell nagy *n* értéket használni. (Tipikus értékek 2-től 6-ig). A fenti valószínűség ekkor a következőképpen számítható ki:

$$P\left(w_i \middle| w_{i-1} \ldots w_{i-n+1}\right) = \frac{N\left(w_i \ldots w_{i-n+1}\right)}{N\left(w_{i-1} \ldots w_{i-n+1}\right)} \qquad (4)$$

ahol *N(.)* a megadott szekvencia előfordulásai száma a tanító szöveganyagban. Ehhez a számításhoz nem kell szegmentált hanganyagot használni, a célra legmegfelelőbbek a nagyméretű szöveges adatbázisok.

2.2…N-gram modellek simítása

A gyakorlatban a lehetetlen méretű adatbázisok készítése helyett az n-gram mo-dellek statisztikai vizsgálata és különböző módszerekkel történő korrigálását alkalmaztuk[6].

A korrigálásra nemlineáris interpolációt használtunk. Mivel utóbbi esetben lényegesen jobb perplexítás-csökkentés érhető el, ezért a nemlineáris interpolációt használtuk [7] , az *absolute discounting* funkciót. Tekintsük most példaként a bigram esetet, ahol a képlet a következő konkrét alakot ölti:

$$\hat{P}\left(w^{(j)} \middle| w^{(i)}\right) = \max\left\{\frac{N\left(w^{(j)}, w^{(i)}\right) - D_i}{N\left(w^{(i)}\right)}, 0\right\} + D_i \frac{|V| - n_0\left(w^{(i)}\right)}{N\left(w^{(i)}\right)} P\left(w^{(j)}\right) \qquad (5)$$

($|V|$ a szótár számosságát jelöli, $n_0\left(w^{(i)}\right)$ pedig azon szavak számát, amelyek egyszer sem követték $w^{(i)}$-t.) A nemlineáris interpoláció esetében a *q(k)* eloszlás súlya arányosan megfelel *(K-n_0)*-val, ami azon különböző események száma, amelyek legalább egyszer láthatóak voltak a szöveganyagban. Ez érdekes dologhoz vezet a feltételes valószínűségek modellezésekor: ha a megelőző szót (*predecessor word*) egy, vagy csak néhány szó követi, akkor a simítás 'kisebb' mértékű lesz, mintha sok szó követné azt. Erre utal a nemlineáris kitétel a módszer nevében. Ha *D=1*, akkor az egyszer látott eseményeket ugyanúgy kezeli az algoritmus, mint az egyszer sem látottak. Ha alkalmazzuk a *Leaving-one-out* elvet, akkor nem jelentkezik igazán lényegi különbség a perplexitásban, ezért a D értékét a lényegesen egyszerűbben kivitelezhető abszolút modell alapján számítjuk, ahol:

$$D_i = \frac{|V| \cdot b}{n_0\left(w^{(i)}\right)}, \text{ ahol } \quad b = \frac{n_1}{n_1 + 2n_2} \qquad (6)$$

Itt n_1 és n_2 azon bigramok száma, amelyek pontosan egyszer, illetve kétszer szerepeltek a betanító korpuszban. Kis betanító anyag esetén $\frac{|V|}{n_0\left(w^{(i)}\right)}$ értéke közelítőleg 1, ezért ilyen korpuszok esetén lehet spórolni a számításokkal és D_i=*b* helyettesítést végezni [6].

3… Nyelvi modell tesztelése, perplexitás vizsgálata

A bigram modellek elkészítéséhez – ún. betanításához – nagy méretű, szöveges, a felismerni kívánt szöveget jól közelítő összetételű és stílusú betanító anyagra van szükség. Esetünkben ez a korábban összegyűjtött és megfelelően feldolgozott (helyesírás ellenőrzés, egységesítés, rövidítések feloldása, fonetikai átírás, stb.) leletanyag volt. Ezt a leletanyagot négy csoportra osztottuk az alábbiak szerint:

1. *SOTE II. sz. Belgyógyászatról származó felső endoszkópiás leletanyag (budapesti gasztroszkópia)*
2. *SZTE Belgyógyászatáról származó felső endoszkópiás leletanyag (szegedi gasztroszkópia)*
3. *SOTE II. sz. Belgyógyászatról származó alsó endoszkópiás leletanyag (budapesti kolonoszkópia)*
4. *SZTE Belgyógyászatáról származó alsó endoszkópiás leletanyag (szegedi kolonoszkópia)*

A fenti négy csoportból természetesen lehetőség van kombinált anyagok összeállítására is, amely így nagy mennyiségű betanító anyagot szolgáltathat a bigram nyelvi modellezéshez.

MKBF akusztikai szint betanításait az: MRBA adatbázis férfi bemondásaival végeztük.

3.1...Tesztelési körülmények ismertetése:

A tesztelés megkezdése előtt felvetődött az a kérdés hogy a rendelkezésünkre álló betanító anyagok közül melyeket használjuk fel a felismerő nyelvi modelljének a betanítására. Az előzetes mérések alapján (1. táblázat) látható, hogy a rendelkezésre álló budapesti és szegedi leletek szókészlete kis mértékben korrelálnak egymáshoz.

1. táblázat: Szókészletek összehasonlítása

		Szegedi colonoscopia		**Szegedi gastroscopia**	
		szó megvan	**szó nincs meg**	**szó megvan**	**szó nincs meg**
Budapesti colonoscopia	**szó megvan**	**1933**	**1174**	**1872**	**1235**
	szó nincs meg	**5089**	**6135**	**7067**	**4157**
		Szegedi gastroscopia		**Szegedi colonoscopia**	
		szó megvan	**szó nincs meg**	**szó megvan**	**szó nincs meg**
Budapesti gastroscopia	**szó megvan**	**2720**	**1594**	**2065**	**2249**
	szó nincs meg	**6219**	**3798**	**4957**	**5060**

A későbbi vizsgálódások azt is megmutatták, hogy a felismerő tesztelésére kijelölt hanganyagok szótárkészlete újabb szavakat tartalmazott az írásos formában rendelkezésünkre álló budapesti- és szegedi endoszkópos leletekhez képest. A fenti táblázat egyértelműen mutatja, hogy a szegedi és budapesti leletek szóhasználata

között milyen nagy eltérés van és a tesztelésre kijelölt annotált felvételek budapesti kórházból származnak. Ennek ellenére a fent említett okok miatt az összesített leletanyaggal való betanítás ígérkezett megfelelőnek.

A tesztelésnél használatos mérőszámok:

Össz _ref: *a felismerendő egységek száma,* ***Össz_ rec:****a felismert egységek száma,*
Helyes: *a jól felismert egységek száma,* ***Ins:****a beszúrt egységek száma,*
Del: *a törölt ,* ***Subs:****a helyettesített egységek száma*

$$\textbf{CORR} = \frac{Helyes}{Össz\_rec}, \quad \textbf{Acc} = \frac{Helyes - Ins}{Össz\_rec}, \quad \textbf{Wer} = 1 - CORR$$

3.2…Perplexitás alapú WER becslés

A gépi beszédfelismerés felismerési pontosságát a szakirodalomban – a fentiekben leírt a Word Error Rate *(WER)* indikátorral szokásos jellemezni. A Word Error Rate egy költséges művelet eredménye, ezért szükségessé vált egy olyan indikátor bevezetése, amely a beszédfelismerés akusztikai szintjétől függetlenül becslést tudna adni a felismerés pontosságára. Így a nagy felismerési idő és a költséges WER számítás kikerülhető lenne a beszédfelismerés nyelvi modelljének vizsgálata esetén. Egy ilyen becslési módszer a – szakirodalomban is jól ismert – perplexitás, melynek segítségével vizsgálni tudjuk a nyelvi modellt. Bár különböző kutatások rávilágítanak hogy készíthető paraméterfüggő (betanítás, nyelvi modell, akusztikai modell) becslési eljárás [3,4] , mégis a konkrét paraméterek ismeretének hiányában a perplexitást találtuk olyan becslési eljárásnak, amely a szakirodalomban elfogadott és számítási módja ismert. A perplexitás számítási módját az alábbi képletben ismertetem:

$$\textbf{PP} = \frac{1}{\left(\prod_{i=1}^{N} P(W_i \mid W_{i-1})\right)^{\frac{1}{N}}} \qquad (7)$$

ahol W_i teszt szekvencia $i.$ **szava ;**, W_{i-1} a teszt szekvencia $i-1.$ **szava**, N a teszt szekvenciát alkotó **szavak** száma.

A perplexitás képletét bigram alapú nyelvi modell formájában használtuk. A **szavak** jelentése lexéma szintű felismerés esetén lexéma, míg morféma alapú felismerés esetén morféma.. A perplexitás értékkészlete egy 1-nél nagyobb valós szám. A tesztanyag nyelvi modul általi felismerése annál tökéletesebb, minél jobban közelít a perplexitás értéke 1-hez. Minél nagyobb a perplexitás értéke, annál kevésbé fedi a nyelvi modell a tesztelő szöveget.

4…A tesztelési eredmények ismertetése

A következő táblázatokban a felismerő tesztelési eredményei láthatóak:

2. táblázat: Gasztroszkópiás felvételek tesztelési eredményei lexéma alapú összegzett betanítású nyelvi modell esetén, orvosok bemondásában

Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER
1173	1580	750	451	22	401	25,4	36,1

3. táblázat: Colonoszkópiás felvételek tesztelési eredményei lexéma alapúösszegzett betanítású nyelvi modell esetén, orvosok bemondásában

Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER
890	1326	504	822	8	370	-35,7	**43,4**

A viszonylag rossz eredmények oka (tipikusan a kötőszavak tévesztése nagy), hogy bár a szótárkészlet ezen betanítóanyag választása mellett biztosítja a legnagyobb fedést, ennek ellenére bigram szókapcsolatok nem fedik a tesztelési bigram szókapcsolatokat. Ha megfigyeljük az 1. táblázatbeli eredményeket, és összevetjük a tapasztaltakkal, akkor megállapíthatjuk következtetésképpen, hogy a teljes anyaggal történő betanításkor (szegedi, budapesti, gasztroszkópiai, colonoszkópiai) olyan nagy mértékű hamis szókapcsolat-statisztikát vittünk be a rendszerbe, hogy az a bigram valószínűségi mezőben zaj keletkezett, így hiába lettek betanítva ezen szókapcsolatok, mégis rossz lett a felismerés. (lásd 1. táblázat budapesti és szegedi szókészlet eltérését.)
Szűkítve a betanítási anyagot, a betanításra csak a budapesti gasztroszkópiás betanítóanyagot az eredmények javulnak, amint azt a 4. táblázatban mutatjuk.

4. táblázat: Gasztroszkópiás felvételek tesztelési eredményei lexéma alapú budapesti gasztroszkópiás betanítású nyelvi modell esetén, orvosok bemondásában

Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1150	1417	799	283	8	343	44,8	30,5	*73,59*

A 2 táblázat eredményei orvosok bemondásai alapján elkészített tesztelési eredmények. A lelet-felvételek meghallgatásakor azt tapasztaltuk, hogy a felvételek igen zajosak és kiejtés szempontjából is igen rossz minőségűek. Így felvetődött azon lehetőség - a nagyobb felismerési pontosság elérése érdekében - hogy limitált szintű zajkörnyezetben felvett felvételekkel teszteljünk és a felvétel során ügyeljünk a helyes artikulációra. Ennek érdekében a budapesti gasztroszkópiás leleteket –amelyeket az orvosok is bemondtak (20 lelet)– bemondtuk a laboratóriumban és ezen felvételeket használtuk fel az MKBF tesztelésére. A tesztelési eredményeit az 5. táblázatban közöljük.

5.táblázat: Kiszajú,laboratóriumi gasztroszkópiás felvételek, tesztelési eredményei lexéma alapú budapesti gasztroszkópiás betanítású nyelvi modell esetén

Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1173	1451	922	280	1	250	54,7	21,3	*73,59*

Az eredményeket összevetve a 4. táblázatéval, látható hogy a szótévesztési arány (WER) javult, mivel az akusztikai szintű felismerés javult, ezáltal az egész felismerés is pontosabbá vált.
Továbbá olyan leletbemondással tesztelünk, ami szerepelt a nyelvi modell betanításában. Ennek megvizsgálása érdekében – a fentiekben említett akusztikai feltételek biztosítása mellett – budapesti gasztroszkópiás leleteket rögzítettünk, amelyeket a budapesti gasztroszkopiás írott leletanyagból olvastunk fel, és a nyelvi modellt a 5. táblázathoz hasonlóan budapesti gasztroszkópiás írott leletanyaggal tanítottuk be. Az eredményeket a 6. táblázatban adjuk meg.

6. táblázat Laboratóriumi, budapesti, gasztroszkópiás felvételek tesztelési eredményei lexéma alapú, budapesti, gasztroszkópiás, betanítású nyelvi modellekkel

Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER	PP
416	444	380	28	0	36	84,6	8,6	9,36

A 6. táblázat egyértelműen mutatja, hogy a szótévesztési arány (WER) javul, hiszen olyan leletanyagot teszteltünk, amely a betanításban szerepelt. Ha a perplexitást vizsgáljuk, akkor 4. és 5. táblázatokban található perplexitás-értékekhez képest nagy arányú perplexitás csökkenést tapasztalhatunk.

4.1 A lexéma alapú tesztelési eredmények kiértékelése

Az eredmények alapján az alábbi következtetéseket vonhatjuk le.

1. A budapesti és a szegedi leletanyag annak ellenére hogy mind a két korpusz azonos témájú, azonban használatban, stílusban jelentősen különböznek, hogy a vegyes anyag alapján készített bigram nyelvi modellel, noha elviekben robosztusabb, a gyakorlatban mégis gyengébb felismerés érhető el. A *WER* eredmény 5.54%-kal rosszabb a csak budapesti anyag alapján tanított bigramhoz képest.
2. A beszédfelismerés során különösen fontos a szöveg gondos, folyamatos bemondása, így a *WER* értéke akár 10%-kal is lejjebb szorítható.
3. A jelenlegi, budapesti gasztroszkópiás leletanyag alapján készült bigram nyelvi modell nem fedi kellőképpen a kívánt alkalmazási területet. Ezt igazolja a szakorvosok és a saját bemondásban készült leletek hibaarányainak nagyfokú korrelációja, azaz a hibaarány jelentős része a bigram nem megfelelő fedéséből, és nem az akusztikai jel minőségéből adódik.
4. Megjegyezzük, hogy a szakorvosok által végzett bemondásokban a PI kódjelű bemondó halk, az általánosan elvárhatónál gyengébb beszédproduktumot adott a felvételek során. Amennyiben az ezekkel a felvételekkel kapott hibaarányt figyelmen kívül hagyjuk, az összesített *WER* értéke 30,52%-ról 24,27%-ra javul, amely utóbbi véleményünk szerint a hitelesebb adat.
5. Az *Acc* paraméter esetenkénti alacsony értéke arra enged következtetni, hogy az adott tesztfelvétel a nyelvi modell számára ismeretlen, vagy a bigram betanító anyagában nem kellő mértékben előfordult szót tartalmaz. Ilyenkor a beszúrások megszaporodnak, amely jellemzően több rövid, felolvasva az elhangzó, de fel nem ismert szóéhoz hasonló hangélményt ad.

4.2 A perplexitás vizsgálata és ennek összevetése a lexéma – valamint morféma alapú beszédfelismerő tesztelési eredményeivel

A 1.2 fejezetben ismertettük és az 1.3 fejezet utolsó eredményeivel szemléltettük, hogy a perplexitással jósolható a felismerés pontossága. Ebben a részben azt vizsgáltuk, hogy a felismerés hatékonysága mennyire jósolható a nyelvi modell szimulálására szolgáló perplexitás-számítással. Ennek vizsgálata érdekében megmértük, hogy mennyire konvergálnak a perplexitás alapú becslési értékek a Word Error Rate értékekkel abban az esetben, amikor olyan anyaggal tesztelünk, ami részét képezi a nyelvi modell betanításának. A budapesti gasztroszkópiás leletekkel kapott mérés eredményeit 1. ábrán a szemléltetjük.

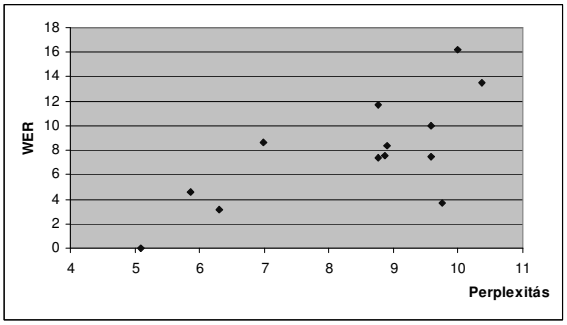


Fig1. Perplexitás és WER értékek korrelációja kiszajú, laboratóriumi, budapesti gasztroszkópiás felvételek esetén, budapesti gasztroszkópia nyelvi modell tanítva

Amint az 2. ábra eredményei szemléltetik, a perplexitás – WER értékpárosok által meghatározott pontok szórnak, de az összefüggés a perplexitás és a WER értékek között jól látható [1]. Az, hogy az értékek szórnak, érthető, hiszen nem a teljes magyar nyelvet vizsgáltuk, hanem csak egy igen szűk tématerületet, ami – a tapasztalatok alapján – a szakterületi mellett hétköznapi, valamint irodalmi nyelvet is tartalmaz. A másik lehetséges oka abban rejlik, hogy a perplexitásszámítás folyamán nem számolunk a fonématévesztéssel és a fonématévesztés hatására eltolódik a WER-Perplexitás kapcsolat [4].

4.3 Lexéma vagy morféma felismerés

A bigram statisztikákat egyszer lexéma egyszer morféma alapokon számítottuk. A tesztelések hasonló eredményekre vezettek mindkét esetben.

Ami a morféma alapú felismerés mellett szól:

Ha a morféma szintű felismerést választjuk a szótárméret jelentősen csökken, így kisebb bigram valószínűségi mezőt kell kezelni. Lexéma alapú betanítás esetén a leletkorpusz alapján 14331 (lexéma) egység jön létre, míg morféma alapú betanítás esetén 6706 egység (morféma).

Mivel a bigram valószínűség mező egy négyzetes diszkrét valószínűségi mező, így tárolási szempontból körülbelül 4,5 szeres tárcsökkenést eredményez, és még a leg-

rosszabb esetben (simítás esetén) is már átlagosan 2,13-szeres valószínűségérték növekedés érhető el.

8. táblázat: tesztelési eredmények, betanítás a budapesti gasztroszkópiás anyaggal, tesztelés kiszajú, laboratóriumi, budapesten felvett gasztroszkópiás felvételekkel

Morféma alap								
Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1631	2045	1241	778	9	355	28,3	23,9	27,31
Lexéma alap								
Össz ref	Össz rec	Helyes	Ins	Del	Subs	Acc	WER	PP
1173	1451	922	280	1	250	54,7	21,3	*73,59*

Ami a lexéma szintű felismerés mellett szól:
Morféma szintű felismerés esetén komoly problémát jelent a toldalékok határain fellépő hasonulások, összeolvadások, hangrendilleszkedések hangzókiesések kezelése. Ennek leírása egyelőre úgy tűnik, csak manuálisan oldható meg.

5 Felismerési pontosság növelése perplexitás alapú szimulálás segítségével

Amint azt az 1. fejezetben ismertettük, a perplexitással becsülni lehet a felismerés pontosságát. A 5. táblázat egy olyan tesztelési eredményeket tartalmazó táblázat, ahol a tesztelésnek kinevezett állomány nem szerepelt a nyelvi modell betanító leletei között. A 6. táblázat viszont olyan tesztelési eredményeket tartalmaz ahol a tesztanyag részét képezte a nyelvi modell betanító anyagának, tehát biztosítottak voltak azon szókapcsolatok betanításai amelyek a tesztelésnek kinevezett anyagokban szerepeltek. Ha a 5. táblázat eredményeit összevetjük a 6. táblázat eredményeivel, azt tapasztalhatjuk hogy a hibaszázalékok kisebbek a 6. táblázatban.
Így felvetődött az a kérdés, hogy ha betanítanánk a 5. táblázathoz tartozó tesztelési mondatokból azon szókapcsolatokat amelyek nem szerepeltek a betanításnál, akkor a felismerési pontosság várhatóan növekedni fog e. Ehhez csupán ezen hiányzó szókapcsolatokat kell megkeresni és a betanítóanyagban elhelyezni.
A hiányzó szókapcsolatok betanításánál azt a technikát választottuk, hogy a tesztelő anyagból azon **szóláncokat** kerestük meg, amely szóláncok bármely bigram szókapcsolatát tekintve, egyik sem szerepelt a betanításban. Tehát a szóláncok kiválasztása a következő:

<utolsó betanításban szereplő szó> <betanításban nem szereplő szó>$^{+}$ <első olyan szó ami a betanításban szerepelt>

+ jel jelenti, hogy 1-nél többször is szerepelhet egymás után betanításban nem szereplő szó, a reguláris kifejezéseknél használt jelöléssekhez hasonlóan

Ezen módszer választása mellett a szándékunk az, hogy a hiányzó bigram valószínűségeket betanítsuk anélkül hogy a már meglévő bigram valószínűségeket jelentősen torzítanánk.

Azt tudjuk, hogy mely részekkel kell a betanítást bővíteni, azonban azt nem, hogy ezen hiányzó részek betanítását hányszor kell megismételni.Az ismétlések számának

meghatározása az általunk használt perplexitás alapú nyelvi modell haté-konyságának becslése alapján történt (6. ábra). Előállítottunk különböző betanító anyagokat, amelyek felépítésüket tekintve a következőképpen alakultak:

Betanítás=<budapesti gasztroszkópiás leletek> + <hiányzó **szóláncok** >*

** jel jelenti, hogy a hiányzó szóláncok 0..n- szer szerepelhetnek a budapesti gasztroszkópiás leletek után, reguláris kifejezéseknél használt jelöléssekhez hasonlóan.*

A perplexitás értékeket különböző betanító anyagoknál nem lehet összehasonlítani, esetünkben viszont az összehasonlítás elvégezhető, mivel a betanítóanyagot csak kismértékben módosítottuk, szókészletek megegyeznek, a bigram valószínűségi mező csupán eloszlási értékeiben csak kismértékben különbözik egymástól.

5.1 A betanításszám meghatározása nyelvi modell szimulálás segítségével

Kiválasztottunk 4 leletet tesztelésre, a betanítóanyaggal összehasonlítva meghatároztuk a hiányzó szóláncokat. A betanításnál ezen hiányzó szóláncokat szerepeltettük 0..n –szer. Ezen leletek mellett figyeltük a többi leletet is, hiszen a cél a bigram valószínűség mező felismerésének erősítése, nem pedig a torzítása. Amint a 2.1 ábrából megfigyelhető a hiányzó szóllánc ismétlési számának növelésével a perplexitás értékek javulnak azon tesztanyag esetén, amely alapján a hiányzó szólánc elő lett állítva. A 2.2 ábrán szemléltettem azon gasztroszkópiás leletek perplexitás értékeinek alakulását, amelyekből nem lett hiányzó szólánc véve. Látható, hogy a 19-20 –szoros betanításig folyamatos perplexitás javulás van, e feletti betanításnál viszont csak romlás tapasztalható.

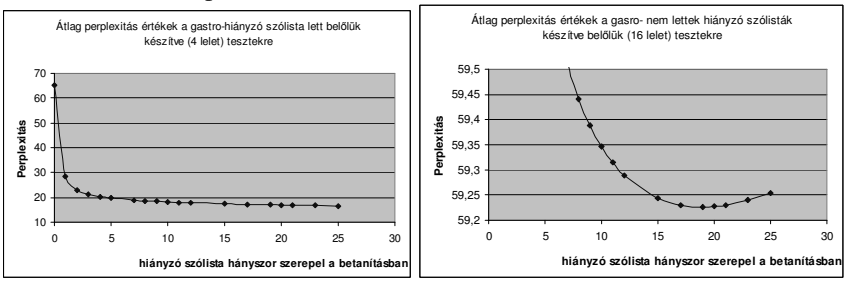


Fig2.1-2.2: *Perplexitás átlagok alakulása*

A hússzoros betanítással kiegészített budapesti gasztroszkópiás anyaggal újra elvégeztük a tesztelést.

Betanítás*:*

1: budapesti gasztroszkópiás írott leletanyagok+<különbözeti szólánc>

*2: budapesti gasztroszkópiás írott leletanyagok + 20*hiányzó szólánc+<különbözeti szólánc>*

A 9.1 táblázat az első betanítás tesztelési eredményeit szemlélteti, míg a 9.2 táblázat a második betanítás eredményeit szemlélteti.

9.1 táblázat: Tesztelési eredmények budapesti gasztroszkópiás írott leletek esetén.

betanítás : hiányzó szólánc nem szerepel a betanításban									
Lelet	**Bemondó**	**Össz ref**	**Össz rec**	**Helyes**	**Ins**	**Del**	**Subs**	**Acc**	**WER**
3	BG	95	116	67	21	0	28	48,4	**29,4**
33	BG	63	80	48	17	0	15	49,2	**23,8**
53	SG	55	68	50	13	0	5	67,2	**9,1**
92	SG	55	62	46	7	0	9	70,9	**16,3**
átlagos_WER:		**19,6%**		átlagos Acc:	**58,9%**				

9.2 táblázat: Tesztelési eredmények budapesti gasztroszkópiás írott leletek + 20*hiányzó szólánc esetén. Tesztanyag: 3, 33, 53, 92

betanítás : hiányzó szólánc 20 szor szerepel a betanításban									
Lelet	**Bemondó**	**Össz ref**	**Össz rec**	**Helyes**	**Ins**	**Del**	**Subs**	**Acc**	**WER**
3	BG	95	110	78	17	1	16	64,2	**17,8**
33	BG	63	69	61	6	0	2	87,3	**3,1**
53	SG	55	58	53	3	0	2	90,9	**3,6**
92	SG	55	61	49	6	0	6	78,1	**10,9**
Átlagos_WER:		**8,9%**		átlagos Acc:	**80,1%**				

A táblázatok azt mutatják, hogy a szóláncok 20-szoros megismétlésével a szótévesztés erősen lecsökkent (9.2 táblázat), **a 19,7%-os szótévesztés 8,9 %-ra javult.**

Azonban kérdéses, hogy a tesztelésnél megjelölt többi lelet esetében mi lett az ilyen betanítás mellett az eredmény.

10.1 táblázat: Tesztelési eredmények budapesti gasztroszkópiás írott leletanyagok esetén. Tesztanyag: 94, 38, 174

betanítás : hiányzó szólánc nem szerepel a betanításban									
Lelet	**Bemondó**	**Össz ref**	**Össz rec**	**Helyes**	**Ins**	**Del**	**Subs**	**Acc**	**WER**
94	SG	51	64	41	13	0	10	54,9	**19,6**
38	ZT	34	41	11	11	2	21	0	**67,6**
174	SG	118	167	70	49	0	48	17,7	**40,6**
átlagos_WER:		**42,6%**		átlagos Acc:	**24,2%**				

10.2 táblázat: Tesztelési eredmények budapesti gasztroszkópiás írott leletanyagok + 20*hiányzó szólánc esetén. Tesztanyag: 94, 38, 174

betanítás : hiányzó szólánc 20 szor szerepel a betanításban									
Lelet	**Bemondó**	**Össz ref**	**Össz rec**	**Helyes**	**Ins**	**Del**	**Subs**	**Acc**	**WER**
94	SG	51	64	41	13	0	10	54,9	**19,6**
38	ZT	34	41	11	11	2	21	0	**67,6**
174	SG	118	166	68	48	0	50	16,9	**42,3**
átlagos_WER:		**43,2%**		átlagos Acc:	**23,9%**				

.A 10.1 és 10.2 táblázatokat összehasonlítva láthatjuk, hogy nem változott a felismerés pontossága, tehát itt is beigazolódott az előzetes becslés.

Tehát a perplexitás elemzéssel, és a betanító anyag egyszerű módosításával egy meghatározott szótárkészletű, nyelvi szöveg felismerését jelentős mértékben javítani tudjuk.

6 Kiértékelés

Tehát a példaként bemutatott perplexitás átlag vizsgálata alapján sikerült a szólláncok ismétlési számát optimálisra beállítani, úgy, hogy a felismerés lényegesen jobb lett, 90% fölötti. A vázolt eljárás sok esetben javíthatja a felismerést.

Figyelembe kell azonban venni, hogy a módszerünk nem ad valóságos megoldást, hiszen a gyakorlatban, az erősen agglutináló nyelveknél tisztán statisztikai n-gram modellel dolgozva, mindig lehet új elem az új bemondások között, ami a betanító anyagban nem szerepelt, és ez hibázáshoz vezet. Azonban, egy közepes szótár méretű, kötött témában kialakítandó felismerő létrehozásában jelentős segítség lehet.

Bibliográfia

1. Máté Szarvas, Sadaoki Furui : Evaluation of the Stoachic Morphosyntactic Language Model on a One Million Word Hungarian Dictation Task. EUROSPEECH (2003) GENOVA, 2297-2300

2 Chen S., Beeferman D., Rosenfeld R. : Evaluation Metrics For Language Models, In: DARPA98 , National Institue of Standards and Technology (NIST),
Elérhető: www.nist.gov/speech/publications/darpa98/html/lm30/lm30.htm

3. Clarkson P., Robinson T. : Towards improved language model evaluation measures. Elérhető: http://Citeseer.ist.psu.edu/clarkson99toward.html

4. Deng Y., Mahajan M., Acero A. : Estimating Speech Recaognition Error Rate without Acoustic Test Data. Elérhető: http://research.microsoft.com/srg/papers/2003-milindm-eurospeech.pdf

5 HUMOR Morfológiai elemző. Elérhető: http://www.morphologic.hu/h_humor.htm

6 Becchetti C., Ricotti L. P.: Speech Recognition,Theory and C++ implementation. Fondazione Ugo Bordoni, Rome, (1999). ISBN 0-471-97730-6

7:Ney, H., Essen, U., Kneser, R.: On Structuring Probabilistic Dependencies in Stochastic Language Modeling. Computer Speech and Language, (1994). 8:1-38.

8 Velkei Szabolcs, Vicsi Klára: Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából, MSZNY (2004). 307-315.

[9] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben, II. Magyar Számítógépes Nyelvészet Konferencia, 2004. 315-319 oldal

Folyamatos beszéd szószintű automatikus szegmentálása szupraszegmentális jegyek alapján

Szaszák György[1], Vicsi Klára[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Kutatólaboratórium
{vicsi, szaszak}@tmit.bme.hu
http://alpha.tmit.bme.hu/speech/

Kivonat: Cikkünkben a folyamatos beszéd szupraszegmentális jegyeken alapuló, szószintű szegmentálási lehetőségeit vizsgáljuk statisztikai megközelítésben, rejtett Markov modellek használatával. A szószintű szegmentálás a folyamatos gépi beszédfelismerés robosztusságát növelheti zajos körülmények között, illetve csökkentheti a keresési teret a dekódolás folyamán. Rendszerünk az alapfrekvencia és az energiaszint értékeit veszi figyelembe, az időtartamok pontos mérése ugyanis felismerési feladatban nehezen kivitelezhető. A rendszert kötött hangsúlyú nyelvekre dolgoztuk ki, és a magyar mellett finn nyelvre is adaptáltuk, illetve vizsgáltuk kétnyelvű rendszerek teljesítményét is, amely a működés hatékonyságát növelte. A statisztikai alapú szegmentáló eredményeit összehasonlítottuk korábbi, szabálybázisú eredményeinkkel, a magyar, illetve a finn nyelv szegmentálási lehetőségeit számos paraméter függvényében vizsgáltuk. Megállapíthatjuk, hogy kísérleteink alapján a kötött hangsúlyú nyelvek esetén a beszéd szószintű tagolása megbízhatóan megvalósítható, ami biztató kilátásokat jelent a kidolgozott rendszer beszédfelismerőbe integrálására vonatkozóan.

1 Bevezetés

A prozódia vagy más néven a szupraszegmentális hangszerkezet az emberi beszéd szerves részét képezi, funkciói részben univerzálisak, részben nyelvspecifikusak. Az univerzális funkciók közül kiemelendő a beszéd értelmezésének megkönnyítését célzó szintaktikai tagolás és a modalitás, de ide tartozik a beszélő érzelmeinek, szándékainak kifejezése is [2]. Ezen univerzális funkciók nyelvenkénti konkrét realizációja már többnyire nyelvspecifikus, míg az "eszközök" sokszor univerzálisak: intonáció, hangsúly, szünetek, ritmus, stb. A szupraszegmentális hangszerkezet segítségével valósíthatja meg a beszélő mondandójának kommunikációs szándékának megfelelő strukturálását. Így, ha a beszéd prozódiai tagolása a szintaxis követelményeinek megfelelően alakul, akkor az egyes szakaszokat prozódiai frázisoknak nevezhetjük. Triviális prozódiai frázis például a két levegővétel közötti beszédszakasz. A műszaki gyakorlatban a prozódiai jegyek reprezentálása három, akusztikailag jól mérhető jellemző révén történhet, ezek az alapfrekvencia, az intenzitás és az időtartam. A

szupraszegmentális hangszerkezet egyes elemei – a prozódiai jegyek – lényegében e három akusztikai jellemző különböző időtartományokra érvényes – értsd szó- vagy mondatszintű – kombinációiként is felfoghatók.

1.1 Rövid történeti áttekintés

A prozódiai jegyek felhasználása beszédfelismerési feladatokban a robosztusság növelésére napjainkban ismét reneszánszát éli. A nyolcvanas évek közepének első próbálkozásai [1,7] során a technikai szint még nem volt adott ahhoz, hogy a kapott eredmények alapján azok a beszédfelismerőkbe is beépíthetőek legyenek. Ez a látszólagos kudarc – Philippe Langlais értelmezésében [3] – elsősorban az alábbi három nehézség miatt következett be:
– a prozódiai tudás jelentős mértékű variáltsága (a beszédtípus, a beszélőtől, a tartalom, a környezet, stb. függvényében);
– a szupraszegmentális szinten hordozott információ és az üzenet nyelvi szerveződési szintjei közötti kapcsolatok bonyolutsága;
– és a prozódiai paraméterek mérésének nehézségei, illetve rendszerbe illesztésük a percepció szintjén.

Néhány korábbi munkában [8,10] a kutatók a prozódiai időtartamok mérésével próbálták meg a beszédbeli határokat detektálni, esetenként a rendszert [8] zajos környezetben működő HMM beszédfelismerő front-end moduljaként megvalósítva. Történtek kutatások [4,5] több prozódiai jellemző alapján készített, folyamatos beszédfelismerő kiegészítő moduljaként működő frázisszintű szegmentálóra is.

1.2 Prozódiai jellemzők a magyar és finn nyelvekben

A beszédképzés során az artikulációs szervek folymatosan mozgásban vannak, amely által folytonos akusztikai jelet hoznak létre. Az ember a beszédértemezés során a szintaktikai és a fonológiai szabályok alapján képes a nyelvi egységek, így a prozódiai frázisok tagolására, a mondatok, szavak azonosítására. Kísérletünkben azt vizsgáltuk, lehetséges-e a szóhatárok nagy hatékonyságú detektálása prozódiai jegyek alapján a folymatos magyar és finn beszédben. Ennek során nagyban kihasználtuk, hogy a mind a magyar, mind a vele rokon finn nyelv kötött hangsúlyú [9]. Emiatt mindig a mondatot felépítő szintagmák első szótagjai kapják a hangsúlyt, amelynek detektálásával kellő pontossággal megtalálhatóak a szószerkezetek, sőt – a kötőszavak, névelők és egyéb hangsúlytalan elemek kivételével –az azokat felépítő szavak határai is. Ilyenformán műszaki szemszögből a tényleges prozódiai frázist szűkebben is értelmezhetjük, amely esetenként egészen az egyes szavak határaiig lebontható alkotó elemekre, a nyelv rétegződésének megfelelően – amelyet kísérleteink tanúsága szerint a magyar nyelv sok esetben még a prozódiával is érzékeltet. A fentiek miatt tartottuk célszerűnek a “szóhatár” kifejezés használatát cikkünkben a “prozódiai frázis” kifejezés helyett.

Kísérleteink során a szabály alapú megközelítés esetéhez hasonlóan [9] ismét az alapfrekvencia és az energiaszint feldolgozását tartottuk célszerűnek és kivitelezhetőnek. Az 1. ábra magyar nyelvű példamondatán jól követhető, ahogyan az alapfrekvencia és az energiaszint a szóeleji hangsúlyoknak megfelelően jelentősen emelkedik a szóhatárok után. A szótagok magánhagzóinak hossza nem mutat egyértelmű

szabályszerűséget. Ebben az esetben az alapfrekvenciát, az energiaszintet és az időtartamokat a szótagok magánhangzóinak közepén mértük.

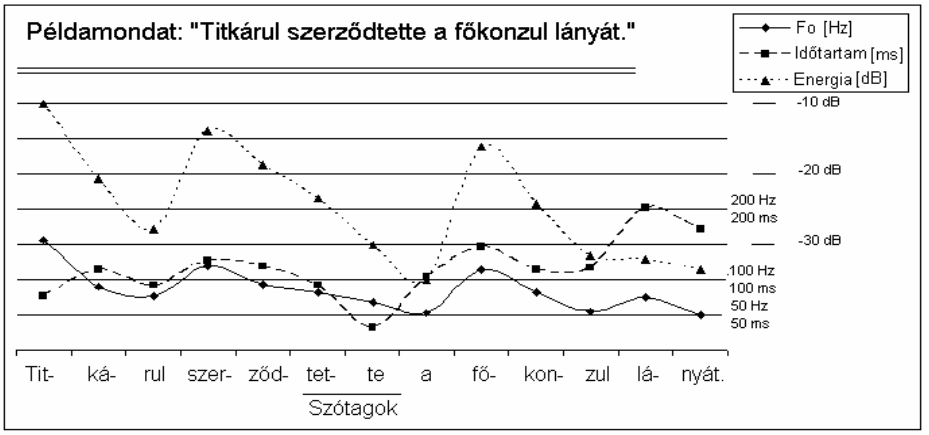


Fig. 1. A magánhagzók közepén mért alapfrekvencia, energiaszint és időtartamok szótagonként a "Titkárul szerződtette a főkonzul lányát" magyar mondatban.

2 Vizsgálati metodika

A szóhatár detektálást a bevezetőben leírtak értelmében prozódiai szegmentálásra vezettük vissza. A korábban elvégzett szabálybázisú vizsgálatok során a hangsúlyt detektáltuk [9], míg a statisztikai alapú automatikus szóhatár meghatározás esetére a HTK [11] rejtett Markov modellek generálására és tesztelésére kialakított fejlesztőkörnyezetet használtuk. Bár a HTK eredetileg beszédfelismerési célokra készült – tehát a fonémák akusztikai HMM modelljeinek elkészítésére és tesztelésére összpontosít –, a benne foglalt HMM implementáció mégis hatékonyan használható más típusú, rejtett Markov modellel leírható osztályozási feladatokra is. Mindezt a laboratóriumunkban kifejlesztett, a HTK programmodul elé illesztett előfeldolgozó egység biztosítja, amely a HTK által értelmezhető formátumúvá konvertálja az adatokat, helyettesítve a beszédfeldolgozás lényegkiemelő modulját.

Vizsgálataink alapjául a BABEL [6], magyar nyelvű beszédadatbázist, és a Helsinki Műszaki Egyetem finn nyelvű beszédadatbázisát (FSD) [12] használtuk fel. Mindkét adatbázis – a későbbiekben részletezendő – prozódiai szintű szegmentálását szakértő végezte. Az adatbázisok magyar nyelvre 22 beszélőtől 1600 mondatot, finn nyelvre 4 beszélőtől 250 mondat tartalmaztak.

2.1 Akusztikai előfeldolgozás

A felhasznált prozódiai jellemzők az alapfrekvencia (Hz) és az energiaszint (dB). Az alapfrekvencia számításakor az autokorrelációs módszert használtuk: az x(n) diszkrét jel autokorrelációs függvénye:

$$R(k) = \sum_{k=N-n-1}^{N} x(n)x(n+k) \quad (11)$$

Az F_0 alapfrekvencia értékét az *i*-edik keretre medián szűrés kapjuk az alábbiak szerint (a keretképzési idő 25,6 ms volt):

$$F_0(i) = \mathbf{med}\ \{ F_0(i-3)\ ,\ F_0(i-2)\ ,\ F_0(i-1)\ ,\ F_0(i)\ ,\ F_0(i+1)\ ,\ F_0(i+2)\ ,\ F_0(i+3) \} \quad (2)$$

Az E(i) energiaszint számítása 100 ms integrálási idővel történt:

$$E(i) = \frac{1}{M+1} \sum_{n=i-\frac{M}{2}}^{i+\frac{M}{2}} x^2(n) \quad (3)$$

ahol M a 100 ms-ra eső minták száma. Az energiaszint értékek keretideje szintén 25,6 ms.

Az így kiszámított alapfrekvencia- és energiaszint-értékeket használjuk a betanításhoz, az első és második deriváltak hozzáfűzése után, illetve a prozódiai alapon működő szó szintű szegmentáló is ilyen bemenő adatokat vár.

2.2 A statisztikai megközelítés

A statisztikai megközelítés során az egyes Markov modellek meghatározott intonációs osztályokra készülnek, amely esetünkben megadja az adott szószerkezet dallamsémáját.

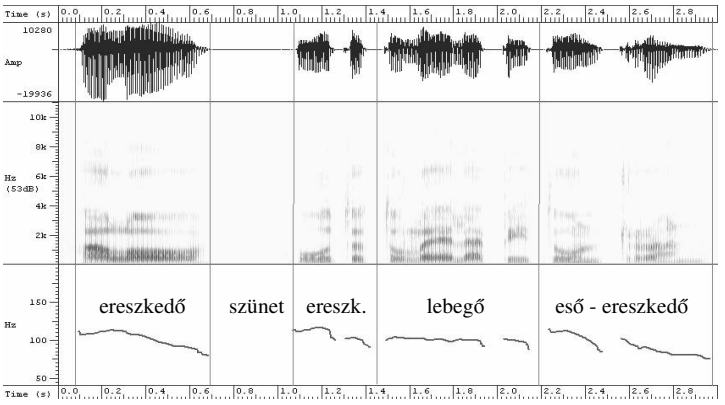


Fig. 2. Néhány tipikus intonációs szegmenstípus magyar nyelvre. Az időfüggvény (felső sáv) és a spekrum (középső sáv) mellett az alapfrekvenciát az alsó sávban láthatjuk.

Kísérleteink során úgy találtuk, hogy célszerű a megkülönböztetett intonációtípusok számát alacsonyan tartani, így csupán az 5, hagyományosnak nevezhető osztályt különböztettünk meg, amelyek az ereszkedő, az emelkedő, az eső, a szökő és a le-

begő. Ehhez hozzávéve a szünetet a kapott 6 féle HMM modellt használtuk vizsgálatainkban. A betanító anyag prozódiai szintű szegmentálásakor ezt a 6 típust jelölték a szakértők ügyelve arra, hogy minden szegmenshatár szóhatárra kerüljön. A 2. ábrán látható példaként néhány tipikus intonációs típus. A szegmentálás során tulajdonképpen az intonációs frázisok határainak bejelölése történt.

A betanítás, illetve az automatikus szószintű szegmentálás blokksémája a 3. ábrán látható. A működés a HMM beszédfelismerő rendszerrel analóg azzal a különbséggel, hogy mások a bemenő adatok, emiatt más az előfeldolgozás, a kimenetből pedig csak a szóhatárok időbeli elhelyezkedése releváns.

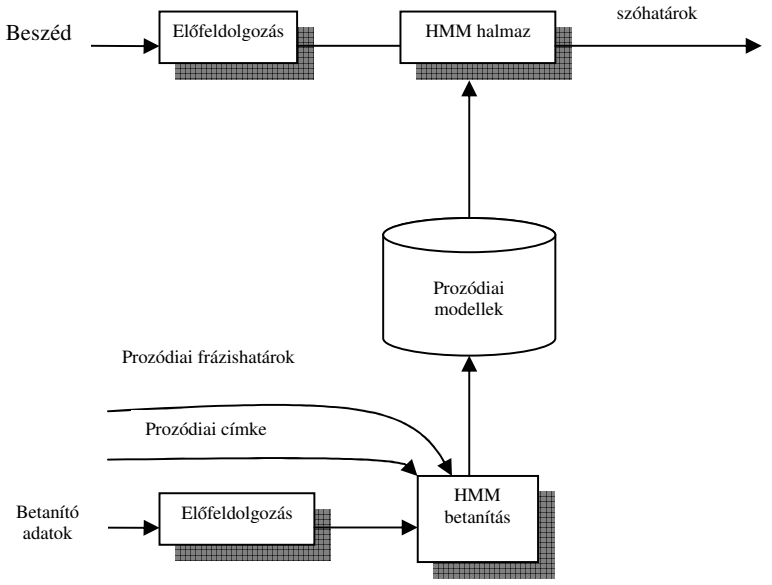


Fig. 3. Az automatikus szószintű szegmentáló vázlatos felépítése, a betanítás és a működés blokksémája

Kiértékelés

Az eredmények kiértékelésénél – a szabálybázisú megközelítéshez hasonlóan [9] – két mutatót használunk. A **pontosság** azt adja meg, hogy a szupraszegmentális jegyek alapján automatikusan detektált szóhatár az esetek hány százalékában valóban szóhatár. Ez a mutató a beszédfelismerésnél elterjedten használt WER mutató inverz megfelelője a prózódia esetére. A másik mutató, a **hatékonyság** pedig az összes szóhatárok közül a megtaláltak arányát adja meg százalékosan.

Nyilvánvaló, hogy a hatékonyságra jóval 100% alatti értékeket fogunk kapni, hiszen az egy hangsúlyozási-hanglejtési szakaszban lévő szókapcsolatokat sokszor nem tudjuk a prozódia alapján elkülöníteni. Szintén nem lesz lehetséges a névelők, a rövidebb kötőszavak, stb. pontos elkülönítése.

A fontosabb mutató a pontosság. Nyilvánvaló, hogy a hibás szóhatár detekció rontaná a csatlakoztatott beszédfelismerő teljesítményét, ezért ennek az értéknek a

maximalizálása kritikus. A szabály alapú megközelítés esetén a pontosság növelése a hatékonyság csökkenését vonta maga után [9], azonban a pontosság értékét ezen az áron is – ésszerű keretek között – maximalizálni kellett.

A szóhatár detekciót akkor tekintettük helyesnek, ha az a valódi szóhatár 100 ms-os környezetébe esett. Az összehasonlítás a betanító anyagban nem szereplő anyag alapján történt, a szószintű szegmentáló kimenetét viszonyítottuk a tényleges szóhatárokhoz.

3. Eredmények

Az eredmények ismertetése mellett szeretnénk kitérni a rendszer főbb paramétereinek optimalizálási lépéseire is, a kapott eredményeket pedig összehasonlítjuk a magyar nyelvre kapott szabálybázisú hangsúly-detektálás esetében kapottakkal, illetve értékeljük a magyar és a finn nyelvre adódott eredményeket.

3.1 A HMM modell struktúrájának optimalizálása

Az alpontban a HMM intonációs modellek két fontos jellemzőjének, az állapotok számának és a kibocsátási valószínűség eloszlást leíró Gauss függvények számának optimalizálási lépését mutatjuk be röviden.

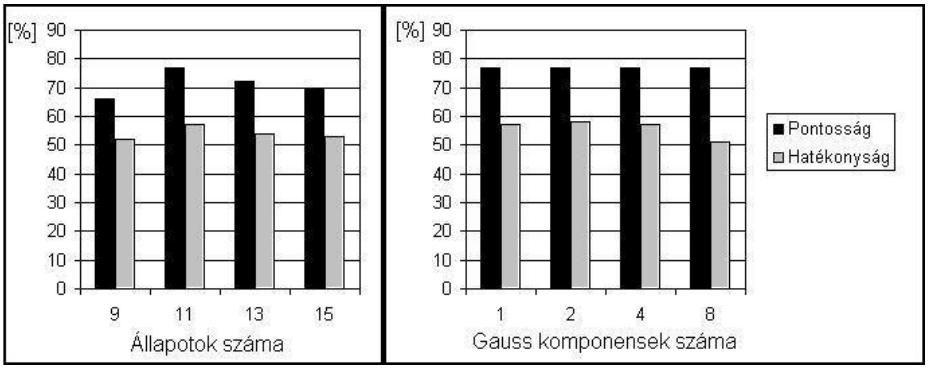


Fig. 4. A pontosság és hatékonyság alakulása az állapotszám függvényében (balra, 4 férfi beszélő, 2 Gauss), illetve a Gauss komponensek számának függvényében (jobbra, 4 férfi, 11 állapot).

A modellek állapotszáma 9 és 15 között változhat, amelyből az első és az utolsó nem kibocsátó állapotok, átmenet mindig csak a következő állapotba lehetséges. Célszerű, ha a modellek legalább 9 állapotúak, hiszen az intonációs frázisok a fonémáknál – amelyeket hagyományosan 5 állapotú modellekkel írnak le 10 ms keretképzési idő mellett – jóval hosszabbak. A 9 állapot kb. 230 ms intonációs frázishossznak felel meg. Esetünkben figyelembe véve a 25,6 ms keretidőt az állapotok számára vonatkozóan ésszerű felső korlát 15 állapot körül van, hiszen minden állápothoz legalább egy keretet hozzá kell rendelnünk. A 15 állapotnak megfelelő intonációs frázis minimálisan szükséges hossza 380 ms, amely tapasztalataink szerint

reális érték. Az optimális állapotszám 11-re adódott (lásd 4. ábra), így a későbbiekben ismertetendő eredményeket is 11 állapotú modellekkel kaptuk. Az egyes állapotokban a kibocsátási eloszlásokat Gauss függvények súlyozott összegével írjuk le [11], a normál függvény komponensek száma esetünkben 1 és 8 között változtatható, tekintettel azonban arra, hogy az alapfrekvencia és az energiaszint menete a beszéd spektrumánál lényegesen egyszerűbb jellemző, elegendőnek bizonyult 1, esetleg 2 Gauss komponens használata (lásd 4. ábra).

3.2 Statisztikai alapú szóhatár detektálás magyar nyelvre

A statisztikai alapú szóhatár meghatározás esetére két betanítási stratégiával kísérleteztünk. Az első esetben vagy csak az alapfrekvencia, vagy csak az energiaszint adataival dolgozott a rendszer, az első és a második deriváltak kiszámítása után csak az egyik prozódiai jellemző (3 elemű jellemzővektor) alapján történt szóhatár detekció. A második esetben mind az alapfrekvencia, mind az energiaszint értékei, első és második deriváltjai alapján történt a betanítás (6 elemű jellemzővektor). Az eredmények a várakozásoknak megfelelően ez utóbbi esetben jobbak, amint azt az 1. táblázatban össze is foglaltuk. A betanítás 14 magyar férfi beszélő anyagával, míg a tesztelés 18 magyar férfi beszélő anyagával történt. A pontosság akkor nagyobb, ha mind az alapfrekvencia, mind az energiaszint típusú értékeket figyelembe vesszük, igaz így a hatékonyság 5-10%-kal csökken.

1. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága magyar nyelvre a bemeneti paraméterek függvényében, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$	magyar	14 férfi	18 férfi	67.4 / 58.4
$E+\Delta E+\Delta^2 E$				67.4 / 63.9
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$				**76.5 / 53.0**

2. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága magyar nyelvre a betanító anyag méretének függvényében, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$	magyar	1 férfi	18 férfi	77.3 / 46.4
		4 férfi		**77.4 / 57.1**
		14 férfi		76.5 / 53.0

Megvizsgáltuk azt is, mekkora betanító adatbázissal lehet a legoptimálisabb eredményt elérni. A betanító anyagot így először 4, majd egyetlen férfi beszélőre szűkítettük, és ugyanazon feltételekkel, ugyanazon 18 férfi beszélő anyagával tesztelést végeztünk. A betanító anyagot ebben az esetben gondosan választottuk ki, különösen ügyelve arra, hogy a betanításhoz használt beszédminták kellően tagoltan, helyes hangsúlyozással beszélő személytől származzanak. Az eredményeket a 2. táblázatban foglaltuk össze. Meglepő, hogy a pontosság gyakorlatilag függetlennek tekinthető a betanító anyagban szereplő beszélők számától, ugyanakkor a hatékonyság már függ ettől, optimálisnak a 4 férfi beszélő anyagával végzett betanítás adódott, ekkor 77,4% pontosságot értünk el 57,1% hatékonyság mellett. Ezek az eredmények felülmúlják a szabálybázisú megközelítéssel kapott értéket, amely esetében 77% pontosság mellett a hatékonyság csupán 23% volt. (vö. [9]). A statisztikai alapú szószintű szegmentáló kimenetére példát a 4. ábrán mutatunk be.

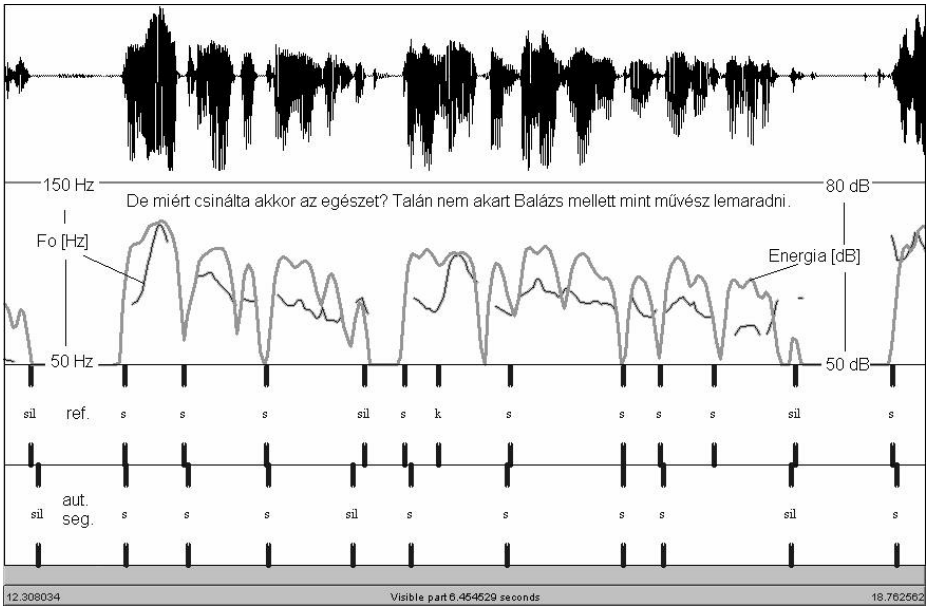


Fig. 4. A prozódiai szegmentálás (3. sáv) és az automatikus szegmentálás (4. sáv) összehasonlítása a „De miért csinálta akkor az egészet? Talán nem akart Balázs mögött mint művész lemaradni.“ szövegrészleten. Az ábra felső részén az időfüggvény (1. sáv), valamint az alapfrekvencia és az energiaszint menete (2. sáv) is látható.

3.3 Statisztikai alapú szóhatár detektálás finn nyelvre

Finn nyelvre a magyar nyelv esetében bemutatott eljárást követve végeztük mind a betanítást, mind a tesztelést. Erre az esetre is a 11 állaptú, 1 Gaussos HMM modellek adták a legjobb eredményt mind az alapfrekvencia, mind az energiaszint, valamint

ezek első és második deriváltjai alapján. A kapott eredményeket a 3. táblázatban foglaltuk össze.

A 3. táblázatból látható, hogy finn nyelv esetén a pontosság alacsonyabb, 69.2%, ugyanakkor a hatékonyság jóval nagyobb, 76.8%, mint a magyar nyelv esetében. Ennek magyarázata az lehet, hogy a kísérleteinkben felhasznált finn beszédet a magyarnál lényegesen lassabb beszédtempó jellemzi, illetve rendkívül gyakoriak a finnben a hosszú, felpattanó zárhangok. Ezeken a helyeken az alapfrekvencia és az energia is hasonlóan viselkedik, mint a szóhatárokon, így az alacsonyabb pontosság abból adódik, hogy a szavak belsejében a hosszú felpattanó zárhangokat is szóhatárként detektálja a rendszer. Mindezt a szegmentáló kimenete is visszaigazolja, hiszen a tévesen detektált szóhatárok finn nyelv eseténben gyakran a hosszan ejtett felpattanó zárhangok zár szakaszára estek. A nagyobb hatékonyság ugyanennek a következménye: a lassúbb beszédtempó miatt a szóhatároknál jobban érzékelhető a szünet, illetve az alapfrekvencia és az energiaszint leesése, így jóval több szóhatárt találunk meg. Véleményünk szerint finn nyelv esetén gyakorlatilag csupán a névelők, kötőszavak előtt, és az egybeolvadásra hajlamos jelzős szerkezetek között nem detektálja a rendszer a szóhatárt, ennek alátámasztása azonban további ellenőrzést igényel.

3. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága finn nyelvre, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel. Az összehasonlításhoz a magyar nyelvű eredményeket is feltüntettük.

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$	finn	4 fő	4 fő	**69.2 / 76.8**
	magyar	4 fő	4 fő	77.3 / 57.1

3.4 Statisztikai alapú szóhatár detektálás kétnyelvű rendszerrel

A módszer más, kötött hangsúlyú nyelvekre való alkalmazhatóságának próbájára magyar anyagon tanított modellekkel finn beszédet szegmentáltunk, illetve ellenkező irányban is végeztünk vizsgálatokat. Megvizsgáltuk továbbá, hogy milyen teljesítményű a mind magyar, mind finn anyaggal vegyesen tanított kétnyelvű rendszer. Az eredményeket a 4. táblázatban mutatjuk be.

A 4. táblázat eredményeiből az tűnik ki, hogy a magyar anyagon tanított, finn nyelvre használt szegmentáló pontossága megegyezik a finn nyelven tanított és finn nyelven tesztelt rendszer pontosságával, a hatékonyság viszont leromlott. A finn anyagon tanított, magyar nyelvre használt szegmentálók esetében a pontosság leromlik, a hatékonyság nem javul. Ezzel szemben a mindkét nyelvű anyaggal vegyesen betanított rendszer pontossága ugyan nem javul az egynyelvű esetekhez képest – magyarra 75%, finnre 69% –, ugyanakkor a hatékonyság jelentősen nagyobb az egynyelvű esethez képest, magyarnál 57% helyett 68%, finn esetében 76% helyett 83%, ami magyar nyelv esetén 19%-os, finn nyelv esetén 9%-os, tehát igen jelentős hatékonyságbeli javulást jelent.

4. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága finn és magyar nyelvre, kétnyelvű rendszerrel.

Prozódiai jellemzők	Betanító anyag	Tesztanyag (4 fő)	Pont. [%] / hat. [%] (11 állapot, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$	magyar (4 fő)	magyar	77 / 57
	magyar (4 fő)	finn	67 / 52
	finn (4 fő)	magyar	70 / 52
	finn (4 fő)	finn	69 / 76
	vegyes (4+4 fő)	magyar	75 / 68
	vegyes (4+4 fő)	finn	69 / 83

4 Összefoglalás

Az alapfrekvencián, és az energiaszinten, mint szupraszegmentális beszédjellemzőkön alapuló automatikus szószintű szegmentálás igen ígéretes eredményeket adott. Ezek alapján statisztikai alapon, az intonációs frázisok rejtett Markov modell segítségével történő leírásával lehetséges a beszédben a szavak határainak megbízható, azaz kellő pontosságú detektálása, jó hatékonysági mutatók mellett. A statisztikai alapú módszer esetén ráadásul kevésbé kényszerülünk kompromisszumot kötni a pontosság és a hatékonyság között, mint a korábbi, szabálybázisú rendszer esetén. Rendszerünket kötött hangsúlyú nyelvekre dolgoztuk ki, és sikerrel adaptáltuk a magyar mellett a finn nyelvre is. A finn és magyar nyelvekre a kétnyelvű rendszer az egynyelvűvel azonos pontosság mellett jóval hatékonyabbnak bizonyult.

A szupraszegmentális beszédjellemzők alapján történő szóhatár detektálás a gépi beszédfelismerők működését javíthatja. Egyrészt hozzájárulhat a felismerés során a keresési tér szűkítéséhez, esetleg lehetőséget adhat a felismerés során futó Viterbi algoritmus szakaszolására. Másrészt zajos körülmények között robosztusabbá teheti a felismerő működését, ez irányban azonban még további vizsgálatok szükségesek. A közeljövőben kísérleteket szeretnénk végezni a szupraszegmentális jegyeken alapuló szóhatár detektálás beszédfelismerő rendszerbe illesztésére vonatkozólag.

Köszönetnyilvánítás

Köszönetünket szeretnénk kifejezni *Péter Attilának*, egyetemünk végzős hallgatójának a kísérletekben való aktív részvételéért, továbbá *Toomas Altosaarnak*, a Helsinki Műszaki Egyetem Akusztikai és Jelfeldolgozási Laboratóriumának vezetőjének hathatós segítségéért, illetve azért, hogy hozzájárult a finn adatbázis használatához.
A kutatást az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretében végeztük.

Bibliográfia

1. Di Cristo: Aspects phonétiques et phonologiques des éléments prosodiques. Modèles linguistiques Tome III (1981) 2:24-83
2. Gósy Mária: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest (2004) 182-243
3. Langlais, P., Méloni, H.: Integration of a prosodic component in an automatic speech recognition system. 3rd European Conference on Speech Communication and Technology. Berlin (1993) 2007-2010.
4. Mandal, S., Datta, A.K. and Gupta, B.: Word boundary Detection of Continuous Speech Signal for Standard Colloquial Bengali (SCB) Using Suprasegmental Features. FRSM
5. Peters, B.: Multiple cues for phonetic phrase boundaries in German spontaneous speech. Proceedings 15th ICPhS, Barcelona (2003) 1795-1798.
6. Roach, P., Vicsi, K. et al.: BABEL: An Eastern European multi-language database. International Conference on Speech and Language Processing, Philadelphia (1996)
7. Rossi, M.: A model for predicting the prosody of spontaneous speech (PPSS model). Speech Communication (1993) 13:87-107.
8. Salomon, A., Espy-Wilson, C.Y., Deshmukh, O.: Detection of speech landmarks. Use of temporal information. Journal of the Acoustical Society of America (2004) 115:1296-1305.
9. Vicsi Klára, Szaszák György, Borostyán Gábor: Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004) 319-326
10. Yang, L.: Duration and pauses as phrase and boundary marking indicators in speech. Proceedings 15th ICPhS, Barcelona (2003) 1791-1794.
11. Young, S. et al.: The HTK Book (for version 3.2). Cambridge University, UK (2002)
12. Vainio, M., Altosaar, T., Karjalainen, M., Aulanko, R., Werner, S.: Neural network models for Finnish prosody. Proceedings of ICPhS 1999, San Francisco (1999) 2347-2350.

Új, zajbecsléssel kombinált, entrópia-alapú beszéddetektálási eljárás a beszédfelismerési hatásfok javítására

Tüske Zoltán, Mihajlik Péter, Tobler Zoltán

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
1117 Budapest, Magyar Tudósok körútja 2,
Hungary
tuske@alpha.tmit.bme.hu
mihajlik@tmit.bme.hu
mgen@freemail.hu

Kivonat: A küszöbszint-alapú beszéddetekció egy új változatát mutatjuk be. Az eljárás energia helyett a robusztusabb spektrális entrópiát használja a beszéd jelenlétének kijelölésére. További különlegessége és újdonsága a megközelítésnek, hogy az entrópiaszámítás előtt minimum spektrális részsáv-energiákon alapuló zajspektrum becslést használ a zaj fehérítésére. Ennek eredményeképp nagymértékben zajtűrő entrópia-alapú beszéddetekciós módszert kaptunk. Ezen állításunkat számos beszédfelismerési kísérlettel támasztjuk alá, melyekben normál és kifejezetten zajos telefonbeszéd-felismerést végeztünk. A javasolt beszéddetekciós eljárás alkalmazásával minden esetben javult a felismerési pontosság (maximálisan rel. 29,5%-kal), míg a felismerendő keretek számát nagyjából az eredeti mennyiség felére szorítva jelentősen csökkent a felismerő terhelése zajban is.

1 Bevezetés

A beszéd-alapú szolgáltatások egyre növekvő száma szükségessé teszi hatékony, zajtűrő beszéddetektorok fejlesztését. A beszéd jelenlétének kijelölése igen fontos például a beszédfelismerőknél és a beszéd telekommunikációs átvitele során.

Előbbi esetben jó beszéddetektálás esetén a felismerő csak a ténylegesen aktív szakaszokat kapja meg, a felismerő kikapcsol, ha a beszélő hallgat. A felismerés pontosabbá válhat, mert ilyenkor a nem-beszédet – amire általában a felismerő nem, vagy csak korlátozott mértékben lehet felkészült – a rendszer nem próbálja a betanított szavak valamelyikéhez hasonlítani, ezáltal a felismerő hatásfoka javul, ráadásul a számításigény is csökken. Tehát egy jó beszéddetektor képes a beszédfelismerő rendszerek pontosságán és működési sebességén javítani.

A második esetben, a beszédátvitel során, a beszéddetektálás fontosságát az adja, hogy sávszélességet spórolhatunk meg, ha a csatornán nem visszük át azokat a szakaszokat amikor a beszélő hallgat.

A távközlésben használt beszéddetektálási algoritmusok azonban nem használhatók közvetlenül a beszédfelismerésben, mert elsősorban nem a beszéd, hanem inkább a csend kijelölése a feladatuk, így nem szűrik ki a beszédfelismerést zavaró zajokat.

Az elmúlt évek során számos detektálási algoritmust dolgoztak ki a beszédfelismerés számára. Ezek az eljárások többé-kevésbé két kategóriába sorolhatók [1]. Az első típusú algoritmus ún. küszöb-alapú [1],[2],[8],[10]. Ebben az esetben a bejövő jelből beszéd/nem-beszéd eldöntésére alkalmas paraméterek kinyerése után adaptív, az idővel változó, a környezethez alkalmazkodni próbáló, vagy globális, előre beállított küszöbérték szerint történik a detektálás.

A küszöb-alapú beszéddetektálás legfontosabb lépései a következők:

- *Paraméter kinyerés*: olyan jellemzők előállítása a jelből, ami mást mutat a zaj- és mást a beszédszakaszokon.
- *Küszöbszint beállítás*: Ez alapján ítélhető meg egy jelszakaszról, hogy azt beszédnek vagy szünetnek tekintsük. Lehet adaptív vagy állandó is.

A másik típusú szegmentálási módszerek mintaillesztéses megközelítést [4] használnak. Ez esetben a beszéd mellett a zajról is szükséges modellt alkotni, és ennek paramétereit megbecsülni. A detektálás hasonlóan történik, mint a felismerési folyamat. A küszöbmódszert alkalmazó detektorokkal összehasonlítva, a mintaillesztésen alapuló eljárások tanító adatokat és nagyobb erőforrásokat igényelnek.

A továbbiakban a küszöb alapján döntő detektorokról lesz szó. Alapvetően egyszerűbbek és gyorsabbak, és jóval szélesebb az alkalmazási körük. Bár a dolgozatban elsősorban a beszédfelismerés hatásfokának javítását célozzuk a zajrezisztens beszéddetekcióval, a lehetséges alkalmazások túlmutatanak a beszédfelismerésen.

2 Energia és entrópia

2.1 Energia-alapú detektorok

Előnyük, hogy a zaj karakterisztikáját nem kell ismerni, viszont érzékenyek a nagy energiájú zajokra, hiszen nem minden beszéd, aminek energiája van, azaz jelentősen csökkenhet a detekció hatékonysága. Alacsony jel-zaj viszony (SNR = Signal to Noise Ratio) esetén pedig a halk beszédszakaszok energiáját teljesen elfedheti a zaj energiája. Tehát az energia-alapú algoritmusok rossz eredményeket mutatnak zajos körülmények között. Az aktuális, T minta hosszú t_0. keretben az energiát a következő módon számoljuk:

$$E_{jel}(t_0) = \sum_{t=t_0 \cdot T}^{t_0 \cdot T+T-1} y^2(t) \qquad (1)$$

A küszöbszint beállítása többféle módon lehetséges. Csúszó ablakos energiaátlagolással, esetleg a t_0-t megelőző rövid időintervallumból a minimális energiaszintet választva. Beszédnek pedig azokat a szakaszokat tekinthetjük, amelyek energiája a küszöb fölé – pl.: min. *6* dB-lel – emelkednek. A fentebb vázolt esetben nincs szükség spektrumszámolásra, aminek számottevő az erőforrás igénye. Bár létezik a spekt-

rum alapján számolt energia-alapú detektálás is, a spektrumból más paraméterek is kinyerhetők, és használhatók az energia mellett illetve helyett.

2.2 Spektrális entrópia-alapú beszéddetektor

E jellemző kiszámolásához szükség van a jel spektrumára. A beérkező jelet átlapolódó blokkokra bontva, és e blokkokon FFT-t (Fast Fourier Transform) végrehajtva kapjuk a jel gördülő spektrumát:

$$Y_{jel}(f,t_0)=\sum_{t=0}^{T-1} y(t_0+t)\cdot h(t)\cdot e^{-\frac{j2\pi\cdot t\cdot f}{T}} \qquad \textbf{(2)}$$

Ahol:

t: a diszkrét idő
y(t): a vizsgált jel
f: frekvencia
t_0: az aktuális keret kezedete
h(t): a súlyozó ablak (általában Hanning)

Amíg a jel-zaj viszony elég magas, addig az energia-alapú detektálás jól használható, de *SNR<0* dB esetén az eredmények elég rosszak, noha a spektrumban még jól látszanak a beszédszakaszok, a spektrum még mutat bizonyos rendezettséget. A spektrum rendezettségének mérésére, az információelméletből ismert Shannon-i entrópia mintájára, [10] bevezeti az amplitúdó spektrum entrópiáját. Ezt a következőképpen definiálja.

Az információ-forrás entrópiája (Shannon) [8]:

$$H(S)=-\sum_{i=1}^{N} P(s_i)\cdot ld\{P(s_i)\} \qquad \textbf{(3)}$$

Ahol s_i a forrásból érkező *i.* szimbólum, $P(s_i)$ az *i.* szimbólum adási valószínűsége. Ezek alapján az *t.* keret *F* frekvencián kiszámolt spektrumának entrópiája [10]:

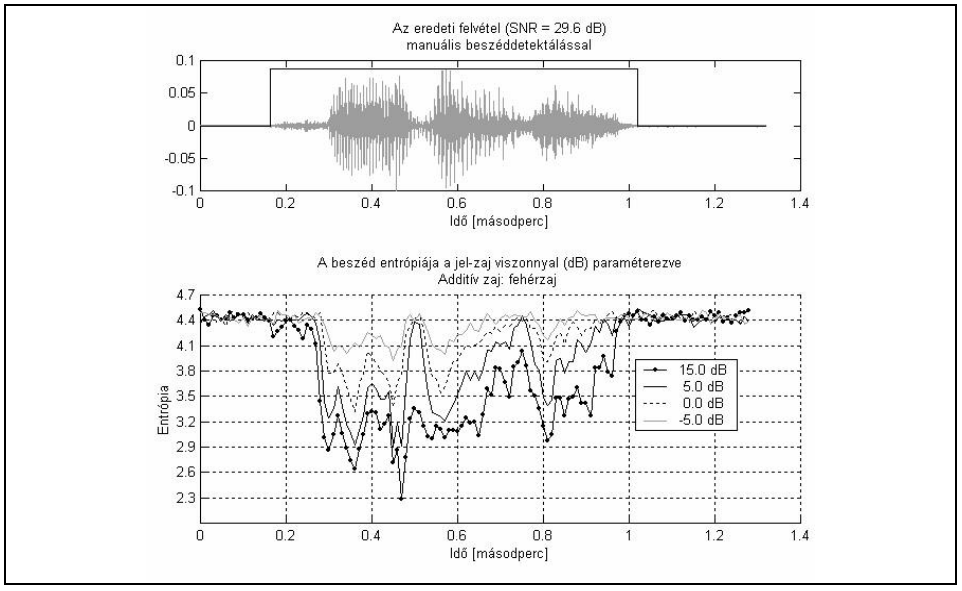
$$H\left(\left|Y_{jel}(f,t)\right|^2\right)=-\sum_{f=1}^{F} P\left(\left|Y_{jel}(f,t)\right|^2\right)\cdot ld\left\{P\left(\left|Y_{jel}(f,t)\right|^2\right)\right\} \qquad \textbf{(4)}$$

Ahol:

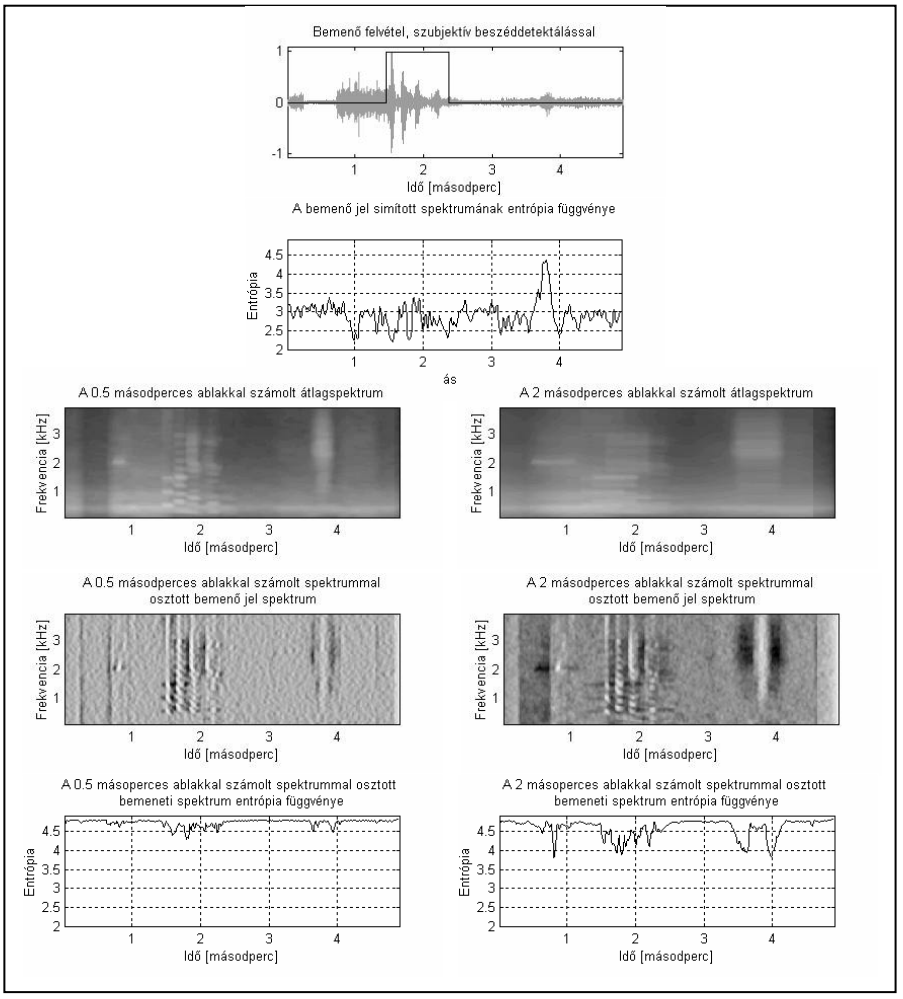
$$P\left(\left|Y_{jel}(f,t)\right|^2\right) = \frac{\left|Y_{jel}(f,t)\right|^2}{\sum_{f=1}^{F}\left|Y_{jel}(f,t)\right|^2} \quad \textbf{(5)}$$

Az entrópia egy véletlen változó bizonytalanságát írja le. Mivel a beszéd és a zaj más-más spektrális karakterisztikával rendelkezik, alkalmas paraméterválasztásnak tűnik a beszéddetektálás döntési kritériumához.

Az entrópia akkor maximális, ha a vizsgált jel fehérzaj, $H_{max} = ld(F)$; minimális, ha a jel tiszta szinusz, $H_{min} = 0$. Fontos, hogy az entrópia értéke a jelszinttől független. Így változó szintű, de állandó spektrális karakterisztikájú zaj esetén a beszéd az entrópiából könnyen kijelölhető. A küszöb meghatározható adaptívan, de létezik statisztikusan becsült megoldás is [10]. Természetesen, ha növeljük a zajszintet, akkor a beszédkeretre számolt entrópia is változik, a zaj spektruma fokozatosan elnyomja a beszédét, a spektrum végül teljesen egyenletessé válik és nem mutat rendezettséget (*1. ábra*).



1. ábra: Beszédjel entrópiájának alakulása növekvő fehérzajban



2. ábra: Az entrópia alakulása átlagspektrummal való osztás hatására

A fent leírt módszer jól használható beszéddetektáláshoz, ha a zaj fehér, azaz a spektruma egyenletes. Színes zaj esetén a zaj spektruma is rendezettebb, ezért nem lesz olyan egyértelmű a beszéd jelenléte az entrópia-idő diagramon.

A [10] az entrópia-alapú detekció egyéb zajokra való kiterjesztéséhez a következőt javasolja. Az aktuális keret spektrumát az entrópia számolása előtt osszunk le a *T* idő alatt számolt átlagoltspektrummal:

$$Y_{átlag}\left(f,t_0\right)=\frac{Y(f,t_0)}{\frac{1}{T}\sum_{t=-T/2}^{T/2}Y(f,t)} \qquad \textbf{(6)}$$

Az így kifehérített spektrumra számoljuk ki az entrópiát kiszámoljuk, és a fehérzajnál alkalmazott detektálási módszer ebben az esetben is használhatóvá válik.

Tapasztalatunk szerint a beszédszakasz spektrumát a körülötte számolt átlagspektrummal osztva lerontjuk a beszéd entrópiáját is. Tehát a zaj spektruma valóban kifehéredik, de a beszéd spektruma is. Így a fehérzajnál alkalmazott detektálási módszer nem lesz elég eredményes színes zaj esetén. (2. *ábra*)

A fenti eljárással az a probléma, hogy az átlagspektrum mindig tartalmazza a beszédspektrumot is, így az azzal való osztás mindig fehérítést jelent a beszédszakasz számára.

Természetesen adódik, hogy ha ismerjük a zaj – legalább közelítő – spektrumát, és a (6) nevezőjében az átlagspektrum helyett alkalmazzuk, akkor csak a zajspektrum fehéredik ki. Meglehet, hogy a beszédspektrum torzul ilyenkor, azonban a rendezettsége megmarad, így az entrópiája is alacsony marad, ugyanakkor a nem-beszéd szakaszok entrópiája közel maximális lesz. Ehhez tehát szükség van a beszéd alatti zaj spektrumának becslésére.

3 Zajbecslés

[7] utal egy olyan fajta zajbecslésre, ami az időben visszatekintve minden frekvenciakomponensnek a minimumát ragadja ki. Az alapgondolat, hogy a beszéd gyorsan ingadozik, szünetekkel tagolt, így megfelelően nagy T időintervallumban a frekvenciakomponensek minimumát kigyűjtve csak a zajra jellemző spektrumot kapunk, ha a zajt lassabban változónak tekintjük a beszédhez képest. A t_0 időponthoz tartozó becsült zaj spektrumát a következő módon kapjuk:

$$Y_{zaj}(f,t_0) = \min_{t=t_0-T\ldots t_0} \{Y_{jel}(f,t)\} \qquad \textbf{(7)}$$

Azonban könnyen belátható, hogy az újonnan belépő zajokkal szemben az eljárás tehetetlen, ezért az általunk javasolt zajbecslés nem csak a múltból, hanem a „jövőből" is vesz mintát a zajspektrum számításához. Természetesen a jövőbeni keretek spektrumának kiszámítása, és felhasználása csak késleltetés árán történhet meg.

A becslés hatásosságának növelésére a becsléshez használt időintervallumot két részre bontottuk, T_1 ill. T_2 hosszú szakaszokra. Mindegyikben külön-külön történt a zajbecslés, azaz két zajbecslővel. Majd a két becsült zajspektrum frekvenciakomponensei közül mindig a nagyobbikat választva került meghatározásra az aktuális keretre vonatkozó zaj spektruma. A becsült zaj t_0 idő pillanatban tehát a következő:

$$\hat{Y}_{zaj}(f,t_0) = MAX\left[\min_{t=t_0-T_1\ldots t_0}\{Y_{jel}(f,t)\}, \min_{t=t_0\ldots t_0+T_2}\{Y_{jel}(f,t)\}\right] \qquad \textbf{(8)}$$

A T_1 és T_2 értékét akkorára érdemes választani, hogy a minimumot kereső ablakban bekövetkezzen beszédhangváltozás, az amplitúdóspektrum átrendeződése. Például egy felpattanó zárhang előtt valószínűleg minden frekvenciakomponens minimumot fog elérni. A múltban működő zajbecsléshez a hosszabb időintervallumot érdemesebb használni, mint a jövő mintáiból való zajbecsléshez, mert ez nem okozhat késleltetést. Viszont a jövőből hosszabb szakaszt venni csak akkor érdemes, ha az algoritmus adatbázison fut, mert valósidejű alkalmazásoknál megengedhetetlenül nagy késleltetést vihetünk be a rendszerbe, ha túl nagy az előretekintés.

4 A javasolt detekciós algoritmus

A bemutatandó beszéddetektor algoritmust NSSE-VAD-nak neveztük (Noise-Suppressed Spectral Entropy-based Voice Activity Detection, [12]), és a következő lépéskből áll. (Lásd még:*3. ábra*)

4.1 Gördülőspektrum-számítás

A bejövő jelet *30* ezredmásodperces keretekre bontva és Hanning ablakot használva, illetve *10* ezredmásodpercenként (a keretek 66.6% átlapolódása) végzett Fourier-transzformálással számoltuk a spektrumot. Az összes beszédminta *fs = 8000* Hz –cel mintavételezett.

4.2 Simítás

Frekvenciában simított spektrumon pontosabban végezhető a zajbecslés, jobban tükrözi a sztohasztikus jelek spektrumát. Például a fehérzaj spektruma ablakozás és Fourier-transzfromálás után nem konstans, míg simítás után jobban közelíti azt. A beszéddetektálást segíti, ha az entrópia görbe gyors időbeli ingadozásait kompenzálandó időben simítjuk a gördülő spektrumot. A két műveletet elvégezéséhez, az amplitudóspektrumot idő-frekvencia síkon egyszerre simítjuk. Ehhez az alábbi 2 dimenziós FIR szűrőt, *S* mátrixot használjuk:

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \frac{1}{35} \qquad \textbf{(9)}$$

$$Y_{simitott}(f_0, t_0) = \sum_{f=-2}^{2} \sum_{t=-2}^{2} Y_{jel}(f_0 + f, t_0 + t) \cdot S(f + 3, t + 3) \qquad \textbf{(10)}$$

Zajbecslés

A zajbecslés a [7] által javasolt elgondolás továbfejlesztett változata (8) alapján történt, hogy a zajbecslő késés nélkül legyen képes követni a hirtelen belépő zajokat. A becsült zaj spektruma a minimum módszerből eredően nem lehet nagyobb egyik frekvencia-komponensen sem, mint az aktuális keret spektruma. A múltbeli zajbecslést a kísérleti tapasztalatok alapján $T_2 = 0.75$ másodpercre, a jövőből becslést pedig $T_1 = 0.25$ másodpercre választottuk.

Zajelnyomás

Az aktuális keret spektrumát (11) alapján fehérítjük. A jelspektrumból azért nem kivonjuk a zajt, mert úgy a maradékspektrum nem lenne fehér, hiszen a becsült zaj csak kisebb lehet, mint a tényleges zaj. Ugyanakkor a becsült zaj spektrumával való osztás után közel konstanssá válik a maradékspektrumban a zaj, ha jó a zajbecslés, és a zaj szerkezetét sikerül megfelelően kinyerni. Tehát az entrópia a maximálishoz közeli lesz a beszédet nem, csak zajt tartalmazó keret esetén.

$$Y_{\text{zajelnyomott}} = \frac{Y_{simitott}}{\hat{Y}_{zaj}} \qquad \textbf{(11)}$$

Spektrális entrópia számítás

Az aktuális, becsült zajjal kifehérített keret spektrális rendezettségét $H(|Y_{\text{zajenyomott}}(f,t)|^2)$ -t a (3), (4) képletek segítségével számoljuk.

Elsőszintű döntés entrópiaküszöb alapján

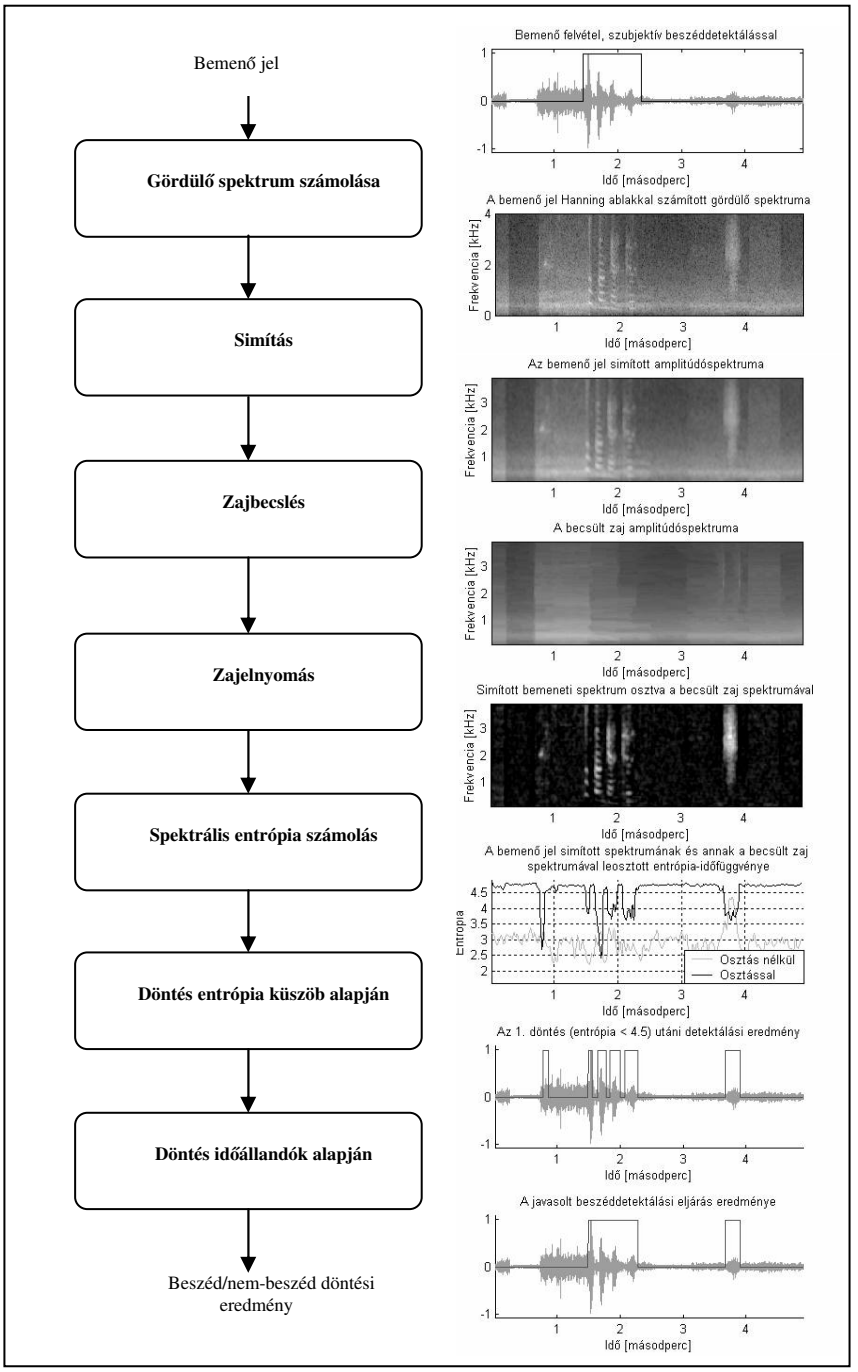
Az entrópia döntési küszöbét 4.5 –nek választottuk. E felett zajnak, alatta beszédnek tekinti a detektor az aktuális keret. Fontos hangsúlyozni, hogy ez a fajta detektálási módszer globális küszöbön alapul. Nincs szükség adaptivitásra, ez a szerep a zajbecslőé. A küszöböt empirikus módszerekkel határoztuk meg.

Második szintű döntés időállandók alapján

A beszédszakasz kijelöléséről az entrópiagörbe küszöb alá kerülésén kívül egy második réteg is dönt a következők szerint.

A beszédszakasz minimális hossza 0.2 másodperc, az ennél rövidebb beszédtartományok nem kerülnek detektálásra.

A beszédben levő szünetek áthidalására a 0.1 másodpercnél kisebb időkülönbséggel rendelkező beszédszakaszok folyamatos szakaszként kerülnek kijelölésre.



3. ábra: A javasolt detektor blokkvázlata és működése

5 Kiértékelés

A szemléltetésnél használt és számos egyéb más beszédmintán végzett kísérletek eredményei jó okot adtak arra, hogy beszédfelismerő rendszerben alkalmazva is megvizsgáljuk a detektor működését, hatását a beszédfelismerésre.

A beszéddetekció hatékonyságát indirekt vizsgáltuk. A tanszéken alkalmazott, nyilvánosan is hozzáférhető beszédadatbázissal [5] betanított beszédfelismerő rendszer felismerési hibaarányait mértük különféle lényegkiemelési beállítások mellett.

5.1 Adatbázisok

Tanításra az MTBA (Magyar nyelvű TelefonBeszéd-Adatbázis) [5] kézzel szegmentált részét használtuk. A teszteléshez két másik telefonbeszéd-adatbázist vettünk igénybe. Elsőként az MTBA-hoz nagyban hasonló Besztel adatbázis „***tiszta***”, vagyis az annotáció során nem zajosként jelölt mintegy 6000 bemondását használtuk. A másik tesztadatbázisunk a nyilvánosan is hozzáférhető Tesztel [6], „***zajos***” telefonbeszéd adatbázis volt. Az ebben levő felvételek szándékosan zajos környezetben (kocsiban, bevásárló központban, utcán, stb.), kifejezetten a zajtűrő beszédfelismerés vizsgálata végett készültek. Itt mintegy 1200 felvételt használtunk a tesztelésnél.

5.2 Vizsgálati módszer

Minden esetben 3 állapotú „balról-jobbra” struktúrájú környezetfüggő rejtett Markov-modelleket használtunk hangmodellekként. Mindkét tesztadatbázison parancsszó felismerést hajtottunk végre a „tiszta” tesztadatbázison 1000 körüli szótárérettel, míg a „zajos” adatbázison 250 körüli szótárérettel a [11] felismerővel.

Az azonos beállítású teszteket mindig párhuzamosan végeztük a két adatbázison. Ezen felül, tekintettel arra, hogy a zajos adatbázis felvételeinek jelentős része AGC (Automatic Gain Control)-torzított, minden beállításnál statikus energiával és anélkül is – az említett hatást kiküszöbölendő – elvégeztük a kísérleteket. Így tehát minden lényegkiemelési módszer esetén négy felismerési tesztet futtattunk. Végül nemcsak a javasolt detektort, hanem az ADSR (Advanced Ditributed Speech Recognition) ETSI szabványban rögzített detekciós eljárást is megvizsgáltuk.

5.3 Lényegkiemelési eljárások

A következő lényegkiemelési konfigurációk mellett végeztünk kísérleteket:

- Alkalmazva az ETSI ADSR lényegkiemelési szabványt, az abban foglalt jelalakformálást, zajelnyomást, vak csatornakiegyenlítést. (ADSR)
- Csak a Mel-frekvenciás kepsztrális együtthatókat számítva. (CC)
- A fenti mellett vak csatornakiegyenlítést is alkalmazva. (CC+BEQ)
- Csatornakiegyenlítést csak a teszteléskor végezve. (CC+fél BEQ)

5.4 Beszédfelismerési eredmények

Először beszéddetekció nélkül mértük az egyes konfigurációk hatásfokát.

1. táblázat: Referencia konfigurációk szó hibaaránya (WER = Word Error Rate) beszéddetektálás nélkül zajos és tiszta adatbázison

Lényegkiemelő	Energiával		Energia nélkül	
	Tiszta	Zajos	Tiszta	Zajos
ADSR	**5,23**	**51,24**	**6,26**	**21,20**
CC	4,78	45,61	5,26	27,33
CC+BEQ	4,76	43,60	5,43	19,97
CC + fél BEQ	4,38	41,87	4,71	20,63

Látható a referenciatáblázatban, hogy a statikus energia elhagyása igen jótékonyan hat a beszédfelismerés hatásfokára zajos esetben. Ez az AGC negatív hatásának kiküszöbölése miatt történhet. Ugyanakkor a tiszta adatokon csökken a hatásfok.

A következő mérési sorozatban a javasolt NSSE-detektor által okozott hatást vizsgáltuk a beszédfelismerés szempontjából, valamint az eredményeket az ADSR saját beszéddetekciós eljárásának erdeményeivel is összevetettük.

2. táblázat: A konfigurációk szó hibaaránya (WER, %) beszéddetektorokkal

Detektor	Lényegki-emelő	Energiával		Energia nélkül	
		Tiszta	Zajos	Tiszta	Zajos
ADSR	ADSR	5,21	51,07	6,26	21,20
NSSE	ADSR	5,11	36,14	5,86	20,54
NSSE	CC	4,66	35,51	5,08	22,77
NSSE	CC + BEQ	4,70	33,83	5,23	18,65
NSSE	CC + fél BEQ	4,27	30,94	4,51	18,48

3. tábázat: A beszéddetektor által okozott relatív százalékos javulás

Detektor	Lényegki-emelő	Energiával			Energia nélkül		
		Tiszta	Zajos	Átlag	Tiszta	Zajos	Átlag
ADSR	ADSR	+0,38	+0,33	**+0,36**	0,00	0,00	**0,00**
NSSE	ADSR	+2,29	+29,47	**+15,88**	+6,39	+3,11	**+4,75**
NSSE	CC	+2,51	+22,14	**+12,33**	+3,42	+16,68	**+10,05**
NSSE	CC + BEQ	+1,26	+22,41	**+11,83**	+3,68	+6,61	**+5,15**
NSSE	CC + fél BEQ	+2,51	+26,10	**+14,31**	+4,25	+10,42	**+7,33**

Látható, hogy a javasolt detekciós algoritmus minden esetben javított a felismerési arányon. Különösen az energiát is tartalmazó zajos eredmények kimagaslóak (maximálisan *29,47%*). Bár a szóhiba-arány eredmények is ígéretesek az NSSE-VAD és az ADSR-VAD összehasonlítást illetően, a két beszéddetektor közti különbség drámaian megnő, ha a „nem-beszéd” keretek eldobási arányait tekintjük (4. ábra).

4. tábázat: A beszéddetektorok által a felismerés során az összes keretből eldobott keretek aránya %-ban

Adatbázis	Detektor	Vektorok száma	Keret dobási arány
Tiszta	ADSR VAD	1.788.101	24,9 %
	NSSE-VAD		60,0 %
Zajos	ADSR VAD	466.332	3,5 %
	NSSE-VAD		52,6 %

7 Összefoglalás

A dolgozat során bemutatott detektálási algoritmus alkalmazásával egyrészt javultak a beszédfelismerési eredmények, másrészt az intenzív kereteldobás következtében jelentősen csökkent a felismerési folyamat erőforrásigénye. Ugyanakkor a zajbecslés az előretekintés miatt *0.25* másodperces késleltetést okoz, ami a valós idejű beszédalkalmazásoknál még megengedhető.

Bibliográfia

1. Abdallah, I., Montrèsor, S., and Baudry, M., “Speech signal detection in noisy environment using a local entropic criterion”, in Eurospeech, Rhodes, Greece, Sep. 1997.
2. Chuan JIA, Bo XU: An Improved Entropy-Based Endpoint Detection Algorithm, ICSLP’02, 2002, Beijing
3. ETSI standard doc., ETSI ES 202 050 v1.1.1.
4. E. Kosmides , E. Dermatas, G. Kokkinakis, “Stochastic endpoint detection in noisy speech”, SPECOM Workshop, 109-114, 1997.
5. http://alpha.ttt.bme.hu/speech/hdbMTBA.php
6. http://alpha.ttt.bme.hu/speech/hdbtesztelen.php
7. Izhak Shafran & Richar Rose: Robust Speech Detection And Segmentation For Real-Time ASR Application
8. Jialin Shen, Jeihweih Hung, Linshan Lee, “Robust entropy based endpoint detection for speech recognition in noisy environments”, International Conference on Spoken Language Processing, Sydney, 1998
9. Péter Mihajlik, Zoltán Tobler, Zoltán Tüske and Géza Gordos; Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, Eurospeech 2005, Lisbon
10. Philippe Renevey and Andrej Drygajlo: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech 2001, Aalborgh
11. T. Fegyó et al. "Voxenter – Intelligent Voice Enabled Call Center for Hungarian", EUROSPEECH, pp. 1905-1908, 2003.
12. Zoltán Tüske, Péter Mihajlik, Zoltán Tobler and Tibor Fegyó; Robust Voice Activity Detection Based on the Entropy of Noisesuppressed Spectrum, Eurospeech 2005, Lisbon

Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához

Tamm Anne, Olaszy Gábor

MTA Nyelvtudományi Intézet, 1068 Budapest, Benczúr u. 33.

Kivonat. Az automatikus szövegelemzés bonyolult kérdésköréből egy részténa vizsgálatát tűztük ki célul, nevezetesen a hangsúly-kategóriák szavankénti kijelölését meghatározott mondatokban. Az eredményeket a gépi beszédszintézis prozódiai támogatáshoz tervezzük felhasználni. A hangsúly-kategóriákat úgynevezett címkékkel jelöljük a szó előtt a szövegben. A célkitűzést két irányból közelítjük: A klasszikus módszernél a címkéket nyelvészeti mondatelemzés eredményéből nyerjük. A másik eljárás lényege nem nyelvészeti központú, hanem egyfajta egyszerű felszíni szövegelemzés, melyben nem használunk nyelvészeti módszereket, csupán szólistákat, táblázatokat, egyszerű szabályokat. Mindkét elemzési formánál alapkövetelmény az algoritmizálhatóság. Az elemzésekhez ugyanazokat a hangsúly-kategóriákat használjuk, így mód nyílik arra, hogy közvetlenül összehasonlíthassuk a nyelvészeti elemzés eredményét a nem nyelvészeti központú eljárásból kapott hangsúly jelölésekkel. Rávilágítunk mindkét elemzésnél, hogy mely problémák miatt nem kaphatunk teljes értékű eredményt sok esetben.

1 Bevezetés

Gépi szövegfelolvasás során a mondatok prozódiája akkor lehet természetes, ha képesek vagyunk egyrészről a mondat dallamának, másrészről a szavak hangsúlyozásának, azaz a prozódiát alkotó két leglényegesebb összetevőnek a szövegbe való jelzés szintű beillesztésére. A mondatok prozódiája egyrészről a szintaktikai szerkezet függvénye, de szemantikai szempontok is meghatározóak lehetnek. Az elmúlt két évtized mondattani, fonológiai és fonetikai eredményei alapján [1], [2], [3], [4], [7] jó eséllyel meg tudjuk jósolni, hogy egy mondatban milyen lesz - vagy lehet - a hangsúlyok eloszlása, és a hangsúlymintázatra milyen intonációs dallam épül. Jelen kutatásban a hangsúlymintázattal, annak is az automatizált megállapításával foglalkozunk. Rámutatunk számos problémára is, ami gátolja a teljes értékű elemzés megvalósítását.

2 Anyag és módszer

A kutatás nyelvi korpuszát médiából gyűjtött híranyag és időjárás jelentés mondatai (500-500 db) képezik. A hangsúlykijelölés alapegysége a mondat. Az algoritmizálás-

ból fakadó kiindulási elv, hogy a mondat minden szavára teszünk hangsúlyjelzést. Ötféle hangsúly kategóriát használunk a szavak besorolására. Ezek jelzései és tartalmuk a következő: [:F]=fókusz, [:E]=kiemelt, [:W]=normál, [:N]=neutrális (hangsúlytalan), [:-] = erősen hangsúlytalan, (esetleg redukált). A fenti kategóriákból három hangsúlycsoport következik: azok a szavak, amelyeken van valamilyen hangsúly (F,E,W jelek), azok, amelyeken nincs hangsúly (N jel), és azok amelyek hangsúlytalan elemeknél is redukáltabban vesznek részt a hangsor felépítésében (negatív jelűek). A hangsúly jelzését a szó elé tesszük az elemzés során. Így a mondat minden szaván lesz jelzés. (A megadott szövegpéldák némelyikében más intonációs jeleket is szerepeltetünk (/26, //11 stb.). Az ilyen jeleket nem kell figyelembe venni, mivel ezek az automatikusan generált prozódiai elemzés nem hangsúlyhoz tartozó jelei.)

Kétféle elemzési elvet vizsgálunk: (A) a szövegek felszíni tanulmányozásából kialakított egyszerű elemző, amelyik nem használ nyelvészeti eszközöket, csak speciális szó- és szókapcsolat-listákat, valamint egyszerű szabályokat; (B) nyelvészeti eszközrendszert felhasználó elemző. A két elemzési módszert összehasonlítottuk olyan formában, hogy vizsgáltuk a szavak elé tett hangsúly jelek helyességét, illetve helytelenségét. Az összehasonlításához egységes alapot teremtettünk percepciós teszttel. A hibákat a zárójelbe tett, helyes jelzéssel érzékeltettük a szöveges anyagban (1), majd összegeztük. A hibákat két fokozatba soroltuk.

Súlyos hiba, ha az elemző a szóra [:E,F] jel tesz, ugyanakkor a szónak [:N], vagy [:-] jelűnek kellene lenni.

Közepes hiba, ha [:W] helyett a szót [:-] jelűként kell jelölni, továbbá, ha [:N] helyett [:W]-t kell szerepeltetni.

(1) */23[:W]reagan /13[:N(**W**)]évek [:N]óta [:pause 1]/24[:E(**N**)]nem [:N]mutatkozott /26[:-]a [:W]nyilvánoss/15ág [:N]előtt[:pause 2800].*

A helyes jelzést a példában kiemelt betűvel jelöltük.

A vizsgálatok eredményeit emberi hangon megszólaló beszédszintetizátorral, hangzó formában is előállítottuk percepciós tesztek végzése céljából.

3. Automatikus hangsúlykijelölés mondatokban

Az automatikus hangsúlykijelölésnél a szöveget vizsgáljuk, és ennek alapján döntjük el az adott szóról, hogy milyen hangsúlyozási fokozatba tartozik. Mindig csak egy mondatot vizsgálunk, mondatok közötti összefüggéseket nem. Mivel nyelvi elemzést végzünk számolnunk kell azzal, hogy lesznek olyan esetek, amelyekben nem tudunk egyértelmű döntést hozni, kompromisszumot kell kötnünk. Erre a következőkben számos példát fogunk látni.

3.1 Hangsúlykijelölés nem nyelvészeti központú megközelítéssel (A)

Ennél a módszernél a szövegek felszíni tanulmányozása alapján alakítunk ki szabályokat, listákat. A szabályok kialakításához felhasználjuk a szövegek felolvasásának hanganyagán végzett fonetikai elemzések (1. ábra) eredményeit is (Olaszy et al. 2001).

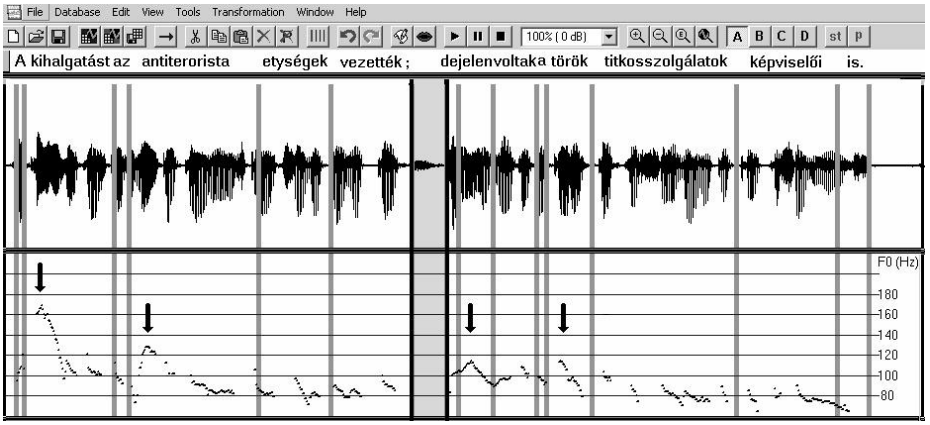


Fig. 1. A hangsúlyos szavak megjelölése a szövegben az alapfrekvencia-görbe (alul) szövegre való visszavetítésével. A nyilak jelzik a hangsúlyokat a szavak első szótagján, a függőleges vonalak a szóhatárokat

Az 1. ábra alapján az elemzett mondatban a szavakra a következő hangsúly-jelöléseket lehet tenni. [:-]A [:W]kihallgatást [:-]az [:W]antiterrorista [:N]egységek [:N]vezették, [:-]de [:W]jelen [:N]voltak [:-]a [:W]török [:N]titkosszolgálatok [:N]képviselői [:N]is.

Hipotézis: A szóhangsúlyozási fokozatok jó hatásfokkal meghatározhatók és jelölhetõk a szövegben szintaktikai elemzésnél egyszerűbb módszerrel, szövegelemek vizsgálatával, és egyszerű szabályok megfogalmazásával is.

A nem nyelvészeti központú hangsúly-meghatározásnál két dolgot kell kiemelni. Az egyik, hogy nem törekszünk teljességre. Elvünk az, hogy a hangsúlyos szavak többségét megtaláljuk a mondatban, továbbá az, hogy lehetőleg ne tegyünk hangsúlyt olyan szavakra, amelyek hangsúlytalanok a kiejtésben. A másik fontos szempont, hogy erősen támaszkodunk a gyakorisági adatokra. Ez azt jelenti, hogy a gyakoribb eseteket vesszük szabálynak, a szabály alóli kivételeket pedig esetlegesen listákban, vagy magában a szabályban adjuk meg. Az eljárásban tehát gyakran kell kompromisszumot kötni.

A szöveg felszíni vizsgálatában a szavakat két kategóriára osztjuk: hangsúly szerinti **tartalmas** szavak, illetve **nem értékes** szavak. Ez utóbbiak azok, amelyekre nem kerülhet hangsúly, vagyis a [:-] jelű szavak (ilyenek például a névelők, a kötőszók). Mindkét kategóriához tartoznak kivételek. A vizsgálatban fontos szerepet tulajdonítunk a mondatban elhelyezett szeparátoroknak (vessző, pontosvessző, kettőspont stb.). A hangsúly-meghatározási eljárás két lépcsős. Az első lépcsőben a megállapított jelöléseket helyezzük el a szavakra, a másodikban az egymás utáni szavakra tett jelzések együttes vizsgálatával (szabályok alapján) az első lépcsőben meghatározott jelöléseket hagyjuk jóvá, illetve változtatjuk meg. Így alakul ki a végleges hangsúly-térkép a mondatban.

3.1.1 Az [:F] jelű kiemelt hangsúlyozású szavak meghatározása

Az [:F] jelű hangsúly a legerősebb a hangsúlyozott szavak között. Ezen szavakat lista alapján jelöljük (például: *nem, ne, nagyon, nincs, soha, semmi, senki, jó, szép, minden, meg kell, mikor, milyen, stb.).* A hangsorban ezek a szavak képviselik a leg-

erősebb hangsúlyt. Fontos szempont, hogy az [:F] jelű szavak után az „irtó" szabályhoz hasonló műveletet hajtunk végre. Ez azt jelenti, hogy ha esetleg a hangsúlykiosztás első fázisában egy [:F] jelű szó utáni szó [:W] jelet kapott, akkor azt törölni kell és [:N]-re kell változtatni. Ugyanilyen szabály vonatkozik az [:F]-jelű szó előtti szóra is. További szabály, hogy [:F] jelzés két egymást követő szón nem lehet. Ha ilyen előfordul, akkor mindig az első tartja meg az [:F] jelzést, a következő [:N]-re íródik át.

3.1.2 Az [:E], illetve a [:W] jelű szavak meghatározása

Az [:E] és [:W] jelzésű szavaknál a hangsúlyozást megvalósító Fo kiemelkedés mértékében van csupán különbség. Az előbbi erősebb hangsúlyt képvisel, mint az utóbbi. A normál szóhangsúlynak a [:W] jelzést tekintjük. Az [:E] jelzést szintén lista alapján osztjuk ki. A [:W] jelzés meghatározásához listát is és szabályokat is alkalmazunk. Ezen szavak kijelölésének a legbonyolultabb a szabályrendszere, ezek empirikus szabályok. Számos olyan tény van, amelyik meghatározza, hogy egy szó az esetek többségében normál hangsúllyal ejtendő. A [:W] szóhangsúly jellel jelöljük meg a listában megadott szavakat, a számok elemeit a *száz, ezer, millió* kivételével, a névelők (*a, az*) utáni szót, a vessző utáni első tartalmas szót, a mondat első tartalmas szavát, a [:-] jelzésű szavak utáni szót, bizonyos szóösszetételek meghatározott szavait (listából), a tulajdon neveket, a személy neveket és a mozaik szavakat (például MTA).

3.1.3 Az [:N], illetve a [:-] jelű szavak meghatározása

Az [:N] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón nem hajtunk végre sem Fo-, sem intenzitás-emelést. A szó a kiejtés szempontjából tehát neutrális (hangsúlytalan), csak a mondatdallamnak engedelmeskedik. Az [:N] jelű szavak kijelölését az ide tartozó lista szerint, valamint a maradék kitöltése elv alapján végezzük. Miután minden jelzést elhelyeztünk a mondatban és a jelzések véglegesítése is megtörtént, akkor az addig nem jelölt szavakra [:N] jelzést teszünk. A [:-] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón csökkentjük az Fo értékét, az intenzitását, valamit a hangok időtartamát. Így egy általános redukciót hajtunk végre a szó hangsorban elfoglalt szerepe szempontjából. A [:-] jelzésű szavakat listából jelöljük ki. Talán ezek a szavak rendelkeznek a legstabilabban a [:-] jelzéssel, mint a hangsúlyozási hierarchia legalsó elemei.

3.2 Vizsgálati eredmények nem nyelvészeti elemzővel

A fenti szabályokkal megvalósított hangsúly-meghatározó algoritmus a Profivox magyar beszédszintetizátorban [6] került megvalósításra a BME Távközlési és Telematikai Tanszékén. A hangsúly-meghatározó szabályok száma mintegy 390 a rendszerben. Az algoritmus döntéseit jelen vizsgálatban egy szűkített korpuszon (50 mondat, összesen 756 szó) vizsgáltuk meg (ugyanezt a korpuszt használtuk a (B) elemző értékelésére is). Minden mondatban elemeztük az algoritmus által a szavakra tett jelzéseket és azok hibás, illetve helyes voltát. Az ítéleteket a mondatok szintetizált formáinak a meghallgatásával végeztük. A mondatokat 3 személy (2 férfi és egy nő, 25, 62, 47 évesek) hallgatta meg és értékelte. A feladatuk az volt, hogy meg kellett hallgatni az adott mondatot, ezzel párhuzamosan tanulmányozhatták a szavakra tett jelzéseket is. Ezután döntötték el, hogy mely jelzések hibásak a mondatban. A hibás

jelzést kijavítva a szövegben a mondatot újra szintetizálták és újbóli meghallgatással ellenőrizték, hogy a rossznak vélt címkékben a javítások a hangzásban is javulást okoztak-e. Példaként bemutatunk egy ilyen tesztelési eredményt. A javított három címkét félkövérrel jelezzük.

Az eredeti mondat:

(2) *A hajnali pára- és köd feloszlását követően ma is sok lesz a napsütés és sokfelé meghaladja a hőmérséklet a 20 fokot.*

A felcímkézett mondat:

(3) */23[:-]a [:W]hajnali [:N**(W)**]pára- [:pause 1]/13[:-]és [:W]köd [:N]feloszlását [:N]követően [:pause 48][:N**(W)**]ma [:pause 1][:-]is [:pause 1]/24[:E**(N)**]sok [:-]lesz /23[:-]a [:W]napsütés [:pause 1]/24[:-]és [:W]sokfelé /23[:W]meghaladja [:pause 1]/24[:-]a [:W]hőmérséklet /26[:-]a [:W]h/15úsz [:N]fokot.*

A szavak száma a mondatban: 21
A hibás jelzések száma:3
A hibák és fajtáik
[:N(W)]pára
[:N(W)]ma
[:E(N)]sok [:-]lesz súlyos hiba

A későbbi (B) eljárás eredményeinek értékelésénél az (A) elemzés percepciós tesztjéből származó jó jelzéseket vettük alapul. Az összesített eredmények szerint a gépi elemzés eredményeit vizsgálva a 756 szóból 97 szón találtak a tesztelők hibás jelzést, ami a teljes szóállomány 12,8 %-a. Az esetek 87,2 %-ában a nem nyelvészeti központú elemző tehát jó hangsúly-kategóriát állapított meg a mondatok szavaira. Ezzel igazolódott a hipotézis. Megvizsgáltuk a hibák összetételét is. Három kategória szerint osztályoztuk: a) amikor a szóra nem tett hangsúlyt a rendszer, noha kellett volna (tipikus esetben az [:N] jelzést [:W]-re kell cserélni); b) amikor a szóra tévedésből hangsúlyt tett, de ez igen zavaró (tipikus eset, amikor a [:W] jelzést [:-]-ra kell változtatni; c) a b) eset enyhébb változata (tipikusan, amikor a [:W] jelzést [:N]-re kell változtatni). Az osztályozás eredménye a következő: a)-ból 77, b)-ből 10 és c)-ből 10 hibát vétett az elemző. A legtöbb esetben tehát a hibás döntés eredménye az volt, hogy a szóra nem tett hangsúlyt az elemző, azt neutrális szintűnek ítélte. Ez összhangban van azzal a korábbi kitétellel, hogy feltételezésünk szerint az a legzavaróbb, ha olyan helyre teszünk hangsúlyt a mondatban, ahová nem kéne. Ilyen hiba mindössze az esetek 2,6%-ában fordult elő.

3.3 A hangsúlyok kijelölése nyelvészeti megközelítéssel (B)

A magyar mondatoknak egy lényegében invariáns hierarchikus szerkezetet [2] tulajdonítunk, s ebből levezethetők a prozódiai szerkezet leglényegesebb komponensei. A mondat topik részre és predikátum részre oszlik. A topik tetszés szerinti (nulla, egy vagy több) igebővítményt és szabad határozót tartalmaz. Bizonyos típusú összetevők (pl. a határozók *szerencsére, valószínűleg, látszólag* típusú mondathatározók) csak a topik részben állhatnak. A topik rész összetevői mind gyenge hangsúlyt viselnek. A

mondat legerősebb hangsúlya és intonációs csúcspontja a predikátumrész első fő összetevőjére esik. A predikátumrész tetszés szerinti és számú (nulla, egy vagy több) disztributív kvantorral (azaz *mindenki, senki, minden előfizető, a posta is* típusú összetevővel) kezdődik. Ezek mindegyike főhangsúlyos. Őket követi a szintén főhangsúlyos, közvetlenül az ige előtti összetevő, mely akár fókusz (***A POSTÁS*** *csengetett be*), akár igekötő (***be-****csengetett*), akár névelőtlen főnév (***levelet*** *hozott*) lehet. Az ezt követő ige hangsúlytalan. Bizonyos mondattípusokban az ige előtti pozíciók üresen maradnak és maga az ige a predikátumrész kezdete: ilyen esetben az ige főhangsúlyos. A főhangsúlyos elemek hangsúlyának erőssége balról jobbra csökken. Az ige utáni fő összetevők attól függően hangsúlyosak, hogy ismert vagy új információt közölnek-e és hogy van-e fókusz a mondatban. Az ige utáni disztributív kvantorok akár hangsúlyosak, akár hangsúlytalanok lehetnek.

A fenti fő elvek alapján egy humán elemző minden lehetséges magyar mondathoz hangsúlyszerkezetet tud rendelni. Ugyanakkor az automatikus elemzést rendkívüli módon megnehezíti, hogy a magyar mondatban lényegében minden mondatpozíció maradhat üresen is, továbbá az igét és az ige előtti pozíciót kivéve a szerkezeti pozíciók több fő összetevővel is kitölthetők. A szerkezeti pozíciók azonosításában tehát nem segít a számolás; irreleváns, hogy egy összetevő hányadik helyen áll. Nem mindig könnyű feladat a fő összetevők határainak automatikus felismerése sem.

A célkitűzésünk megvalósítására a nyelvészeti eszköztárakból a következőket használjuk: morfológiai elemző; NP elemző; fókusz szabályok (azon belül azonosító szabályok és hangsúlytörlő szabályok); egyéb erős hangsúlyt adó szabályok és környezetük; topik szabályok; határozói szabályok ; a szintaktikai egységeken, frázisokon belül működő balszél szabályok; listák (szavak, kifejezések) és egyéb szabályok (például: szöveg- vagy mondat típusától függő szabályok).

3.3.1 Morfológiai elemző (szószablya) - minden szóra megállapítja annak a kategóriáját (pl. hazarendelték = haza[PREF]+rendel[vrb]) és az alaktani alakját (pl. +[PAST INDIC DEF PL 3]). A morfológiai elemző különösen fontos a fókuszos mondatok elemzésében, ahol a ragozott ige és az igerészek helyétől függően a mondatban több frázison is törlődik a hangsúly.

3.3.2 NP elemző (INTEX-alapú, és a Nyelvtudományi Intézetben fejlesztett) – megadja a mondatok határait (a cimkéje: {S} a mondatok elején és a végén). Mondatok (esetleg a mondatrészek) határjelei között keresi a szövegben rejlő NP-ket és ellátja az azonosított elemeket az "NP" címkékkel (4).

(4) {S}*Váratlanul hazarendelték* [np *konzultációra* np] [np *Irakból* np] [np *az ország amerikai polgári kormányzóját* np]. {S}

Ha az NP elemző talál egy főnevet (névszót), akkor megállapítja ennek a közvetlen környezetéhez való viszonyát, és ezután dönti el, hogy az NP-hez tartozik ez a környezet. A főnévi csoport (az NP) határait azért releváns megkeresni, mert a frázisra adott erősebb hangsúly egy főnévi csoportban csak az első „tartalmas" szóra esik. A többi szó az NP-ben semleges hangsúlyt kap vagy azt a hangsúlyjelölést, amelyet a listák alapján előírják neki. Névelőtlen főnevek helye a ragozott igéhez képest viszont fontos az ige utáni frázisok hangsúlyadásban.

3.3.3. Ige kereső - a hangsúlyadás szempontjából fontos az ige helye bizonyos más szavakhoz képest, de azt a tényt is kell megállapítani, hogy van-e ige a mondatban és

ha van, akkor milyen az ige alakja. Tehát az egyik legfontosabb szabályunk az igeszabály. Az igeszabály egy ragozott igét keres, kétfajta kimenetet ad (talált ilyent, illetve nem). Ha a program talál egy ragozott igealakra utaló jelölést (5), akkor ez a mondat egy további igeazonosítás-szabály bemenete lesz (6).

(5) {S}*Váratlanul hazarendelték* **[vrb]** [np *konzultációra* np] [np *Irakból* np] [np *az ország amerikai polgári kormányzóját* np]. {S}

Ha a program nem talál ilyent, akkor a mondat elemzése a névszói állítmányos szabály alkalmazásával folytatódik (6).

(6)
morfológiai elemző → NP-elemző → igekereső → igeazonosítás-szabály
→ névszói állítmányos szabály

3.3.4 Fókusz kereső - az igeszabályokat követő lépésben a fókuszt és a fókusztól függő hangsúlyadást lehet megjelölni. A fókusz lehet egy vagy több szóból álló csoport (egy frázis). A fókusz (F, amit [:F] jellel jelölünk a szó elött) a legerősebb hangsúly a mondatban. A fókusz hat a környezetére is: hangsúlyt irt. Az őt követő ige mindig hangsúlytalan [2]. A fókuszt több ágon lehet meghatározni. A fókuszt kereső szabályok összetettek, egymást követően alkalmazhatók, amíg megtaláljuk a fókuszt. A fókuszt kereső szabályokból jelenleg 5 van. Az első szabály a névelőtlen főnév, igekötő vagy azzal azonos státusú igerész előfordulásnál alkalmazható. Gyakran van egy mondatban egy hátravetett igerész, igekötő vagy névelőtlen főnév. Ha névelőtlen főnév (NP, pl. ***könyvet***), igekötő (***be***) vagy azzal azonos státusú igerész (***haza-***) közvetlenül az ige után helyezkedik el, akkor az a frázis, amelyik az ige előtt helyezkedik el, fókusz (NAGY betűvel jelöltük a példákban) (7). A példában a szabály megtalálja a ragozott, igekötős ige (*berontott*) igekötőjét (a példában: ***be***) az ige után, azért az a frázis, ami az ige előtt van, fókusz (*valamivel 10 óra előtt*). Az [:F] jelzés helyes elhelyezését majd a későbbi balszél szabály fogja kijelölni.

(7) *A gazdagréti bankfiókba VALAMIVEL 10 ÓRA ELŐTT rontott* ***be*** *a símaszkos rabló.*

Gyakran nincs a mondatban hátravetett igerész, igekötő vagy névelőtlen főnév, akkor nem tudjuk helyesen megállapítani az ige előtti és utáni hangsúlyeloszlást. Hosszabb, összetett mondatokban ez viszont lényeges. Olyan esetekben más "kapaszkodókat" használunk. A második szabály: ha a létige bővítménye közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz. A harmadik szabály a negatívan minősítő határozószók esetén találja a fókuszt, a negyedik akkor, ha van egy frázis a kvantorok és az ige között, az ötödik akkor, ha van egy frázis kezdetén a "csak"-szó. Egy példának legyen itt a negatívan minősítő szavak fókuszszabálya. A negatívan minősítő szavak kevésre értékelt számosságú vagy kevésre értékelt mennyiségű dolgot, kis gyakoriságot, kis fokot, mértéket, vagy kevésre értékelt módot jelölnek (vö.. É. Kiss (1998: 48)). Az időjáráskorpuszban gyakran előforduló példák: *rossz, kevés, ritka, kevéssé, ritkán, rosszul*. A szabály szerint ha a negatívan minősítő határozószó, pl. *kevés, ritkán, rosszul*, áll közvetlenül az ige előtt, akkor ez a határozószó fókusz. Ugyanezzel a szabállyal lehet kijelölni a problémás névszói állítmányos

mondatokban vagy mondatrészekben is a fókuszt.(8).

(8) *.... a magas hőmérséklet miatt a lehullott csapadék hamar elolvad, így KEVÉS az esély a fehér karácsonyra.*

Segítség a fókusz megállapításához az is, ha van egy szó vagy szavak csoportja (egy frázis) a kvantorok és az ige között (9), illetve ha a frázis kezdetén a "csak"-szó van, akkor ez a frázis fókusz (10).

(9) *A hajnali pára- és köd feloszlását követően ma is SOK lesz a napsütés...*

(10) *...CSAK ELSZÓRT ZÁPOROK valószínűek.*

Ha semelyik fókuszkereső részprogram nem talált fókuszt, akkor fókuszhangsúlyt adó szabályokra nem kerül sor és tovább lehet lépni a nem-fókuszos főhangsúly, úgynevezett erős hangsúly (E, amit [:E] jellel jelölünk a szó előtt) keresésre. Ha megtaláltuk a fókuszt, akkor lehet tovább lépni a hangsúly-írtó szabályokra.

3.3.5. Erős hangsúly keresése - az erős hangsúlyú frázisok, az „E-elemek" a hangsúly szempontjából a fókuszra hasonlító, de nem teljes fókusz hatáskörrel felruházott frázisok. Az E-elem hangsúlya vagy a fókusz hangsúlya jelöli a predikátumrész kezdetét a mondatban, pl.: *János* ***mindig beteg.*** (a predikátumrészt félköver betűkkel jelöltük). Az E-elem hangsúlya abban hasonlít a fókuszhangsúlyhoz, hogy erős. Két szempontból viszont különbözik a fókusztól, az egyik, hogy hangsúlyirtó hatása nincs, a másik, hogy egy mondatban vagy mondatrészben több E-elem is előfordulhat. A fókuszszabályokkal ellentétben itt fontos az alkalmazási sorrend is. Ha a szabályok nem találtak egyetlen E-elemet se, csak akkor lehet az igét E-elemként címkézni (azaz az ige lesz az E-elem az utolsó azonosítószabály szerint). Az erős hangsúly keresésére is több szabály vonatkozik. Az egyik szabály például azt állapítja meg, hogy ha a névelőtlen főnév (*orvos*), igekötő (*meg-*) vagy azzal azonos státusú igerész (*haza-*) közvetlenül az ige előtt vagy az ige részeként helyezkedik el (11), akkor ez a frázis E-hangsúlyt kap, pl. ***Orvos*** *lett,* ***meg****lett,* ***haza****rendelték,* ***szükség*** *van*.

(11) *[:E]Boát [:-] loptak az állatkertből.*

Egy másik szabály azt állapítja meg, hogy ha egy disztributív kvantor található a mondatban, akkor ez a frázis E-hangsúlyt kap (*János* ***mindig*** *beteg*). Ha egy frázis egy „*is*" szót tartalmaz, akkor ez disztributív kvantor és erős hangsúlyt kap (***János is*** *beteg*). Ha a szó tagadószó, akkor erős hangsúlyt kap, de magába olvasztja a következő szót [2]. A létige ige előtti bővítménye (kiegészítője, pl. ***szerencsés*** *volt,* ***orvos*** *volt*) is erős hangsúlyt kap.

Az erős hangsúlyt adó szabályok után következnek "finomabb" szabályok, topikszabályok, határozószabályok és egyéb szabályok. A konkrét hangsúlyok adása mellett ezeknek a szabályoknak később, a "korrigáló" szabályoknál is fontos szerepük van.

3.3.6. Topik szabályok. Meg kell jelölni a topikrész végét ahhoz, hogy a topikhangsúlyt adó szabályok és néhány egyéb mondatprozódiai szabály működni tudjon. Ezt a feladatot a topikszabályok látják el. A topik szűkebb értelemben egy olyan vonzat, amely a mondatban az ige előtt áll. Az itt alkalmazott "topikrész" alatt

viszont azt a részt értjük a mondatban, amely több frázist is tartalmazhat. A topikrész alatt a predikátumrész előtti részt, technikailag az első [:E] jelű szó előtt vagy a fókusz előtt levő szövegrészt értjük. Tehát, az első [:E] és [:F] jel előtti szövegrész a mondatrészkezdet-jelzésig topikrész, beleértve a határozókat is. A topikrészhez tartozik minden olyan NP és hatarozó, amely az ige előtt áll és nem egy [:E] jelű elem, és nem egy [:F] jelű elem. Ha „topikos" a mondat, akkor a frázis vagy a frázisok topikhangsúlyt kap(nak), vagyis a [:W] jelzés alkalmazható a mondat elején az első tartalmas szón, utána az [:N] a predikátumrész kezdetéig. Több frázist tartalmazó, hosszabb mondatokban, olyan mondatokban, amelyekben van fókusz, de az algoritmusnak nincs egyetlen formai „kapaszkodója" sem" (pl. a hátravetett igekötő vagy a "*csak*"-szó) a fókusz megtalálásához is releváns megjelölni a topikrészt. Például a mondatban nagy valószínűséggel van fókusz akkor, ha a mondat predikátumrész lényegesen „nehezebb" a topikrésznél, azaz több NP-ből és határozófrázisból áll (12).

(12) *Clinton otthon lábadozik.*

Topikrész: *Clinton otthon*, predikátumrész: *lábadozik*. →
Topikrész: *Clinton,* predikátumrész: [:F]*otthon lábadozik.*

3.3.7. Határozó-azonosítók – a főfeladatuk a szövegben rejlő határozók azonosítása, címkézése, hangsúlystruktúrájuk azonosítása és cimkézése. Ehhez 3 külön szemantikai-prozódiai leírással rendelkező listát használunk. Mondat- és módhatározókat különböztetünk meg: pl. *tényleg* mondathatározó, *gyorsan* viszont módhatározó. Ha a határozószó mondathatározó (*esetleg, állítólag*, *okvetlenül, feltétlenül, tényleg*), akkor a mondat topikrészében van. Ez azt jelenti, hogy a lista alapján már el lehetne dönteni, hogy a határozószó a mondat kezdeti topikrészhez – nem a predikátumrészhez – tartozó szavak csoportjai között van és semleges topikhangsúlyt kap vagy nem. Mondathatározó listákból kettő van: hangsúlyosak és hangsúly nélküliek. Ha egy hangsúly nélküli mondathatározó a mondat első szava, akkor – annak ellenére is, hogy ez a pozíció általában hangsúlyos és egy [:W] jelzést kap – a hangsúly nélküli mondathatározó nem kap a topik elején megjelenő hangsúlyt. Tehát egy mondat elején megjelenő mondathatározó nem lesz mindig [:W] , hanem [:N] (13).

(13) *[:N]Állítólag [:W] kínvallatás alkalmazását is engedélyezte az amerikai védelmi miniszter a guantánamói amerikai támaszponton fogva tartott feltételezett terroristákkal szemben.*

3.3.8. Az egyéb szabályok összetettek, legtöbbjüknek lokális, lista-alapú jellege van. Néhány példa: ha a szó kötőszó (*hogy, ami, de, hanem*), akkor törlődik a hangsúly ([:N] lesz), az *egy* szó hangsúlyos, ha utána mértékegység jön (egy *fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, kilométer, milliméter, centiméter, méter*-sorozat*, láb, mérföld, gramm-sorozat, deka,* stb.). A mértékegységeken viszont törlődik a hangsúly ([:N] lesz). Ha van kis fokozatot jelölő összetevő [7] szerint:, *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi, némileg*), akkor törlődik a hangsúly ([:N] lesz). A szemantikailag kiüresedett bővítmények esetében [7] törlődik a hangsúly [:N] lesz, pl. *bizonyos, valóságos, szegény, kis*. Ha címek és rangok vannak a tulajdonnevek előtt, akkor törlődik a hangsúly, [:N] lesz: *úr, néni, bácsi, út, köz, utca,*

doktor stb. Páros kötőszók esetén alkalmazható a [:W] jelzés (*nem.. hanem, akár... akár, vagy .... vagy, mind ... mind*). Ha a frázisban található egy listázott hangsúlykerülő, akkor [:-] jelzésű hangsúlyt kell alkalmazni. A hangsúlykerülők pl. *akar, érint, fog, folyik, talál, kell, szabad, szeretnék,* stb.

A szövegtípusból adódó hangsúlyszabályok közé tartozik például, hogy nagyobb hangsúlyt, [:W]-t kap az ige, ha a predikátumrész rövidebb a topik résznél, azaz kevesebb NP-ből és határozófrázisból áll.

Egy vessző utáni „mondásige" (*mondta, döntött,* stb.) a következő hangsúlymintát kapja: az ige és az igemódosító [:-] jelet, a többi frázisba egy [:W] - [:N] típusú hangsúlyminta kerül (a hangsúlyt a frázis első tartalmas szava kapja a balszél-szabály szerint), a többi rész [:N jelet kap. Ez akkor is érvényes, ha egy hátravetett igemódosító és főnévi csoportok vagy határozók állnak mögötte. Ha nincs fókusz vagy E-elem a mondatban, akkor is a frázisokon balszél-szabály szerinti fenti hangsúlymintázat lesz a topik, a határozók és az erős hangsúlyú ige után.

Egyéb szabályok alkalmazásánál gyakran számít a sorrendjük. A balszélszabályt az elemzés legvégén alkalmazzuk. A frázisra adott erősebb hangsúly egy lineáris szavak csoportjában csak az első „tartalmas" szóra esik. A többi szó a frázisban semleges hangsúlyt kap vagy azt a hangsúlyjelölést, amelyet a listák alapján előírják neki. Ezek a szabályok sok esetben korrigálhatnak a lista alapú megközelítéssel kapott eredményt.

3.4 Vizsgálati eredmények a nyelvészeti elemzővel

A nyelvészeti központú hangsúlykijelölő elemzés jelöléseit humán erővel, az algoritmus figyelembevételével helyeztük el a vizsgált mondatokban. Itt is ugyanazt az 50 mondatot használtuk, amelyeket az (A) elemző értékelésénél (lásd 3.2 pont). Az értékelés során megvizsgáltuk, hogy a nyelvészeti elemzőből adódó jelzésekben hol van hiba (ez szintén humán elemzéssel történt). Az összesített eredmények szerint a nyelvészeti elemzés során a 756 szóból 90 szón találtunk hibás jelzést, ami a teljes szóállomány 11,9 %-a. A nyelvészeti központú elemző tehát közel annyiszor vétett, mint a gépi elemző. Itt is megvizsgáltuk a hibák összetételét is. Az a)-ból 66, b)-ből 5 és c)-ből 19 esetben vétett a nyelvészeti elemző. A legtöbb esetben tehát a hibás döntés eredménye itt is az volt, hogy a szóra nem tett hangsúlyt az elemző, azt neutrális szintűnek ítélte.

4 Összefoglalás

A kapott számszerűsített eredmények azt mutatják, hogy minkét elemző hasonlóan jó hatásfokkal végzi az elemzést. Az eredmények hasonlóságánál látnunk kell, hogy a számok mögötti hibák típusai különbözhetnek a két eljárásban. Ezért külön elemezzük az eredményeket. Az (A) eljárásnál tapasztalt hibák legtöbbje abból adódik, hogy nincs szisztematikus fókusz keresés, ezért az irtó szabály hatóköre sok esetben szűkebb, mint kellene, továbbá az ige meghatározás hiánya is sokszor hibát okoz. Az NP elemző és a balszél szabály nélkülözése túl sok felesleges [:W] hangsúlyhoz vezethet. A (B) típusú elemző hibái abból adódnak, hogy a jelenlegi szabályok túl általánosak, az NP elemzést finomítani kell. Ha az NP elemző rosszul azonosítja az NP határokat, akkor a többi, erre épülő elemzés hibás eredményt ad. Az elemzés hatékonyságának növeléséhez ezen kívül bővíteni kell a hangsúlykerűlő elemek szótárát. Be kell építeni továbbá mondatrészhatár azonosítót is a rendszerbe. A jelenlegi tapasztalatok alapján tehát úgy látjuk, hogy a meghallgatásos vizsgálatokra alapozva, valós elhangzó minták vizsgálatának támogatásával finomítani és bővíteni lehet az elméleti nyelvészeti kutatásokat is.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki az MTA Nyelvtudományi Intézet munkatársainak Gábor Katának és É. Kiss Katalinnak az elemzésekhez nyújtott segítségükért, továbbá a BME Távközlési és Médiainformatikai Tanszékről Kiss Gézának, aki az (A) elemző algoritmusából működő programot készített.

Ezt a kutatást az NKFP 2. programja (2/034/2004 sz.) támogatta.

Hivatkozások

1. É. Kiss, K.: The Syntax of Hungarian. Cambridge Syntax Guides. Cambridge: Cambridge University Press (2002)
2. É. Kiss, K., Kiefer, F., Siptár, P.: Új magyar nyelvtan. Budapest: Osiris (1998)
3. Hunyadi, L.: Hungarian sentence prosody and Universal Grammar. New York, Peter Lang (2002)
4. Olaszy, G.: The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4) (2002) 277-306
5. Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, Cs., Gordos, G.: Profivox – a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4. Kluwer Academic Publishers (2000) 201-215
6. Olaszy, G., Németh, G., Kiss, G.: Hungarian audiovisual prosody composer and TTS development tool. In: Prosody 2000. Editors: Puppel Stanislaw, Grazina Demenko. Poznan (2001) 167-178
7. Varga, L.: Intonation and Stress: Evidence from Hungarian. New York: Palgrave Macmillan (2002)

Beszéd a szavakon túl

Ruttkay Zsófia

PPKE ITK, Budapest
University of Twente, Enschede, The Netherlands,
z.m.ruttkay@cs.utwente.nl

Kivonat: Az emberi kommunikációban a beszélt nyelven túl nagy jelentősége van az azt kísérő nemverbális jeleknek: arckifejezéseknek, gesztusoknak, a tekintetnek, melyeket mindenképpen figyelembe kell venni virtuális emberek kommunikációjának tervezésekor is. A cikkben áttekintést adunk a nemverbális jelek jelentéséről, fajtáiról, morfológiai jellemzőiről. Az arc mimika lehetséges hatásait két kísérlet eredményeivel illusztráljuk. Majd bemutatjuk a GESTYLE szöveg annotáló nyelvet, melynek segítségével virtuális emberek beszédét kísérő nemverbális jeleket lehet több szinten, nemdeterminisztikus módon definiálni.

1. Bevezetés

Az emberi kommunikáció legfontosabb formája a beszélt nyelv. A kimondott, szó szerinti tartalmat modulálja, vagy akár meg is változtatja az, hogy miként hangzik el a mondat: gúnyosan, kétkedve, meglepett hangon. A beszéd jellegzetességei mellett hasonló szerepük van a nemverbális csatornákon közvetített jeleknek: a beszélő arckifejezésének, gesztusainak, tekintetének, sőt testhelyzetének. A nemverbális jelek alapvető velejárói a beszédnek: a beszélő még akkor is gesztikulál és használja mimikáját, ha a hallgató nem láthatja azokat, például. telefonos beszélgetéskor. Továbbá időnként kizárólag nemverbális jelekkel kommunikál az ember. Például a hallgató bólogatással jelzi, hogy követi a beszélő által mondottakat. A nemverbális jelek és a beszéd viszonyát tekintve ma is vitatott, hogy a beszéd és nemverbális jelek egy tőről és időben keletkeznek-e, és együtt, egyenrangú módon közvetítik a kifejezendő tartalmat, vagy a nemverbális kommunikációt alárendelt szerepet tölt be oly módon, hogy a gesztusok segítik a beszélőt a beszéd produkálásában, kiegészítik, modulálják a beszédben közölt tartalmat [8].

A továbbiakban *gesztus* alatt a test egy vagy több részének olyan, többé-kevésbé meghatározott, összehangolt mozgását értjük, amely egy közösségben jelentéssel bír, használati kontextusa körülhatárolt. E definíció egyes aspektusait majd még részletesen szemügyre vesszük, egyelőre néhány példa: felhúzott szemöldökök és tágra nyílt szemek csodálkozást, illetve kérdést fejeznek ki, vagy kiemelik a beszédben elhangzottat. A jobb kéz mutató és középső ujja kb. a váll magasságban, V alakban tartva győzelmet, helyeslést fejez ki.

Minket az emberi gesztusok egy speciális szempontból érdekelnek. Nevezetesen, virtuális embereket készítve, arra vagyunk kíváncsiak, hogy őket az emberi gesztusrepertoár mely elemeivel, és milyen módon ruházzuk fel. Egy *virtuális ember* (VE, angol megfelelői: virtual humans, embodied conversational agents) olyan, megjelenésében az emberre hasonlító számítógépes modell, mely a hétköznapokban megszokott, természetes, emberi módon képes kommunikálni [1]. Napjainkban az egyre jobb minőségű szintetikus beszéd mellett nagy figyelmet szentelnek a nemverbális modalitások használatának [2, 17]. Noha kérdésfelvetésünk alapvetően pragmatikusnak tűnik, az emberi nemverbális kommunikáció durván szólva ilyetén reprodukálása alapvető elméleti kérdéseket is felvet, különböző tudományágak terén:

1. **A nemverbális kommunikáció jelenségei** Mik a gesztusok szerepe a mindennapi kommunikációban? A beszélgető felek, illetve a környezet egyes jellemzői miként befolyásolják a nemverbális jelek használatát?
2. **Gesztusok szerepe ember-virtuális ember párbeszéde esetén** A funkciók és befolyásoló tényezők azonosak-e akkor is, ha az egyik beszélgető fél egy virtuális ember? Fontos-e, hogy egy virtuális ember nemverbálisan is tudjon kommunikálni? Egy-egy alkalmazási szerepkörben milyen gesztusokra kell hogy képes legyen a virtuális ember?
3. **Gesztusok számítógépes modellezése** Ha a fenti kérdéseket tisztáztuk, a számítógépes megvalósítás újabb, immár a gesztus formálás és mozgás részleteinek modellezésre vonatkozó morfológiai problémákat vet fel: hogyan jellemezhető egy gesztus, mik az időbeli változások jellemzői? Mennyire kötöttek ezek a jellemzők, hogyan tehetők a gesztusok egyedivé? Mik a beszéd és gesztusok szinkronjának elvei?
4. **Gesztushasználat számítógépes vezérlése** A virtuális embert a fenti elvek alapján akarjuk vezérelni, lehetőleg magas szinten és a gesztusokat automatikusan előállítva. Milyen reprezentációt és számítógépes vezérlési mechanizmust használjunk?
5. **Gesztusok számítógépes megjelenítése** Interaktív alkalmazások esetében elengedhetetlen, hogy elfogadható válaszidőn belül előálljon a verbális és nemverbális reakció, és a virtuális ember gesztusai simán és megfelelően időzítve jelenjenek meg a képernyőn.

A fenti kérdések megválaszolásához a társadalomtudomány – pszichológia, nyelvészet, szociális antropológia – , a számítástudomány – mesterséges intelligencia, számítógépes nyelvészet, grafika és animáció, gépi látás –, sőt időnként a művészetek – animációs film, festészet, színjátszás – művelőinek összefogása szükséges. Másrészt a számítógépes modellek, és a minden részletében kontrollálható virtuális ember kísérleti médiumként is szolgál, például pszichológusok, pszichiáterek számára ahhoz, hogy részletekbe menően feltérképezzék az emberi nemverbális kommunikáció jelenségeit.

A cikkünkben a fenti kérdéscsoportokat vesszük sorra, az utolsó, számítógépes grafikait kivéve. A 2. fejezetben először az emberi gesztikuláció jelenségeit tekintjük át. Majd arra keressük a választ, hogy mi a hatása egy virtuális ember az igazi emberéhez többé vagy kevésbé hasonló gesztikulálásának. Saját kutatásunkból idézzük a szemöldökmozgás lényegkiemelő funkciójára és a tekintet mint személyiségjegyre vonatkozó, virtuális emberrel végzett kísérleteink eredményeit. A 3. fejezetben egyetlen arckifejezés, a mosoly esetében mutatjuk be a modellezés megannyi elvi kérdését. A 4. fejezetben az általunk kifejlesztett GESTYLE nyelvre kerül sor, mely lehetővé

teszi, hogy gesztusok morfológiáját és egy virtuális ember által gesztikulációs szokásait deklaráljuk, és ennek alapján az elmondandó szövegbe szúrt, jelentést meghatározó címkéket automatikusan gesztusokra lefordítva, egyedileg gesztikuláló virtuális lényt hozzunk létre.

Végül a záró fejezetben a sok nyitott kérdés közül olyanokat sorolunk fel, melyek a számítógépes nyelvészet, illetve a magyar nemverbális kommunikáció szempontjából érdekesek.

2. A nemverbális kommunikáció sajátosságai

2.1 Az emberi gesztusok jellemzői

Az emberi gesztusokat különböző szempontok szerint vizsgálhatjuk [4, 7, 14]. A legfontosabb szempont a *funkció* vagy *jelentés*: mit fejez ki egy gesztus? Egy gesztus *biológia szükségletet teljesíthet* (például a pislogás alapvetően a szemgolyó nedvesen tartását szolgálja), *tagolhatja a beszédet* (például felsorolás jelzése számokat mutató gesztusokkal, vagy ellentét jelzése a fej jobbra illetve balra döntésével vagy a jobb és bal kéz kinyitó mozdulatával), *szabályozhatja a diskurzus menetét* (például ha a szót át kívánjuk adni a partnerünknek, ezt ráemelt tekintettel tesszük, különben csak röviden pillantunk a hallgatóra, hogy időről időre ellenőrizzük, ért-e bennünket). Egy-egy gesztus, illetve a gesztikulálás módja jelezheti a beszélő *érzelmi, szellemi vagy fizikai állapotát*. Például az elégedettség mosolyban nyilvánul meg, míg a szomorúság a lefelé görbülő száj, mint arckifejezés mellett a kéz és testmozgások lassúbb és ritkább voltában is, de ez utóbbiak a fizikai fáradtság jelei is lehetnek. Egy emlék felidézését, illetve a mondandó megformálásának keresését ferdén felfelé emelt tekintet kíséri. Egy gesztus jelezheti a beszédben említett *tárgy vagy esemény bizonyos jegyeit*, mint méret, alak, hely, időbeliség, ismétlődés. Például egy hatalmas farkasról beszélve, a mesélő szeme kitágul, szemöldökét felemeli, és két kezével is jelezheti a farkas méretét. Egyes gesztusok teljes, *absztrakt fogalmakat jelölnek* (lásd a korábban említett győzelem jele). Mások *modulálják az elmondottakat,* annak bizonyosságot-bizonytalanságot, illetve különböző imperatív jelleget adva. Például egy alku során az „ezer forint" puhatolózó, kérő vagy megmásíthatatlan tény jellege nemcsak intonációban, hanem kérdő, kérő vagy közlő arckifejezésben is megnyilvánul.

A *beszéd és gesztus viszonyát tekintve*, bizonyos gesztusok *beszédtől függetlenül*, önmagukban is használhatók. A legtöbb gesztus azonban beszéddel együtt használatos. Ezek vagy redundáns módon *megerősítik, vagy kiegészítik* a beszédben foglalt információt. Speciálisan, a rámutatásnak a beszédben nem szereplő referenciák megadásában, illetve többértelmű referenciák feloldásában lehet szerepe. Például a „Nyisd ki az ablakot!" felszólításban a megszólítottat és a kinyitandó ablakot is rámutatás jelezheti a beszélő, ha több személy és ablak is szóba jön. A rámutatás nemcsak kézzel, hanem tekintettel, sőt fejbiccentéssel is történhet.

Egy másik osztályozási szempont az, hogy egy gesztus *milyen körben ismert,* illetve milyen helyzetben használatos. Vannak csak egy-egy etnikai, foglalkozási illetve társadalmi körön belül használatos gesztusok. Görögországban a tagadást fejezi ki a nyugat-európai igenlő bólintás. Egy másik példa az üdvözlés gesztusai: a partnerek

neme, kora, társadalmi helyzete, ismeretsége és nem utolsó sorban etnikai hovatartozása határozza meg, hogy például a tekintettel történő biccentés és (adott számú és sorrendű) arcon csókolás közti lehetőségek közül melyiket használják.

A gesztusok *morfológiájukat tekintve unimodálisak* vagy *multimodálisak*, *egyszeriek* vagy*ismétlődők, statikusak illetve dinamikusak.* Statikus gesztus például a V jel mutatása, míg dinamikus a hangsúlyozó leütő kézmozdulat. Egy gesztust az arc egyes jegyeinek, illetve a kéznek a koordinált mozgása jellemez. Egy-egy gesztus többféle változatban is használatos, mely változatokat lényegében azonosként érzékelünk és értelmezünk. Pl. a V jel mutatása esetén a kézfej helye nem pontosan meghatározott, a kézforma a karakterisztikus jellemző, melyet látható magasságban kell felmutatni. Viszont a hangsúlyozásra a fejbiccentés és kézleütés két, modalitásában és morfológiájában különböző gesztus. Másrészt egyetlen gesztus többféle jelentést is kifejezhet, lásd a már említett felemelt szemöldök. Tehát a gesztusok és jelentések között *több-többértelmű a megfeleltetés*.

Ezzel eljutottunk a két utolsó, a gesztikuláló személyéhez kapcsolódó szempontig. Az, hogy valaki a morfológiailag és mozgáskarakterisztikájában különböző alternatívák közül milyen gesztusokat használ, *személyiségétől is függ*. Egy nyitott, magabiztos személy nagyobb gesztikulációs teret tölt ki maga körül, több és nagyobb intenzitású gesztust használ, mint egy zárkózott ember. Továbbá a gesztikulálásban, beleértve az arc mimikáját is, mindig van több-kevesebb *egyéni jellegzetesség*.

2.2 Hogyan értelmezzük egy virtuális ember gesztusait?

Szükséges-e virtuális embereket a valódiak körében megismert gazdagságban és részletességben gesztusokkal felruháznunk? Nem elegendő-e csak a szigorúan funkcionális, referenciákat feloldó vagy információt hordozó gesztusokra szorítkoznunk, és azokat egyetlen prototípus formájában használni? Az utóbbi választásra az adhatna alapot, hogy a számítógép előtt egy virtuális emberrel beszélgető felhasználó tudatában van annak, hogy nem egy másik emberrel, hanem egy többé vagy kevésbé élethűnek tűnő számítógépes szimulációval cserél eszmét, akitől nem is vár el gazdag és élethű gesztikulálást. Ám hogy ez nem így van, újabb és újabb kísérletek erősítik meg.

Egyrészt az ember igen érzékeny a mindennapi életben nagy változatosságban és tömegben tapasztalt gesztusokat illetőleg. Mechanikus, nem az emberi mozgásdinamikát követő, és mindannyiszor pontosan azonos formában ismétlődő mozgások unalmassá és zavaróvá válnak, és a virtuális lény elveszti az élő lény illúzióját. De nem elég csupán változatos és a mozgást tekintve emberszerű gesztikulációról gondoskodni. C. Nass és kollegái egy sereg vizsgálatot végeztek, melyek azt bizonyították, hogy a virtuális lény kommunikációját meghatározó paraméterek közül egyet megváltoztatva, más lett a virtuális lény által keltett szubjektív benyomás, és ami még fontosabb, objektív hatás [9, 10, 13]. Például, gesztusokra szorítkozva, attól függően, hogy a virtuális lény milyen (mozdulatlan) testhelyzetben jelent meg, a felhasználók többé vagy kevésbé adtak hitelt a tanácsainak. A preferencia attól függött, hogy a felhasználó maga extrovertált vagy introvertált személyiségű-e. Egy másik példa azt mutatta, hogy pusztán a szemöldököt kissé összevonva, alig észrevehetően szigorúbb kifejezést adva egy virtuális lénynek, a kísérleti alanyok jobb hatásfokkal végzik el az általa kirótt feladatot, viszont kevésbé élvezik a feladatvégzést [20]. Tehát a gesztusok részleteit is értelmezi az ember, és azok alapján (is), jóllehet öntudatlanul, reagál.

Egy virtuális lény nemverbális kommunikációját ezért igen körültekintően kell megtervezni. A nemverbális gesztusoknak összhangban kell lenniük továbbá az egyéb modalitások (kinézet, nyelvhasználat, hanghordozás) által keltett benyomásnak, hogy az összbenyomás egy konzisztens személyiség legyen.

2.3 A szemöldök szerepe kiemelésben

Ebben a fejezetben egy saját kutatási eredményt ismertetünk annak illusztrálására, hogy egy virtuális ember nemverbális kommunikációjának egyetlen részlete is miféle hatást vált ki. Azt vizsgáltuk, hogy egy beszélő fej esetében [16] a megemelt szemöldök jelzi-e a dialógus kontextusában az információ fontosságát, újdonságát (prominence) [5]. A kísérlet keretében egy beszélő fej hollandul a „blauw vierkant" (kék négyzet) jelzős szerkezetet ismételte különböző változatokban: az első, a második, illetve mindkét szót hangsúlyozva, és mindhárom esetben további 4 variációban aszerint hogy kíséri-e megemelt szemöldök az első, a második, mindkét vagy semelyik szó kiejtését. A nézőnek két kérdésre kellett válaszolnia két-két variáció megtekintése után: melyiket tartja természetesebbnek? Melyikben kap hangsúlyt a jelző, illetve a jelzett szó?

A kísérlet két tanulsággal szolgált. Egyrészt a vokális hangsúllyal egybeeső megemelt szemöldökös változatokat tekintették a nézők a legtermészetesebbnek. Ha mindkét szó hangsúlyozott volt, akkor csak az elsőt kísérő szemöldökmozgást kedvelték.

Továbbá a megemelt szemöldök igenis megerősítette a vokálisan is kiemelt szó prominenciáját, és „leárnyékolta" a megelőző vagy követő szón lévő vokális hangsúly kiemelő szerepét.

A kísérlet részletei megtalálhatók a cikkekben. Itt csak azt emeljük ki, hogy nem ismert hasonló adat igazi beszélőket tekintve, aminek egyik oka, hogy igen nehéz egy embernek úgy beszélnie, hogy vokálisan és vizuálisan más szót hangsúlyozzon. A mesterséges beszélővel kapott eredményeket tekintherjük az emberi beszédre is jellemzőnek.

2.4 Extrovertált virtuális ember tervezése

A mindennapi életben tapasztalt jelenség, hogy egy nyitott, extrovertált személyiségű embernek élénkebb a beszéde és az arc mimikája, mint egy zárkózott, introvertált személyiségé. Virtuális emberek esetében is fontos, hogy a lény konzisztens személyiségjegyeket mutasson, ezek megléte – aszerint, hogy a virtuális lény személyiség illek-e a felhasználóéhoz – kedvezően vagy éppen negatívan befolyásolja a virtuális lény megítélését, és közvetett módon, az általa képviselt szolgáltatás igénybevételét vagy kirótt feladat elvégzésének sikerességét [9].

Az előzőt követő kísérletünkben arra kerestünk választ, hogy a beszédet milyen arc mimikai jellegzetességekkel kell kísérni, hogy a virtuális lény extrovertált vagy introvertált benyomást keltsen [5]. Három paraméter változtattunk egy beszélő fejen: a szintetizált beszéd, a tekintet és a szemöldökmozgás mindegyikét elkészítettük extrovertált és introvertált formában. A szemöldök introvertált esetben nem mozgott, extrovertált esetben kétszer megemelkedett két vokálisan hangsúlyozott szón. Az introvertált tekintet azt jelentette, hogy a fej 300 ms alatt lefele-balra mozgatta a sze-

mét, és csak a beszéd vége felé emelte fel a tekintetét ismét (ld. 1. ábra). Extrovertált esetben a karakter mindvégig a hallgató szemébe nézett, pislogásokkal megszakítva. A 8 lehetséges változatban jelentéssel nem bíró modern verssorokat mondattunk a.beszélő fejjel. A vokális jellemzők és a kísérlet részletei megtalálhatók a cikkben.

Fig. 1. Introvertált és extrovertált beszélő fej.

A teszt személyeknek a beszélő fejet két szempontból kellett minősíteniük: mennyire természetes, és milyen személyiségű. Az utóbbi kérdés meglepő eredménnyel szolgált. A beszélő fej már akkor is extrovertált benyomást keltett, ha a 3 jegyből kettő volt extrovertált. A megítélést csak kissé javította, ha mindhárom modalitásban extrovertált jeleket produkált a beszélő fej.

Ez az eredmény egyrészt felveti a kérdést, hogy mi ennek a jelenségnek az emberi kommunikációban és percepcióban rejlő, mélyebb oka. Másrészt elvi alapot ad arra, hogy tudatosan válasszuk ki azt a két modalitást, mellyel sikeresen kelthetjük az extrovertált személyiség benyomását. Például jó minőségű, paraméterezhető beszédszintetizátor hiányában (ami gyakori eset) elég az arc mimikára nagyobb gondot fordítani, hogy kompenzálódjon a szintetizált beszéd semleges vagy akar introvertált jellege.

3. Gesztusok modellezése

Ha egy virtuális lény arc vagy kéz gesztusainak megtervezéséhez kezd az ember, három alapvető problémával találja szembe magát:

1. Milyen valójában a reprodukálandó gesztus? Mik a morfológiai és dinamikai karakterisztikái? Milyen apró eltérések lehetségesek, sőt, szükségesek, hogy életszerű és nem mechanikus, robot hatást keltsenek?
2. Egy gesztust különböző helyzetekben, például más és más beszédet kisérve vagy más-más célpontra mutatva akarunk használni. Ez a morfológiai és dinamikai jellemzők paraméterezésével érhető el. Mik az alkalmas paraméterek, és ezek lehetséges értékei?
3. A gesztusokat szeretnénk alkalmazni különböző virtuális lényekre, ahelyett hogy minden esetben újból létrehozzuk a gesztus repertoárt.

A 3. probléma a szabványos virtuális lény modellek és animációk témájához vezet, mellyel itt nem foglalkozunk Az 1. és 2. problémát alább egyetlen gesztus, a mosoly példáján tárgyaljuk.

3.1 A mosoly dinamikája és fajtái

Ahhoz, hogy modellezni tudjuk a mosolyt, igazi emberi mosolyokat kell részleteiben megfigyelnünk. Erre, durván, két út van: a tudományos kísérletezés és leírás, illetve a művészi absztrahálás. Mi korábbi munkánk során mindkettőt alkalmaztuk, amikor mosolyt terveztünk [3]. A tudományos módszer bevett gyakorlata szerint igazi emberek arcán előhívott mosolyt tapogattunk le, az arcon megjelölt, a mosoly során mozgó pontok koordinátáinak időfüggvénye formájában.

Amint az a 2. ábrán látható, az egyes pontok (miként az azok mozgását előidéző megfelelő arcizmok) koordináltan, többé-kevésbé szimmetrikusan mozognak. Elkülöníthető a mosoly *megjelenésének, megtartásának és eltűnésének* 3 fázisa. Noha mi is és mások is tapasztalták a szájcsücskök mozgásának „kettős hegycsúcs" alakú függvénygörbéjét, mind a mai napig bevett gyakorlat, hogy a mosolyt (és a többi arckifejezést is) trapéz alakú aktivációs függvényekkel jellemzik, az arckifejezés egy ideig mozdulatlanul van jelen az arcon. Nem tisztázott, hogy az ábrán is látható görbének mik az általános jegyei, mik az egyéni jellegzetességek, és mi a mintavételi vagy a kísérlet egyéb körülményeiből adódó zaj?

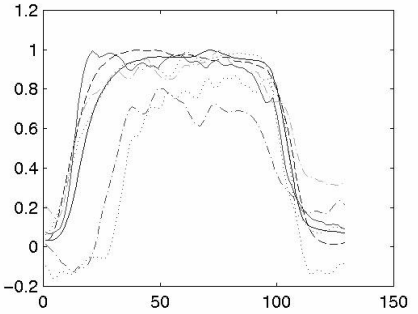


Fig. 2. Egy emberi mosoly során az arcon megjelölt pontok mozgása.

Az egyszerű, trapéz alakú aktivációs függvényt feltételezve, máris számtalan kérdést kell tisztáznunk ahhoz, hogy ilyen függvényekkel illúziókeltő mosolyt fakasszunk egy virtuális arcon. Milyen rövid, illetve hosszú lehet egy mosoly? Ha időben skálázunk egy mosolyt, hogyan változik a 3 szakasz időtartama? Mi a helyzet az intenzitást illetőleg, milyen az átsuhanó mosoly és széles mosoly szájmozgás függvénye? Szimmetrikusnak kell-e lennie az arc mozgásának?

Mindezek megválaszolásában a fő nehézség, hogy nincs elég, természetes helyzetben felvett mosolyról adatunk. Márpedig a spontán helyzet alap követelmény (lenne), mivel ismert, hogy a parancsra kiváltott mosoly különbözik az igazi érzelmet tükrözőtől. Egy másik probléma, hogy mi is az az érzelem, ami a mosolyt kiváltja? Az elégedettség, viszontlátás öröme, csodálat, káröröm, gúny mind mosollyal járnak – melyik ezek közül az az érzelem, amely *a* mosoly? Megfelel-e ez az általunk megcél-

zott esetben, amikor is a virtuális lény például eladó, tanár vagy pszichológus szerepét kell hogy betöltse?

Számunkra igen tanulságos volt, amikor az arckifejezések letapogatásának nehézségeivel szembesülve, egy grafikusművészt kértünk meg, hogy fejből tervezzen arckifejezéseket. Ő a fentiek szerint minősített mosolyok sokaságát tervezte, melyek mind a szubjektív szemlélő, mind a számítógépes felismerő rendszer számára karakterisztikusabbak és részletgazdagabbak voltak, mint a kísérleti körülmények között előhívott igazi mosolyok.

3.2 Parametrizált gesztus repertoár

Egy virtuális lény gesztus repertoárját szisztematikusan, és egy-egy alkalmazási területnek megfelelően építhetjük fel. Gesztusokat *gesztus primitívek párhuzamos és szekvenciális komponálásával* határozhatunk meg. Pl. az ijedt arckifejezés három primitív, a szemöldökemelés, szem és szájkerekítés. Egy ilyen arckifejezést numerikus vagy kvalitatív módon paraméterezhetünk (pl. intenzitás és 1 között, kissé ijedt vagy rettentően ijedt). A gesztusok időzítését rendszerint beszédhez igazítjuk. Itt merül fel, hogy milyen hosszú is legyen-lehet egy arckifejezés: egy-egy szótag, szó, vagy mondat időtartamú? Láttuk, hogy egy-egy gesztus, különösen kéz gesztusok, többféle változatban is kivitelezhető. A mozgás amplitúdója, simasága, a dinamikai jellemzői (lásd indulatos gesztikulálás) mind paraméterekkel vezéreltek kell hogy legyenek, mivel a gesztus ilyen részletei például érzelmi vagy egyéni jelleget tükröznek. Az általunk kifejlesztett, alább bemutatásra kerülő nemverbális kommunikációt deklaráló GESTYLE nyelvben párhuzamos és szekvenciális kompozícióval definiálhatók összetett arc és kéz gesztusok. A gesztus egyes fázisainak hossza, a mozgás milyenség, pontossága, arckifejezések szimmetriája mind paraméterekkel szabályozhatók [11].

4. Multimodális kommunikáció GESTYLE-ban

4.1 A vezérlés szintjei: gesztustól a stílusig

A GESTYLE nyelv egy olyan szöveg annotáló nyelv, mellyel az előbb felsorolt követelményeknek megfelelően, több szinten és több modalitást tekintve lehet előírni a beszédet kísérő (vagy helyettesítő) nemverbális jeleket [11].

1. **Gesztus repertoár definiálása** Elemi unimodális gesztusokból multimodális, összetett gesztusokat lehet definiálni. A virtuális lényt megjelenítő animáló motornak végül pontosan időzített elemi gesztusokat kell majd megjelenítenie.
2. **Gesztus könyvtár megadás** Egy-egy gesztus (egy vagy több) jelentést fejezhet ki. A jelentés-gesztus több-többértelmű, és a választások tekintetében nemdeterminisztikus hozzárendelés egy-egy könyvtárba összegyűjtve szerepel. Egy ilyen könyvtár valamilyen foglalkozási, etnikai vagy egyéb csoport körében honos gesztushasználatot rögzíti.

3. **Stílus** A virtuális lény stílusát a használatos gesztus könyvtárak, valamint a gesztikuláció specifikus jegyeinek megadásával definiáljuk. Pl. egy introvertált tanár, akire aszimmetrikus szemöldökmozgás jellemző, a tanári foglalkozás és introvertált személyiség gesztus könyvtárait fogja használni úgy, hogy az ottani, szemöldök mozgást magukba foglaló gesztusok m módosulnak.
4. **Szöveg annotálás címkékkel** A (beszédszintetizátor által majd elmondandó) szöveget magas szinten, *jelentés címkékkel* tűzdelhetjük meg, melyek értelmezése a használatra kerülő gesztus könyvtárak alapján képeződik le egy-egy gesztusra, mely időzítése a szintetizált beszédhez igazítva, automatikusan történik. Emellett *gesztus címkékkel* gesztusok közvetlenül, a beszédhez igazítva vagy abszolút időzítéssel is előírhatók.
5. **Dinamikus hatások** Az érzelmi és egyéb dinamikus címkékkel egy VL gesztikulációja dinamikusan is változtatható, a stílusban definiáltakat időlegesen felülírva. Például szomorú érzelmi állapotban, a gesztusok száma és intenzitása csökken, az arckifejezésnek, szemmozgásnak lesz nagyobb szerepe a kéz gesztusokkal szemben.

4.2 Egy példa

Az előadás során látható és hallható lesz egy virtuális lény, melyet az alább annotált szövegrészlettel vezéreltünk:

```
1 <StyledText>
2 <StyleDeclaration>
3 <OrderedElements>
4 <Style aspect="PERSONALITY" dict="Extravert" />
5 </OrderedElements>
6 </StyleDeclaration>
7 <TextBody>
8 <Meaning name = "Happy">
9 <Meaning name="Greet">  Hello </Meaning>
10 I am a gesturing avatar.
11 </Meaning>
12 <Meaning name = "Sad">
13 Sorry for
14 <Meaning name = "EmphasizeMild"> not </Meaning>
15 being properly dressed ...
16 </Meaning>
17 Do you allow
18 <Gesture name="PointAtSelf" intensity="INTENSE”> me </Gesture>
19 in?
20 </TextBody>
21</StyledText>
```

A fenti példában a VE személyisége az egyetlen, amely meghatározza a gesztikulációját. Csupán a 4. sorban a PERSONALITY címke megváltoztatásával introvertált személyiséggé, a VE ugyanezt a szöveg más, introvertált gesztikulációval fogja előadni (ld. 3. ábra). Megjegyezzük, hogy a GESTYLE fordító többször egymás után esetleg más-más gesztusokat rendel a jelentés címkékhez, a könyvtárban szereplő valószínűségi megfeleltetés alapján. Így pl. verbálisan adott üdvözlés többszöri ismét-

léskor többé vagy kevésbé más gesztusokkal történik, hasonlóan az emberi szokásokhoz.

Láthatjuk azt is, hogy a szöveg során a VE kedve megváltozik (8, 12 sorok). Ennek eredményeként pl. a 9 sorban az üdvözlést nagyobb amplitúdójú integető gesztus fejezi ki. A jelentés címkék mellett a 18 sorban egy intenzitás paraméterrel megadott, a szöveghez szinkronizált gesztus szerepel.

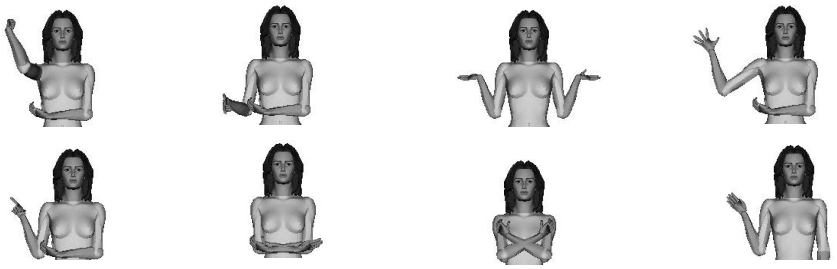


Fig. 3. Gesztusok extrovertált és introvertált személyiségű VE esetében.

6. Nyitott kérdések

Végül a sok lehetséges nyitott kérés közül olyanokat sorolunk fel, amelyek számítógépes nyelvészek számára érdekesek lehetnek.

A fentiekben azt vázoltuk, hogy miként lehet címkézett szövegből kiindulva előállítani a szöveget kísérő gesztusokat. Kérdés, hogy egy szöveg szintaktikus és szemantikus elemzésével generálhatók-e a jelentés címkék automatikusan? Milyen módon, nyelvtannal írhatók le a címkékkel megtűzdelt magyar mondatok?

A nemverbális stílus generálását össze kéne kapcsolni a nyelvben és diskurzusban megjelenő stílussal is [19]. Legvérmesebb álmunk egy olyan, stílusában paraméterezhető VL, mely Queneau „Stilusgyakorlatok"-jának mintájára egyetlen történetet különböző nyelvi, narratív és nemverbális stílusban képes előadni [12].

Szinte semmit nem tudunk a nemverbális jelek és a magyar beszéd szinkronizálásáról. Vannak-e, és mik a magyar nyelvből adódó sajátosságok? Az idézetthez hasonló kísérlet megismétlését tervezzük magyar résztvevőkkel – érdekes lenne közel azonos kísérleti környezetben összehasonlítani pl. holland és magyar alanyokkal kapott eredményeket.

Nem tértünk ki arra, hogy az egyéniség, érzelmi állapot a beszédben is kell hogy tükröződjön. Vannak kezdeti kutatások arra nézve, hogy miként lehet, a gesztusokhoz hasonló módon, a szintetikus beszédet is például szomorúra vagy izgatottra hangolni. A GESTYLE nyelv elvben képes ilyen változtatásokra is [15, 18]. A Profivox TTS rendszerrel lehetne a magyar nyelvre hasonló környezetet kialakítani. A munka dandárját az érzelmes beszéd analitikus leírása és modellezése jelentené.

Végül, egy igen érdekes, hasznos és szakmailag is figyelemreméltó alkalmazás lenne különböző nyelvi, nemverbális és diskurzus stratégiákat megtapasztalni, kipróbálni programozható virtuális lényekkel. Például sokan nem is veszik észre, hogy milyen zavaró hatású lehet, ha túl kevés vagy túl sok szemkontaktust tartanak beszélgető partnerükkel. Egy virtuális lénnyel beszélgetve ezt gyakorolni, tanulni lehetne.

A legutolsó kérdés elvezet a VL alkalmazásának legkritikusabb problémájához: a mai virtuális lények elég emberszerűen tudnak előadni, percepciós képességük, beszédértésük viszont igen korlátozott.

Köszönetnyilvánítás

A cikk a Magyar Felsőoktatási Minisztérium Szent-Györgyi Albert Ösztöndíjának támogatásával készült.

Irodalomjegyzék

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E. "*Embodied Conversational Agents*", MIT Press, Cambridge, MA. 2000.
2. Chi, D., Costa M., Zhao L., Badler N. "The EMOTE Model for Effort and Shape", *Proc. of Siggraph*, 2000. pp. 173-182.
3. Hendrix, J., Ruttkay, Zs. Exploring the space of emotional faces of subjects without acting experience, *CWI Report INS-R0013*, Amsterdam, 2000
4. Kendon, A. "Human Gesture", In: Ingold, T. and Gibson K. (eds.) *Tools, Language and Intelligence*, Cambridge University Press, 1993.
5. Krahmer, E., S. van Buuren, Ruttkay, Zs., W. Wesselink: Audio-visual Personality Cues for Embodied Agents: An experimental evaluation, Proc. of the AAMAS03 Ws on "Embodied Conversational Characters as Individuals", 2003, Melbourne, Australia
6. Krahmer, E., Ruttkay, Zs., Swerts, M., Wesselink, W. Audiovisual Cues to Prominence. In: Proceedings International Conference Spoken Language Processing, Denver, CO, 2002, pp. 1933-1936.
7. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press. 1991.
8. McNeill, David (Ed.) *Language and gesture*. Cambridge: Cambridge University Press, 2000.
9. Nass C., Isbister K., Lee E-J. "Truth is Beauty: Researching Embodied Conversational Agents," In: [1], pp. 374-402.
10. Nass, C.I., Steuer, J., Tauber, E. "Computers are Social Actors", *Proc. of CHI'94*, Boston, MA, 1994.
11. Noot, H. Ruttkay, Zs. Variations in Gesturing and Speech by GESTYLE, International Journal of Human-Computer Studies, Special Issue on 'Subtle Expressivity for Characters and Robots', to appear in 2005.
12. Queneau R., Wright B. (Translator): Exercises in Style, New Directions 1981.
13. Reeves, B., Nass, C. The Media Equation – How People Treat Computers, Television and New Media Like Real People and Places, CUP, 1996.
14. Ruttkay, Zs., Pelachaud, C., Poggi, I., Noot, H. Exercises of Style for Virtual Humans, In: L. Canamero, R. Aylett (Eds.), Animating Expressive Characters for Social Interactions, Advances in Consciousness Research Series, John Benjamins Publishing Company, to appear.

15. Ruttkay, Zs., V. van Moppes, Noot, H. The jovial, the reserved and the robot, Proc. of the AAMAS03 Ws on "Embodied Conversational Characters as Individuals", 15th July, 2003, Melbourne, Australia
16. Ruttkay, Zs., Noot, H. Animated CharToon Faces, *Proceedings of NPAR 2000 - First International Symposium on Non Photorealistic Animation and Rendering,* pp 91-100, June 2000.
17. Ruttkay, Zs., Pelachaud, C. (Eds.), From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer, 2004.
18. Van Moppes, V. Improving the Quality of Synthesized Speech Through Mark-up of Input Text with Emotions, Master Thesis, VU, Amsterdam, 2002.
19. Walker, M., Cahn, J., Whittaker, S. "Improvising Linguistic Style: Social and Affective Bases for Agent Personality", *Proc. of Autonomous Agents Conf.* 1997.
20. Walker, J., Sproull, L., Subramani, R. "Using a Human Face in an Interface", *Proc. of CHI'94*, pp. 85-91.

VIII. Poszter prezentációk

A Szeged Korpusz és Treebank verzióinak története

Csendes Dóra[1], Alexin Zoltán[1], Csirik János[1], Kocsor András[2]

[1] Szegedi Tudományegyetem Informatikai Tanszékcsoport
6720 Szeged, Árpád tér
{dcsendes, alexin, csirik}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Aradi vértanúk tere 1.
{kocsor}@inf.u-szeged.hu

1 Bevezetés

A Szegedi Tudományegyetem Informatikai Tanszékcsoportján 1998 óta folytatott természetesnyelvi kutatások és fejlesztések egyik fő célja egy nagyméretű, kézileg annotált szöveges adatbázis kialakítása volt. Tettük ezt azért, hogy további számítógépes nyelvészeti kutatásokhoz jó minőségű alapot biztosítsunk, ill. számítógépes tanuló algoritmusok számára egy nagy megbízhatóságú adatbázist hozzunk létre. A munkálatok eredményeként mára a szegedi szövegállomány négy különböző verziója készült el Szeged Korpusz 1.0, Szeged Korpusz 2.0, Szeged Treebank 1.0, ill. Szeged Treebank 2.0 néven. Az összes verzió kialakításánál egy automatikus előelemzési fázist egy részletes kézi ellenőrzés és javítás követett. Az alábbiakban röviden beszámolunk a szövegállomány fejlesztési munkálatairól. A négy verzióban szereplő fájlok XML formátumúak, belső szerkezetüket a TEIXLite, ill. a TEI P4 DTD séma írja le.

2 Rövid történeti áttekintés

2.1 Szeged Korpusz 1.0

A Szeged Korpusz [1] szövegeinek gyűjtése 1999-ben kezdődött meg. A vállalkozásra a MUTEXT-EAST projekt [4] ösztönözte az akkori konzorciumot[79], amelynek keretében megszületett a TELRI korpusz magyar változata, és megtörténtek az első kísérletek a szövegek automatikus elemzésére. Ennek továbbfejlesztésére vettük célba egy bővebb szövegállomány összeállítását, hogy a további számítógépes nyelvészeti kísérleteket már egy reprezentatívabb korpuszon végezhessük. A szövegek kiválasztása során a legfőbb szempont az volt, hogy tematikailag a lehető legkülönfélébbek legyenek. Végül öt különböző témakörből választottuk ki a szövegeket, nevezetesen: szépirodalmi írásokból, 14-16 éves tanulók fogalmazásaiból, napilapokban és folyóiratokban megjelent újságcikkekből, jogi szövegekből, és számítástechnikai szövegekből, témakörönként kb. 200 ezer szó terjedelemben. Az így összegyűjtött korpusz 1 millió szövegszót és további 200 ezer írásjelet tartalmaz. Természetesen ez a mennyiség nem elegendő ahhoz, hogy a teljes mai magyar nyelv szókészletét és nyelvtani

[79] Szegedi Tudományegyetem Informatikai Tanszékcsoport, MorphoLogic Kft.

struktúráját lefedje, de elég reprezentatívnak bizonyul abból a szempontból, hogy számítógépes nyelvészeti kutatásokat lehessen rá alapozni.

A Szeged Korpusz első verziója egy morfo-szintaktikailag elemzett és kézzel egyértelműsített természetesnyelvi szöveges adatbázis, amely 139.000 különböző szóalakot tartalmaz. A szövegek morfo-szintaktikai annotálásához a nemzetközileg elfogadott MSD (Morpho-Syntactic Description) kódrendszert használtuk. A korpusz jelen verziója a többjelentésű szavak esetében csak a kiválasztott morfo-szintaktikai kódokat tartalmazza, a lehetségeseket nem tünteti fel.

2.2 Szeged Korpusz 2.0

A Szeged Korpusz 2.0 verziója az 1.0 verzió kibővítésével keletkezett. A meglévő 1 millió szavas szövegállományt az akkori konzorcium[80] egy 200 ezer szavas rövidhír részkorpusszal bővítette ki, amely elsősorban gazdasági és pénzügyi rövidhíreket tartalmaz. Így a korpusz 1,2 millió szövegszavasra nőtt, amely 155.500 különböző szóalakot tartalmaz, és további 250 ezer írásjelet is magában foglal. A korpusz második verziója az elsőhöz hasonlóan egy morfo-szintaktikailag elemzett és kézzel egyértelműsített természetesnyelvi szöveges adatbázis. A méretnövekedésen kívül az első verziótól abban tér el, hogy a kontextusnak megfelelően kiválasztott morfo-szintaktikai kódok mellett a lehetséges kódok is szerepelnek az adatbázisban, így hatékonyan alkalmazható automatikus szófaji annotáló módszerek tesztelésére.

2.3 Szeged Treebank 1.0

A természetesnyelvi feldolgozás fontos lépése a szintaktikai elemzés és annotálás, azaz a különböző szintaktikai egységek, pl. főnévi vagy melléknévi csoportok, névutós szerkezetek bejelölése. Mivel a mondatok többségében az egész mondat jelentése szempontjából a főnévi csoportok (NP-k) kulcsfontosságú szerepet játszanak, ezért a Szeged Treebank 1.0 [2] verziójában ezeknek a szerkezeteknek a bejelölése volt az elsődleges cél. Ezen kívül, ugyancsak a mondatok tartalmának értelmezhetősége szempontjából, fontos szerepe van a tagmondatok (CP-k) elkülönülésének és egymáshoz való viszonyának (alárendelés, mellérendelés), ezért ezeket is jelöltük a szövegeken. A főnévi csoportok és tagmondatok bejelölését a Szeged Korpusz 2.0 állományán, 82.000 mondaton végeztük.

Számos olyan alkalmazásról tudunk, ahol elegendő a szövegek részleges szintaktikai elemzése (shallow parsing). Ilyen pl. az automatikus információkinyerés (information extraction) vagy kivonatolás (text summarisation) is. Az itt leírt Szeged Treebank 1.0 verzió ilyen alkalmazásokban került felhasználásra, ill. további elemzéshez szolgál kiindulópontként.

2.4 Szeged Treebank 2.0

A Szeged Treebank 2.0 [3] az első verziónál gazdagabb elemzést és annotációt tartalmaz. Jelen verzió magában foglalja az összes előző verzió eredményeit (morfo-szintaktikai, NP- és CP-annotálást), és ezt kiegészíti további szintaktikai elemzéssel, amely a melléknévi, határozószói csoportok, névutós szerkezetek, igék, stb. bejelölését foglalja magában. A treebank kialakításakor a már ismert forrásmunkákra és meglévő elméletekre támaszkodtunk. Ezek tanulmányozásával és összevetésével nyelvész

[80] SZTE Informatikai Tanszékcsoport, MorphoLogic Kft, MTA Nyelvtudományi Intézete

szakértőink egy konzisztens szintaktikai szabályrendszert dolgoztak ki a generatív szintaxis szabályainak megfelelően. A használt szintaktikai címkék a nemzetközi szabványnak megfelelőek, és lehetővé teszik az adott szintaktikai szerkezetre vonatkozó attribútumok tárolását is. A Szeged Treebank 2.0-ra vonatkozó statisztikai adatokat az alábbi két összefoglaló táblázat mutatja.

1. Táblázat: A szintaxisfák mélység szerinti eloszlása a treebank mondataiban

Szintaxisfa-mélység	**1**	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**
Fogalmazások	141	2922	7898	8388	3942	1380	62	0
Jogi szövegek	2	110	687	1554	2127	3346	1337	115
Újságcikkek	29	577	1466	2469	2545	2567	534	24
Üzleti hírek	0	75	864	2396	2844	2933	455	10
Szépirodalom	493	4649	5230	4170	2373	1495	152	2
Számítástechnika	9	541	1133	2413	2654	2638	373	7
Összes	**674**	**8874**	**17278**	**21390**	**16485**	**14359**	**2913**	**158**

2. Táblázat: A szintaxisfák szélesség szerinti eloszlása a treebank mondataiban

Szintaxisfa-szélesség	**1**	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**	**21-50**	**50-**
Fogalmazások	25	126	319	578	1109	2811	4738	11309	3667	51
Jogi szövegek	20	56	60	72	48	147	429	2640	5153	653
Újságcikkek	1	83	97	120	156	438	1000	3693	4401	222
Üzleti hírek	1	0	2	11	158	114	502	3741	5006	42
Szépirodalom	15	434	1099	1336	1397	2691	3095	5487	2864	146
Számítástechnika	104	142	108	80	130	266	681	3643	4430	184
Összes	**166**	**841**	**1685**	**2197**	**2998**	**6467**	**10445**	**30513**	**25521**	**1298**

3 Gépi tanulási alkalmazások

A részletes kézi annotálásnak köszönhetően a Szeged Korpusz és Szeged Treebank különböző verziói megbízható tanulási és tesztelési adatbázisként szolgálnak számítógépes tanulóalgoritmusok számára. A Nyelvtechnológiai csoport a következő területeken kísérletezett tanuló algoritmusok alkalmazásával:

- morfo-szintaktikai elemzés és egyértelműsítés,
- részleges és teljes szintaktikai elemzés,
- információkinyerés részleges szemantikai információk (szemantikai keretek) segítségével.

3.1 Morfo-szintaktikai egyértelműsítés (POS tagging)

Nagy sikerrel alkalmaztunk különböző tanulóalgoritmusokat és azok kombinációját a szófaji elemzés és egyértelműsítés területén [6]. A magas találati arányok (97% feletti találati pontosság) figyelemre méltó, hiszen a magyar nyelv gazdag és változatos morfológiája meglehetősen nagy kihívást jelent az automatizált módszerek számára.

Ezen kívül azért is jelentősek az elért eredmények, mert a felhasznált morfo-szintaktikai kódrendszer (MSD) nagyon részletes, így a többértelmű szavak aránya eléri az 50%-ot.

3.2 Szintaktikai elemzés

A Szeged Treebank első verzióját numerikus, szabály-alapú és statisztikai tanulóalgoritmusok, ill ezek kombinációjából összeálló módszerek tanítására használtuk [5]. A cél az volt, hogy automatikus információkinyerés támogatása céljából a tanulóalgoritmusok minél pontosabban tudják beazonosítani a mondatokban szereplő főnévi csoportokat, ill. minél jobban el tudják találni a tagmondatok határait. A legjobb pontossági eredmények 85-92% között vannak, de az algoritmusok teljesítménye nagyban függ a feldolgozott szöveg típusától (pl. jogi vs. újságnyelvi szöveg) és a mondatszerkezetek bonyolultságától.

Folyó kutatásaink egy automatikus szintaktikai elemző kifejlesztésére irányulnak a Szeged Treebank 2.0 verziója alapján, vagyis részletes szintaktikai annotálás felhasználásával. Már a legkorábbi eredmények elérték a 80-85% körüli találati pontosságot, amit bíztató előjelnek tekintünk a további fejlesztések szempontjából. Egy teljes szintaktikai szerkezeteket felismerni és elemezni képes program nemcsak az automatikus információkinyerést segítené nagyban, hanem gépi fordítórendszerek is jól hasznosíthatják.

Bibliográfia

1. Alexin Z., Csirik J., Gyimóthy T., Bibok K., Hatvani Cs., Prószéky G., Tihanyi L.: *Manually Annotated Hungarian Corpus*, in Proc. of the Research Note Sessions of EACL'03, Budapest, Hungary, pp. 53-56 (2003)
2. Csendes, D., Csirik, J., Gyimóthy, T.: *The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus* in Proc. of TSD 2004, Brno, Czech Republic and LNAI vol. 3206, pp. 41-49 (2004)
3. Csendes D., Csirik J., Gyimóthy T., Kocsor A.: *The Szeged Treebank*, in Proc. of TSD 2005, Karlovy Vary, Czech Republic and LNAI vol. 3658, pp. 123-132 (2005)
4. Erjavec, T., Monachini, M.: *Specification and Notation for Lexicon Encoding*, Copernicus Project 106 „MULTEX-EAST", Work Package 1 – Task 1.1, Deliverable D1.1F (1997)
5. Hócza, A., Felföldi, L., Kocsor, A.: *Learning Syntactic Patterns Using Boosting and Other Classifier Combination Schemas* in Proc. of TSD 2005, Karlovy Vary, Czech Republic and LNAI vol. 3658, pp. 69–76 (2005)
6. Kuba, A., Csirik, J., Hócza, A.: *POS tagging of Hungarian with combined statistical and rule-based methods* in Proceedings of TSD 2004, Brno, Czech Republic and LNAI vol. 3206 (2004)

Skálázható szöveg-alapú nyelvazonosító módszer beszédszintézis céljára

Kiss Géza, Németh Géza

Budapest Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}@tmit.bme.hu

Kivonat: Szövegek nyelvének automatikus azonosítása nagyon fontos több alkalmazásterületen. E cikkben áttekintjük a szövegből történő nyelvazonosítása (language identification, LID) használt főbb módszereket és leírjuk legfontosabb tulajdonságaikat. Ezek egyes, nagyon rövid szövegekre helyes kezelését is igénylő alkalmazásterületeken – mint például a beszédszintézis – jelentkező hiányosságai kezelésére egy új módszert mutatunk be, amely változó hosszúságú N-gramok használatán alapuló, tisztán statisztikai módszer, emellett tetszőleges szöveg helyes azonosítására betanítható, jól skálázható, és viszonylag kis számítási kapacitást igényel az azonosítási fázisban. Bemutatjuk hatékonyságát a tanító- és attól független tesztanyagon, különböző méretű szövegtörzseken való tanítás esetén, kevés és nagyon nagy számú nyelven való működés esetén is. Az eredmények igazolják a megközelítés életképességét.

1 Bevezetés

A szövegből történő automatikus nyelvazonosítás (Language Identification, LID) még mindig fontos kutatási terület, bár sok fejlődés történt az elmúlt évtized folyamán.

A számítógépen tárolt szövegek nyelvének automatikus azonosítására számos alkalmazási területen szükség van. Ilyenek például a webes kereső motorok [1], valamint más webes alkalmazások, amelyek az internetet tudásbázisként használják, például ontológiák tanulásához [2], több nyelven elérhető szövegek gyűjtéséhez számítógéppel segített fordítás céljára [3]. Számos természetes-nyelv feldolgozási eljárás alkalmazásához is szükség van a szöveg nyelvének megbízható előzetes megállapítására, pl. kérdés-válasz (question answering) rendszerekben, automatikus fordításnál [4]. Az is igazolást nyert, hogy a nyelvazonosításra használható módszerek esetenként más szempont szerinti kategorizálásra is használhatónak bizonyulnak, mint például téma vagy szerző szerinti osztályozásra [5].

Egy más jellegű, de szintén fontos használati terület a többnyelvű vagy poliglott beszédszintézis, mivel kevert nyelvi környezetekben (pl. elektronikus levelek vagy webes szövegek felolvasásakor) szükséges, hogy pontosan ismerjük a szöveg nyelvét, különben a létrehozott beszéd érthetetlen vagy legalábbis rendkívül kellemetlen hangzású lesz. Ennél az alkalmazási területnél rövid szövegekre, mondatokra, sőt egészen

a szavak szintjéig megbízhatóan meg kell állapítanunk a nyelvet. Ennek oka egyrészt az, hogy esetenként nem áll rendelkezésre hosszabb szövegrész (pl. sms-felolvasás esetén), másrészt az, hogy mondandónk gyakran tartalmaz beékelt idegen nyelvű szavakat, kifejezéseket (pl. személyneveket, művek címét, idegen eredetű szavakat, szakkifejezéseket).

Jelen cikkben egy olyan újonnan kidolgozott nyelvazonosítási módszert mutatunk be, amely tisztán statisztikai alapon működik, de megfelelő méretű tanító szövegtörzs használatával tetszőleges megbízhatóságú nyelvazonosítás elérhető, a szavak szintjén is. Fontos tulajdonsága, hogy jól használható felismerési arány eléréséhez is csekély tárolási és számítási kapacitást igényel az azonosítási fázisban, és jól skálázható a két szempont bármelyike szerint.

2 A módszer bemutatása

2.1 A jelenleg használt nyelvazonosítási módok áttekintése

Az használt technikákban több csoportra oszthatók. Legnagyobb részük az egyes nyelvekre jellemző írásmód felszíni jelenségeit ragadja meg különböző statisztikus jellemzők használatával; ilyenek például a leggyakoribb szavak listájának használata[6], N-gram alapú módszerek [5], vektortér modellek [7], döntési fák [8], vagy neurális hálók [9]. Ezeknek gyakran hosszabb szövegrészre van szüksége a megbízható nyelvazonosításhoz, de a szavak szintjén való azonosításhoz nem elég megbízhatóak. Emellett azok, amelyek a dokumentumot előzetesen tanító szövegtörzsből nyelvenként készített „nyelvi profilokhoz" hasonlítják (pl. [5], [7]), gyakran számottevő számítási kapacitást igényelnek az azonosítási fázisban is. Ez lényeges szempont a felhasználhatóság szempontjából, míg a betanítási fázishoz szükséges számítási kapacitásnak, a tanító algoritmus futási idejének – ésszerű határok között – nincs jelentősége, főként ha az utóbbi rovására az előbbi csökkenthető.

Másik csoportjuk, főként a beszédszintézis említett jellemzői miatt a szószinten való helyes azonosításra törekedve részletes morfológiai elemzést alkalmaz, pl. DCG-k (Definite Clause Grammar) használatával [10], esetleg közvetve egy helyesírás-ellenőrző használatával [11]. Egy köztes megoldásban nem történik valódi morfológiai elemzés, hanem szótárak (szó és szóelem-listák) elemeire való illeszkedés alapján következtetnek a szavak nyelvére, kiegészítve ezt statisztikai módszerekkel [12].

Összefoglalásként elmondható, hogy a jelenleg használt, tisztán statisztikai alapú megközelítések általában nem adnak eléggé pontos nyelvazonosítást rövid szövegeken, és/vagy nagy számítási kapacitást igényelnek az azonosítási fázisban, míg a részletes morfológiai elemzés végzése nehezen kivitelezhető, főként nagyszámú nyelvre, valamint problémát okozhat egyes alkalmazásokban a szükséges számítási kapacitás.

2.2 A probléma választott megközelítése

A célunk olyan megoldás kidolgozása volt, amely lehetővé teszi nagyon rövid szövegek helyes azonosítását is, akár a szavak szintjéig, és amely kézben tartható abban az értelemben, hogy be lehet tanítani tetszőleges bemenet helyes azonosítására, de egy-

ben általánosító képességgel is rendelkezik, azaz nem látott szavak nyelvét is képes helyesen felismerni a tanítóhalmaz szavaihoz való hasonlóság alapján. Emellett célunk volt a működéshez szükséges adatbázis méretének korlátok között tartása is.

Ennek a célnak megfelel, ha a P (szó | nyelv) valószínűséget egy előzetesen rögzített kritériumnak megfelelő pontossággal becsüljük meg, majd arra a nyelvre döntünk, amelyhez tartozik, ill. homográfok (több nyelvben előforduló szóalak) esetén arra a nyelvre, amelyben a legnagyobb az előfordulásának valószínűsége. A szavakra meghatározott nyelvi címkék alapján dönthetünk a szövegrész nyelvére. A szavak kontextusa alapján számított nyelv-valószínűség figyelembe vételével akár szószinten helyes nyelvazonosítást is kaphatunk, még homomorf szavak esetén is. Ez azt is lehetővé teszi, hogy egy egynyelvű szövegbe beszúrt idegen nyelvű szó a valódi nyelvének megfelelő azonosítást kapja, szemben a környezet alapján determinisztikusan döntő naiv megközelítéssel.

Megfelelő valószínűség-becslési módszerrel az ismert szavak írásmódja alapján képesek lehetünk korábban nem látott szavakra is becsülni ezt a valószínűséget. Ez a megközelítés megőrzi a szó-alapú módszerek előnyét, a kézbentarthatóságot, kiterjesztve azt általánosító képességgel, és szóalapon is helyes működést tesz lehetővé.

2.3 A kidolgozott módszer leírása

Az általunk kidolgozott módszer változó méretű N-gramok használatán alapszik. Míg a szokványos Markov-modellt használó megoldásban rögzített hosszúságú előzményt használunk egy karakternek az előzőek utána való következése valószínűségének becslésére, a javasolt módszerben többféle hosszúságú előzményt használunk, amely hosszt minden környezetre egy tanítási folyamat során határozunk meg. A tanítás 0 hosszúságú karakter környezettel indul minden karakterre (ez a karakter előfordulásának valószínűsége), majd ezt a hosszt egyes környezetekben növeli a megcélzott valószínűség-becslési kritérium elérésére, amely lehet pl. a leggyakoribb szavak helyes felismerése. A folyamat korlát nélküli folytatása a láncszabályt adja, ezzel pedig nyelvenkénti szó-valószínűséget, ezért a tanító folyamat tetszőleges tanító halmaz esetén jobb szó-valószínűség becsléshez, így korrekt azonosításhoz konvergál a tanítóhalmazra. Hosszabb N-gramokat tartalmazó, nagyobb méretű felismerő adatbázis használatával pontosabb azonosítási eredmény érhető el megfelelő tanítás esetén.

Az N-gram környzetekhez tartozó feltételes valószínűségeket fában tárolva úgy is felfoghatjuk a módszert, hogy egy fajta döntési fa tanítását jelenti a szóvalószínűségek becslése céljából. A fa bővítésének irányát a bővítésnek a becslési kritérium szempontjából meghatározott „hasznossága” szerint határozzuk meg. Több ilyen hasznosság-függvénnyel dolgoztunk, melyeket a tanító halmazon való helyes felismerési arány (recall) és az attól független teszt-halmazon való eredmény (precision), azaz az általánosító képesség mellett az alapján is vizsgáltunk, hogy mennyire tömör adatbázist. A tömörséget nem pusztán a mérettel jellemeztük – hiszen nem közömbös, hogy milyen felismerési arányt ad a tömörebb adatbázis – hanem a felismerési/méret hányadossal, LID adatbázis teljesítményének nevezünk. A legjobb adatbázis méretet adó függvény a feltételes valószínűségek logaritmusának nyelvek közötti eltérését, míg a legjobb általánosító képességet adó emellett az N-gram előfordulásának valószínűségét is figyelembe vevő, entrópia-jellegű mennyiség.

A módszerben újítás még, hogy a nyelvek független szemlélése helyett a nyelvenkénti valószínűségek eltérésének helyes becslésére törekszünk, amelytől kisebb adat-

bázis méretet várunk, hiszen így a tanítás során a két nyelvet megkülönböztető jellemzőkre való „koncentrálásra” késztetjük az algoritmust.

3 Eredmények

Több tesztet végezünk eltérő méretű tanító és felismerendő beszédkorpuszon. Először három nyelvre (angol, német, magyar) végeztünk betanítás nagyméretű korpuszon (British National Corpus, Project Guttenberg DE, Magyar Elektronikus Könyvtár), azoknak a hozzávegyült idegen nyelvű részektől való megtisztítása nélkül, a leggyakoribb szavak 90%-ának helyes felismerésére. A tesztet az előzőtől független szöveghalmazon végeztük (Project Guttenberg, online magyar újságok). Az [5]-ben bemutatott módszer egy web-en megtalálható implementációjához[81] használt, 77 nyelvhez tartozó kis méretű (5 kilobájt) szövegre is elvégeztük a betanítást.

A helyes azonosítás százalékos eredményeit az 1. táblázat tartalmazza. Az osztályozott szövegek áttekintése azt mutatta, hogy az első esetben a más nyelvűnek osztályozott szövegek gyakran valóban nem a csoportjuknak megfelelő nyelvhez tartoztak, vagy kevert nyelvűek voltak, valamint hogy valóban pontos szó-alapú működéshez szükség van egyes (formátumukat tekintve) nyelv-függetlennek tekinthető kifejezések azonosítására, melyekre példák a római számok, internet és e-mail címek, dátumok, nemzetközi szavak (pl. „tel.”, „fax.”), rövidítések, mértékegységeket tartalmazó kifejezések (pl. „2 cal”).

Vizsgáltuk a szavak környezete alapján számított nyelv-valószínűség figyelembevételének hatását is. Ehhez az azonosító által nyelvi címkézett szövegből kalibráltunk nyelv-valószínűség becslő szabályokat, amelyek a szomszédos szavak nyelvével való egyezés valószínűségét adták. Ezekkel a szószintű azonosítási eredménye a harmadik adatbázist használva a német korpuszon a korábbi 45%-ról 65%-ra növekedett, majd a folyamatot ismételve 70%-ra, igazolva az iteratív tanítás-finomítás létjogosultságát.

1. Táblázat: eredmények különböző tanító halmazok esetén,
egy azoktól független 3 nyelvű teszt szöveggel, szó és mondat szintű azonosításra

Tanító (szavak)	Nyelvek	Adatbázis	Tanító, szó	Teszt, szó	Teszt, mondat
20641974-89138922	3	54 kbájt	99,6%	94,2%	98,5%-99,5%
580-694 (5 kbájt)	3	7,4 kbájt	95,5%-97,8%	79,6%-87,4%	91,7%-97,2%
465-1681 (5 kbájt)	77	5,4 mbájt	70,0%-99,8%	30,1%-59,6%	71%-84,0%

4 Konklúziók

Bemutatottunk egy új szöveg-alapú nyelvazonosítási módszert, amely a nyelvenkénti szóvalószínűségek eltérésének döntési fa-alapú becslésén alapszik. A módszer statisztikai alapú, így viszonylag könnyen létrehozható nagy számú nyelvre működő változata, jól skálázható a felismerési arány és az adatbázis méret viszonylatában,

[81] http://odur.let.rug.nl/~vannoord/TextCat/Demo/

tetszőleges szó kívánt azonosítására tanítható, és az azonosítási fázisban relatíve kicsi számítási kapacitást igényel. További kutatást igényel a bemutatott, leggyakoribb szavakra való tanítás összehasonlítása a döntési fa adott pontosságú feltételes-valószínűség ill. ezek nyelvek közötti eltérésének becslésére való tanítás. A módszer várhatóan más kategorizálási feladatokban való felhasználása is lehetséges, például szófaj-címkézés (POS tagging).

Bibliográfia

1. Risvik, K. M., Michelsen, R.: Search engines and Web dynamics. Computer Networks, Vol. 39, Issue 3. (2002) 289-302
2. Kilgarri, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Computational Linguistics, 29(3) (2003) 333-348
3. Volk, M.: Using the Web as Corpus for Linguistic Research. In: Pajusalu, R., Hennoste T. (eds.): Tähendusepüüdja. Catcher of the Meaning. Festschrift for Professor Haldour Õim. University of Tartu, Estonia: Publications of the Department of General Linguistics 3 (2002)
4. Bond, F.: Toward a Science of Machine Translation. Proc. of the MT Roadmap Workshop at TMI-2002, Keihanna, Japan (2002)
5. Canvar, W. B., Trenkle, J. M.: N-gram based Text Categorization. Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas (1994) 161-176
6. Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: The Design, Implementation and Operation of a Hungarian E-mail Reader. International Journal of Speech Technology, Kluwer Acedemic Publishers, Vol. 3, Numbers 3/4. (2000) 217-236
7. Prager, J. M.: Linguini: Language Identification for Multilingual Documents. Proc. of the Thirty-Second Annual Hawaii International Conference on System Sciences, Vol. 1. (1999) 2035
8. Häkkinen, J., Tian, J.: N-gram and Decision Tree-based Language Identification for Written Words. Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio Trento, Italy (2001)
9. Tian, J., Suontausta, J.: Scalable neural network based language identification from written text. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 1 (2003) 48-51
10. Pfister, B., Romsdorfer, H.: Mixed-lingual text analysis for polyglot TTS synthesis. Proc. of Eurospeech 2003 (2003) 2037-2040
11. Halácsy, P., Kornai, A., Németh, L.,, Rung, L., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. Proc. of LREC 2004 (2004) 203-210
12. Marcadet, J. C., Fischer, V., Waast-Richard, C.: A Transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. Proc. of Eurospeech 2005 (2005) 2249-2252

Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból

Miháltz Márton[1], Pohl Gábor[2]

[1]MorphoLogic, Orbánhegyi út 5, 1126 Budapest
mihaltz@morphologic.hu
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

A cikkben bemutatunk egy kísérletet, melynek célja, hogy automatikus módszerekkel annotált tanítókorpuszokat állítsunk elő angol-magyar, illetve magyar-angol fordítórendszerben működő jelentés-egyértelműsítő modul számára. A tanítópéldákat, melyekben a forrásnyelven többértelmű (tehát több lehetséges fordítással rendelkező) szavakat célnyelvű fordításaikkal látunk el, nagyméretű, mondatszinten szinkronizált párhuzamos korpuszból nyerjük ki. Az annotáló algoritmus kétnyelvű szótárak és statisztikus heurisztikák alkalmazásával működik.

Egy olyan szabályalapú gépi fordítórendszerben, mint a MetaMorpho megértéstámogató fordítóprogram [6], jelentős kihívást jelent a többértelmű lexikális elemek kezelése. A forrásnyelvi nyelvtani elemzés során csak korlátozott mértékben van lehetőség a forrásnyelven többértelmű, következésképpen a célnyelven is általában több különböző fordítással rendelkező szavak egyértelműsítésére. Többértelmű főnévi, melléknévi és gyakran igei elemeknél a rendszernek szüksége van külső segítségre, melyhez egy statisztikai gépi tanuláson alapuló jelentés-egyértelműsítő alrendszert fejlesztettünk [3]. Ez a modul a forrásnyelven többértelmű szó eredeti kontextusában (a fordítási egység bekezdésében) megfigyelt szemantikai és szintaktikai információk alapján hoz döntést a legvalószínűbb célnyelvi fordításról.

Minden, a forrásnyelven többértelmű szóhoz külön osztályozót használunk, melyek annotált tanítópéldákból betanított modelleken alapulnak. Megfelelő tanítópéldák előállítására korábban angol nyelvű lexikális erőforrásokkal (Princeton WordNet) annotált korpuszokat használtunk, melyekben az angol jelentés-címkéket magyar fordításokra képeztük le. Mivel ilyen korpuszok véges mennyiségben állnak csak rendelkezésre, a rendszer felskálázásához további megoldásokra van szükség. Az egyik lehetőség angol korpuszokból kigyűjtött példák kézi annotálása a többértelmű szavak magyar fordításaival. A kézi annotálás azonban rendkívül időigényes, és ezért költséges folyamat.

Egy másik, kedvezőbb alternatíva párhuzamos korpuszok felhasználása. Mivel a jelentés-egyértelműsítő modulnak a mi esetünkben eleve célnyelvi fordításokkal annotált tanítópéldákra van szüksége, a kétnyelvű, mondatszinten szinkronizált szövegben a többértelmű szavakat a másik oldalon megtalálható fordításaikat azonosítva juthatunk megfelelő tanítóanyaghoz ([1], [5]).

A Hunglish projektben elkészített Hunglish kétnyelvű angol-magyar párhuzamos korpusz 44,6 millió angol, illetve 34,6 millió magyar szövegszót tartalmaz [7]. A szabadon felhasználható, nagypontosságú mondatszintű illesztéssel ellátott korpuszt szeretnénk felhasználni mind többértelmű angol, mind többértelmű magyar szavak előfordulásainak automatikus annotálásához.

A korpusz angol oldalán automatikus szófaj-egyértelműsítőt (POS-tagger) [2] alkalmazunk, mivel a MetaMorpho rendszerben egy adott szóalak különböző szófajú előfordulásaihoz külön jelentés-egyértelműsítő modell betanítása szükséges. A többjelentésű szó lehetséges fordításait tövesítve keressük a fordításban, ha több változatot is találunk, a mondatpárt nem egyértelműként jelöljük meg. Az ilyen mondatpárok esetleg elhagyhatók, kézzel egyértelműsíthetők, vagy ha túl gyakoriak (gyakori többértelmű igék esetében), automatikus módszerrel is megpróbálhatjuk egyértelműsíteni őket: a szó (szavak) környezetét szótárral, illetve szószintű szinkronizációs algoritmusokkal leképezve a mondatpárban. Ha nem találunk egy keresett szóhoz fordítást, megvizsgáljuk, hogy ismert kifejezés részét képezi-e a mondatban (ezekkel nem foglalkozunk). A szavak ismeretlen fordítású előfordulásait tartalmazó mondatpárokban megpróbálunk statisztikai módszerekkel [4] valószínű fordításokat keresni, a jó találatokkal bővítjük a lehetséges fordítások halmazát, majd megismételjük az eljárást.

Az algoritmus eredményeinek kiértékelésére a következő metodológiát tervezzük. Az angol-magyar irány ellenőrzéséhez kiválasztunk 10, a British National Corpusban gyakori és a WordNet szerint többértelmű szót, és ezekhez a Hunglish korpuszban 50-50 véletlenszerűen kiválasztott példában meghatározzuk a magyar szövegben a fordításaikat (ha vannak). A magyar-angol irányban hasonló módon hozunk létre kiértékelő halmazt, csak a WordNet helyett egy magyar-angol középszótár segítségével kiválasztva a többértelmű, a Magyar Nemzeti Szövegtárban gyakori szavakat. Ezeken a halmazokon futtatva az algoritmust meghatározzuk a humán annotációhoz képesti pontosságot és a lefedettséget.

Bibliográfia

1. Diab, M.: Relieving the data acquisition bottleneck for Word Sense Disambiguation. In Proceedings of ACL (2004)
2. Giménez, J., L. Márquez: SVMTool: A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004 .
3. Miháltz, M.: Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő modullal. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004)
4. Och, F. J., Ney, H.: Improved Statistical Alignment Models. In Proceedings. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.
5. Specia, L., M. G. Volpe Nunes, M. Stevenson: Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria (2005)
6. Tihanyi, L.: A MetaMorpho projekt 2004-ben. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004)
7. Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón: Parallel corpora for medium density languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria (2005)

Morfológiai idioszinkrázia többszavas kifejezésekben

Oravecz Csaba, Varasdi Károly, Nagy Viktor[1]

[1] MTA Nyelvtudományi Intézet,
Budapest 1068, Benczúr u. 33.
{oravecz,varasdi,nagy}@nytud.hu

Kivonat: A dolgozatban megvizsgáljuk, hogy magyar nyelven egyes szókapcsolatok morfológiailag idioszinkratikus viselkedése, mint lehetséges információforrás, használható-e többszavas kifejezések korpuszból történő kinyerésére. Megmutatjuk, hogy legalábbis egyes TSZK típusok esetén, a toldalékolás idioszinkráziája jól jelzi a szókapcsolat TSZK státuszát illetve idiomatikusságát.

1. Bevezető

A számítógépes nyelvfeldolgozásban az utóbbi időkben számos módszert fejlesztettek ki többszavas kifejezések (TSZK) korpuszból történő kinyerésére illetve azonosítására [3]. Többségük a korpuszból kinyert pozíciós illetve relációs jelöltlisták [2] tagjait rangsorolja valamilyen asszociációs mérték segítségével. Olyan nyelvek esetében azonban, melyek morfológiája pl. az angolnál sokkal gazdagabb információforrást jelent, a kutatás éppen csak elkezdődött az „együtt előfordulás" mellett egyéb információ felhasználására [1]. Magyar nyelvre egyes szókapcsolatok morfológiailag idioszinkratikus viselkedése természetesen adódik, mint lehetséges további információforrás, melyet TSZK-k bizonyos csoportjainak azonosításában fel lehet használni.

A dolgozatban megvizsgáljuk a szókapcsolat jelöltek tagjainak toldalékeloszlásából kinyerhető információ felhasználhatóságát, és esettanulmányokon keresztül megmutatjuk, hogy legalábbis egyes TSZK típusok esetén, a toldalékolás idioszinkráziája jól jelezheti a szókapcsolat TSZK státuszát illetve idiomatikusságát.

2. A kivonatoló módszer

Egy szósorozatot akkor tekintünk morfológiailag illetve morfoszintaktikailag idioszinkratikusnak, ha egyes tagjainak toldalékeloszlása az adott szókapcsolatban jelentősen eltér a tagok összes előfordulásra vetített toldalékeloszlásától. Ez a megközelítés bizonyos mértékben eltér [1] módszerétől, ahol adott inflexiós jegyek csupán a már azonosított TSZK-n belül kerülnek összehasonlításra, és a jegyek egyes értékeinek (pl. egyes vagy többes szám) aránya a TSZK morfoszintaktikai preferenciájának jelzésére szolgál. Az általunk alkalmazott eljárás más megközelítésben, általános módszerként kívánja felhasználni a szókapcsolaton kívüli illetve belüli toldalékelosz-

lást, és az ebből kinyert információ segítségével próbálja azonosítani a TSZK-t. Ezáltal független osztályozóként az együtt előforduláson alapuló mértékek helyett, és nem utánuk, mint további feldolgozó lépés kíván szerepelni.

A munkahipotézis a következő. A jelöltlista valamilyen szintaktikai viszonyban álló 2 szavas kombinációkat tartalmaz, ahol szabad morfoszintaktikai jegyeknek nevezzük azokat a jegyeket, amelyeket nem ez a viszony kényszerít ki (egyeztetéssel vagy kormányzással). Ezek akkor vagy a tagok inherens jegyei, vagy a mondatszerkezet másik frázisa írja elő meglétüket. Pl. a „bedobja ... törölközőt" TSZK-ban a tárgyrag nem szabad, mert az állítmány-tárgy viszony írja elő, viszont a törölköző szám- stb. jegye szabad. Az ige minden lehetséges jegye szabad. A hipotézis az, hogy egy TSZK tag szabad jegyekre vett statisztikai eloszlása eltér az ő összesített (itt a szótő vagy a nem szabad jeggyel ellátott szótő összes előfordulását tekintjük) eloszlásától, ha a TSZK tagjaként fordul elő, és ez az eltérés jól jelzi a szemantikai átlátszatlanságot. Lehetséges viszont, hogy pusztán az is megváltoztatja a jegyeloszlást, hogy valamilyen szintaktikai viszonyban áll a tő. Ezért szűkebb környezetre kell az eloszlást vizsgálni, és a csupán az ugyanazon szintaktikai viszonyban álló alakok eloszlásának különbözőségét figyelembe venni. A tesztet a TSZK mindkét tagjára külön végre lehet hajtani, és így azt is megkaphatjuk, melyik tag jelentése változott meg a TSZK-ba kerüléskor.

3. Statisztikai vizsgálat

Az inflexiós elemzést az 1. táblázat szerint osztjuk fel dimenziókra.

1. táblázat. A különböző szófajoknál figyelembe vett inflexiós jegyek

Szófaj	**Dimenziók**				
Névszók	szám	birtokos szám/személy	anafonikus possessivus	eset	fok (melléknévnél)
Igék	mód/idő	határozottság	szám/személy	–	–

Minden potenciális többszavas kifejezésben (C) a tagok inflexiós eloszlását ezen jegyek mentén parametrizáljuk. Egy paraméter egy jegy (F) – érték (v) párt képvisel. Minden paraméterhez hozzárendeljük a jegy-érték pár relatív gyakoriságát:

$$P(F_i = v_j \mid w_k, C) = \frac{c(F_i(w_k) = v_j \ ha \ w_k C \ tagja)}{c(C)} \quad (1)$$

Nyilvánvalóan fennáll a következő összefüggés: $\sum_j \frac{c(F_i = v_j)}{c(C)} = 1$. Ezt az eloszlást kell összehasonlítani az összesített $P(F_i = v_j \mid w_k)$ eloszlással, vagyis amikor a tagszó előfordulásait nem korlátozzuk arra, hogy tagja legyen a többszavas kifeje-

zésnek, viszont feltételül kell szabni, hogy ugyanolyan szintaktikai szerkezeti pozícióban legyen, mint C-ben (pl. ha a TSZK-ban a tagszó főnevet módosító melléknév, akkor az összesített eloszlásban nem vesszük figyelembe azokat az előfordulásokat, amikor állítmányi szerepű).

$$P\ (F_i = v_j \mid w_k) = \frac{c\ (F_i(w_k) = v_j)}{c\ (w_k)} \tag{2}$$

A vizsgálatokban több szókapcsolatjelölt toldalékolási mintáját elemeztük. Számos esetben volt felfedezhető összefüggés az eloszlás egyenetlensége és a szósorozat idiomatikussága között, mely mutatja, hogy a tárgyalt megközelítés mindenképpen biztató eredményeket ad. További kutatást igényel viszont az eloszlások összehasonlítását végző legjobb mérték kiválasztása, illetve a poliszémiából származó torz adatok ki-küszöbölésének módja is.

Bibliográfia

1. Evert, S., Heid, U., Spranger, K.: Identifying morphosyntactic preferences in collocations. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 907–910
2. Evert, S., Krenn, B.: Computational approaches to collocations. Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003) (2003) Vienna.
3. Krenn, B.: The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 7. PhD thesis, Universität des Saarlandes, Department of Computational Linguistics (2000)

WordNet relációk szerepének vizsgálata a jelentés-egyértelműsítésben

Szarvas György[1], Csendes Dóra[1] és Kocsor András[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2., Hungary
{dcsendes, szarvas}@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
6720 Szeged, Aradi vértanúk tere 1., Hungary
kocsor@inf.u-szeged.hu

1 Bevezetés

A jelentés egyértelműsítés feladata az egyes szóalakok konkrét jelentésének meghatározása, a szövegkörnyezet segítségével. A lehetséges jelentések halmaza általában rendelkezésre áll valamilyen elektronikus szótár, vagy ontológia/nyelvi adatbázis formájában. Ebben a cikkben ismertetjük az angol nyelvű WordNet [1] ontológia különböző relációinak jelentés-egyértelműsítésben való felhasználhatóságát vizsgáló kutatásunkat.

Gyakran felügyelt tanulási módszereket alkalmaznak a szóalakok adott szövegkörnyezetben legvalószínűbb jelentésének kiválasztására, melyhez kézileg egyértelműsített példákat tartalmazó korpuszra van szükség. Ilyen korpusz angol nyelvre a SemCor [1] korpusz, mely ingyenesen hozzáférhető. Kísérleteink során ezt a korpuszt fogjuk használni a paraméterek hangolására, illetve tesztelési célokra.

A másik gyakori megközelítés nem igényli előre annotált korpusz meglétét [5, 7, 8], helyette az elektronikus szótárban kódolt információ (leggyakrabban a szóalakok definíciói, glosszái, illetve a szótári egységek között definiált relációk) képezi az egyértelműsítés alapját. Ezek a módszerek a szavak glosszái közötti egyezéseket, illetve a relációs gráfban a szóalakok között értelmezhető utak hosszát, mint szemantikai értelemben vett távolságot használják fel.

Természetesen találhatóak a sokféle heurisztikát ötvöző jelentés-egyértelműsítő rendszerek is [6].

2 Kísérletek

A folyó szövegek szavaihoz azok WordNet ontológiabeli jelentésének hozzárendelésére a mondat szavainak egymástól vett ontológiabeli távolságát vizsgáltuk. Módszerünk mögött az a feltevés húzódik meg, hogy a mondatok szavai szemantikailag koherens struktúrát alkotnak az ontológia gráfjában, azaz a jelentés-hozzárendelés elvé-

gezhető a szóalakok valamely távolságmetrika mellett vett legközelebbi rendszerének megkeresésével.

Számos szemantikai hasonlóságot gráfbeli távolsággal definiáló módszer létezik, azonban ezek legtöbbször az ontológia hierarchikus (hipo-, hipernim) relációit veszik figyelembe [3], vagy más területre – weblapok rendszerezett adatbázisán – lettek kifejlesztve [4]. Munkánk célja a sokféle WordNet-reláció fontosságának kivizsgálása a jelentés azonosításban való hasznosságuk szerint, hogy a megfelelő súlyozással egy, az eddigieknél hatékonyabb, gráfbeli távolságalapú egyértelműsítő heurisztikát kapjunk (esetleg azonosítsuk az egyértelműsítés szempontjából haszontalan relációkat).

Kísérleteinkhez a WordNet ontológiát, kiértékelésre a jelentés-egyértelműsített SemCor korpuszt használtuk, a megfelelő magyar erőforrások hiányában. A kidolgozott módszerek magyar WordNet birtokában magyar nyelvre is átültethetők.

Az általunk bemutatott módszer a Bevezetésben tárgyalt két típus között helyezkedik el, hiszen az egyértelműsítéshez használt metrika az utóbbi, felügyelet nélküli módszerekkel rokon, azonban célunk volt az eddig az egyértelműsítés során kiaknázatlan relációfajták felhasználása, illetve vizsgálata a hatékonyságra gyakorolt hatásukat illetően, melyhez címkézett adatokat használhatunk validációs célokra. A szerzők tudomása szerint nem készült az összes, WordNet ontológiában tárolt relációt felhasználó távolságalapú jelentés egyértelműsítő módszer. Az egyetlen, a hierarchikus relációkon kívül mást is felhasználó eljárás Hirst és St. Onge módszere [2], valamint annak különböző változatai.

Az elvégzett kísérletek során a mondat szavainak összes jelentését figyelembe véve kiszámítjuk a szavak által kifeszített részgráfban a csúcsok távolságának összegét. A legkisebb távolságösszeggel rendelkező részgráf kijelöli a mondat szavainak az adott kontextusban legvalószínűbb jelentését. A távolság számításához az egyes relációkat különböző súlyokkal vehetjük figyelembe, ekkor minden lehetséges súlyozás egy-egy jelentés-egyértelműsítő heurisztikát definiál. A súlyok hangolásával kereshetjük a többértelműségek feloldásához legalkalmasabb súlyozást.

A lentebb található ábrán a „NOR COULD HE CALL UP (#3) MEMORY-PICTURES (#1) OF CLOSE (#2) FRIENDS (#1) OR RELATIVES (#1).” angol mondat szavai által definiált szemantikustávolság-mátrix látható. A mátrix egyes soraiban az egyértelműsítendő szavak egyes lehetséges jelentéseinek a többi szó különböző alternatíváitól való távolságait mutatja. A cél minden szóhoz egy lehetséges jelentés hozzárendelése úgy, hogy a kijelölt sorok, és oszlopok metszeteiben álló elemek összege legyen minimális. A fenti példa mondat esetén a legegyszerűbb, minden relációt egyforma, egységnyi súllyal figyelembe vevő heurisztika egyértelmű optimumot definiál, mely pontosan a jelentések helyes hozzárendelését adja.

Jelenleg is folyó kutatásunkban a következő feladatokra koncentrálunk:

- Az ábrán látható típusú mátrixokban a legkisebb össztávolságot definiáló hozzárendelést megadó hatékony algoritmus kifejlesztése
- Egyértelmű (egyetlen lehetséges jelentéssel bíró) szavak kitüntetett szerepének vizsgálata – javíthat-e a pontosságon vagy sebességen az ilyen szavak kitüntetett kezelése

- A minimális össztávolságú struktúra meghatározásakor a mondat szavainak az összes többitől vett távolságát célszerű vizsgálni, vagy elegendő egy bizonyos sugarú környezetben (a mondat távoli részei közt is mutatható ki szemantikai értelemben vett kohézió, vagy csak az egymáshoz közeli szavak között)
- Szintaktikai elemzés információinak felhasználása a keresés szűkítésére milyen hatással van az egyértelműsítésre (a keresést nem a szó szoros környezetére, hanem a vele nyelvtani kapcsolatban álló szavakra korlátozzuk)
- Változó élsúlyok használata a távolság számításakor (pl. egy hipo-hiperním kapcsolatot leíró él nem azonos szemantikai távolságot definiál a hierarchia alsóbb és felsőbb, absztraktabb fogalmakat leíró részeiben)

	call up # 1	*call up # 2*	call up # 3	*call up # 4*	memory picture # 1	*close # 1*	close # 2	*close # 3*	*close # 4*	*close # 5*	friend # 1	*friend # 2*	*friend # 3*	*friend # 4*	*friend # 5*	relative # 1	*relative # 2*
call up # 1					8	9	8	9	11	9	6	8	6	7	8	6	5
call up # 2					9	8	7	8	10	8	5	6	5	6	7	5	6
call up # 3					**6**	5	**4**	5	8	5	**5**	7	5	5	7	**5**	6
call up # 4					9	8	8	8	7	8	6	6	6	7	8	6	6
memory picture # 1	8	9	**6**	9		10	**9**	10	12	10	**7**	9	7	7	9	**7**	8
close # 1	9	8	5	8	10						6	8	6	7	8	6	7
close # 2	8	7	**4**	8	**9**						**5**	7	5	6	7	**5**	6
close # 3	9	8	5	8	10						6	8	6	7	8	6	7
close # 4	11	10	8	7	12						7	9	7	8	9	7	8
close # 5	9	8	5	8	10						6	8	6	7	8	6	7
friend # 1	6	5	**5**	6	**7**	6	**5**	6	7	6						**2**	3
friend # 2	8	6	7	6	9	8	7	8	9	8						4	5
friend # 3	6	5	5	6	7	6	5	6	7	6						2	3
friend # 4	7	6	5	7	7	7	6	7	8	7						3	4
friend # 5	8	7	7	8	9	8	7	8	9	8						4	5
relative # 1	6	5	**5**	6	**7**	6	**5**	6	7	6	**2**	4	2	3	4		
relative # 2	5	6	6	7	8	7	6	7	8	7	3	5	3	4	5		

1. ábra: A „Nor could he call up memory-pictures of close friends or relatives.” mondat szavainak szemantikustávolság-mátrixa. A nem megfelelő jelentések dőlt betűvel, a megfelelő jelentések szemantikus távolságai vastag, nagyobb számokkal szerepelnek.

3 Összegzés

Az általunk ismertetett szóalakok jelentésének WordNet-synsetek szerinti egyértelműsítése évtizedes múltra visszatekintő kutatási irányzat. Az irodalomban fellelhető egyértelműsítő eljárások egyik nagy csoportját a WordNet struktúrán, mint

címkézett, irányított gráfon értelmezett szemantikai távolság/hasonlóság metrikán alapuló algoritmusok adják, a dolgozatban ismertetett módszer is ezek közé sorolható.

A szerzők által bemutatott jelentés-egyértelműsítő az eddig ismerteken túlmutat abban, hogy a WordNet összes relációját felhasználja a gráfbeli utak hosszának vizsgálatakor, hiszen az összes ontológiai reláció szemantikai értelemben vett kapcsolatot teremt az egyes fogalmak között, így figyelembevételük a szemantikai távolság számításakor indokolt. A vizsgálatok eredményeként nemcsak egy új jelentés-egyértelműsítő heurisztikát kapunk, de a távolságmérték számításához felhasznált súlyok hangolásával a WordNet relációinak értékelését is megkapjuk, mely megmondja, hogy a kérdéses relációtípus mennyire jól használható szavak többértelműségének feloldására.

Bibliográfia

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press, USA (1998)
2. Hirst, G., St. Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.), WordNet: An electronic lexical database, pp. 305–332., MIT Press, (1998)
3. Lin, D.: An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conf. on Machine Learning, Madison, Wisconsin (1998)
4. Maguitman, A.G., Menczer, F., Roinestad, H., Vespignani, A.: Algorithmic Detection of Semantic Similarity. Proceedings of the 14th International World Wide Web Conference, Chiba, Japan (2005)
5. McCarthy, D., Koeling, R., Weeds, J.: Ranking WordNet senses automatically, Technical Report CSRP 569. University of Sussex (2004)
6. Mihalcea, R.F., Moldovan, D.I.: A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation. International Journal on Artificial Intelligence Tools, Vol. 10, No. 1-2 (2001)
7. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003)
8. Pedersen, T., Banerjee, S., Patwardhan, S.: Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report UMSI 2005/25 (http://www.msi.umn.edu/general/Reports/rptfiles/2005-25.pdf) University of Minnesota, Duluth (2005)

Beás nyelvű morfológiai elemző problémái a hunlex-hunmorph rendszerben

Szeredi Dániel[1]

[1] MTA Nyelvtudományi Intézet – Eötvös Loránd Tudományegyetem BTK
Elméleti Nyelvészeti Tanszéki Szakcsoport
Budapest 1068, Benczúr u. 33.
dani@szeredi.hu

Kivonat: A beás nyelv flektáló nyelv, így morfológiája nagy mértékben különbözik az agglutinatív nyelvekétől, amelyekre a hunmorph és a hunlex rendszerek a legkönnyebben alkalmazhatóak. Ennek következtében többféle probléma merül fel az elemző készítése során, ám a hunlex-ben található eszközök újféle használatával ezek kezelhetőek.

A rendszer. A hunmorph morfológiai elemző [3] felépítése leginkább az agglutinatív típusú nyelveket preferálja: az egyes szóalakokat szótőre és affixumokra választja szét. A hunlex lexikonkezelő [2] pedig előállítja a hunmorph működéséhez szükséges nyelvi erőforrást egy szabályrendszerből és egy lexikonból. A hunlex már rendelkezik olyan eszközökkel, amelyekkel nemkonkatenatív szabályok is leírhatóak, hiszen karaktereket vághat le, a töveket képes reguláris szabályok illesztésével változtatni. Kérdéses, hogy a hunmorph hogyan képes kezelni a flektáló nyelveket, amelyekre nem a könnyen szegmentálható affixumok a jellemzőek, hanem az inflexió során az egyes toldalékok összeolvadnak egymással, illetve a tővel, tehát a kimenetek lineárisan nem szegmentálhatóak.

A beás nyelv. A morfológiai elemző és a lexikonkezelő ilyen típusú nyelveken való tesztelésére a beás igen alkalmas, mivel erősen flektáló jellegű. A beás nyelvet Magyarországon élő romák beszélik, főleg a déli, délnyugati megyékben. A magyarországi cigányoknak körülbelül 5 százaléka beszéli a beást, amely a románnak közeli rokona, tehát nagyon messze áll az ismertebb, ind eredetű lovári nyelvtől. A beás formális leírása csupán a legutóbbi időkben indult meg [1], így a rendelkezésre álló adatok még nem teljeskörűek, főként az igeragozás tekintetében.

Problémák. A beás nyelvnek a hunlex–hunmorph rendszerben történő morfológiai elemzése során tehát a legfontosabb probléma az, hogy a flektáló nyelvben nem különülnek el a szótő és a toldalékok egymástól. A beásban egy lexémának több, egymásból szabályosan képezhető töve van. Ezt egy konkatenatív szabályokat alkalmazó keretrendszerben nem lehet leírni, minden egyes affixhoz meg kellene adni az adott tőváltakozást. Ez azonban egy ilyen nyelvben értelmetlen, tehát meg kell kísérelni először a töveket levezetni, majd ezekből származtatni a különböző alakokat. Így míg az agglutináló magyarban az egyes szóalakok levezetése, elemzése során a hunlexben megadott szabályok nagyrészt megfelelnek az egyes toldalékmorfémák-

nak, egy beás szóalakban nem ilyen egyértelmű ez. Kérdés, milyen szabályok alkalmazódnak pl. a *fracijê* alakra, amely a *fratje* 'fiútestvér' szó többes számú, távolra mutató határozott alakja? A levezetés melyik pontján kell levágni a szóvégi magánhangzót, és hol zajlik a *tj* □ *c* hangváltozás?

Bonyolítja a helyzetet, hogy egy adott toldalékmorféma allomorfjai gyakran a lexémák eltérő töveihez járulnak, ebből következően egy adott tőalternáció más-más paradigmatikus alakokban jelenik meg szavanként. Például az első palatalizációnak nevezett alternáció a főnevek körében a nőneműek többes számú alakjaiban, és az *–ă* végű nőnemű főnevek egyes szám birtokos esetében zajlik le. Ebben az esetben a probléma az, hogy az egyes szám birtokos esetet képző szabály hogyan oldja meg, hogy az egyes lexémák más alternációt mutatnak ebben az alakban? Amennyiben az ilyen problémák az agglutinatív elemzőkhöz hasonlóan kezelődnének, hamar konfúzzá válnának a szabályok és kezelhetetlenné a morfológiai elemzés.

Megoldási javaslatok. Ezeknek a problémáknak az orvoslásához el kell szakadni attól az elvtől, hogy az egyes `hunlex`-szabályok megfelelnek egy-egy morfémának, tehát az egyes szabályokat kisebb operációk elvégzésére érdemes használni. Így az elemző képessé válik az egy szón megjelenő többféle alternáció kezelésére. Így a fent említett *fracijê* alak létrehozásához először egy szabály (`NOUN_PL_CHOP`) levágja a szóvégi magánhangzót, majd a `NOUN_PL_DIST` osztályozószabály megállapítja, hogy ennek a szónak alakja és lexikai tulajdonságai szerint milyen többes számú alakja van. Ezek alapján továbbküldi a `SEC_PAL` nevű szabályba, amely a *tj* ~ *c* alternációt kezeli, amely maga után von bizonyos, a *fratje* szóban nem megjelenő magánhangzó-váltakozást (pl. *lat* 'széles' ~ t.sz. *lec*), amelyet a `RAISE_MID` kezel. Itt ismét egy osztályozószabályba (`RAISE_MID_DIST`) megy a szóalak, mivel ezeknek az alternációknak akár az igei paradigmából is lehetett bemenete, tehát nem egyértelmű, hogy melyik szabályba kell visszatérni.

Itt ismét nehézségekbe ütközne az agglutinatív nyelvek kezelésére használt módszertan használata, hiszen nem lehet eldönteni, hogy az ilyen alternáció utáni osztályozószabály melyik gyűjtőszabályba utalja az aktuális alakot. Ám megoldást adhat a szálak szétbogozására, ha a `hunlex`ben az egyes lexémákhoz adható morfofonológiai jegyekkel élve a generálódó alakhoz belső használatú jelzőket (pl. `_noun_plur`, vagy `_verb_sg_2`) rendel az a szabály, amelynek a `SEC_PAL` a kimenete. Ezek után a `RAISE_MID_DIST` egyszerűen meg tudja vizsgálni, hogy az adott alak milyen paradigmából érkezett, és milyen szabályba kell küldeni.

Ezután tehát a `NOUN_PL_COLL` gyűjtőszabályban újra összefutnak a különböző többes számú alakokat képző „szabályszálak", és innen már a `hunmorph` számára kimenetként képződik a létrejött alak. A gyűjtőszabályból pedig még továbbhalad a forma különböző, a többes számú tőből képzett alakokat létrehozó szabályokba.

Több esetben nem tűnik szükségesnek a különböző részszabályok szétválasztása, így például az egyes számú birtokos eset gyűjtőszabályának (`NOUN_SG_GEN_COLL`) bemenete minden esetben egy szuffixum-hozzáadó szabály (`NOUN_SG_GEN_SUFF`), így a kettő összevonható lenne. Azonban a szétválasztás egyrészt könnyebben áttekinthetővé teszi a rendszert, másrészt pedig a további bővítéseket (például az igei paradigmák implementálását) is megkönnyíti. Hosszabb távon pedig a szabályrendszer sablonszerűsége is egyszerűsítheti a nyelvtan megalkotását, és segíthet más (főleg flektáló) nyelvek e rendszerben történő leírásában is.

Összefoglalás. A flektáló nyelveket a `hunmorph` elemző szótövek és affixumok leválasztására épülő architektúrája nem képes a leghatékonyabban elemezni. A

`hunlex` lexikonkezelő rendszere így szintén az agglutinatív nyelvek sajátosságain alapul, ám ez a rendszer alkalmassá tehető flektáló nyelvek, így a beás elemzésére, ha a szabályoknak külön-külön egyértelmű funkcióik vannak, amelyek közül a legfontosabbak tehát:

- affixumot hozzáadó szabályok
- tőbeli alternációt kezelő szabályok (mindegyik alternációtípust külön szabály kezel)
- osztályozószabályok, amelyek különböző „szálakra" terelik a különböző fonológiai, vagy lexikális jegyekkel rendelkező lexémákat
- gyűjtőszabályok, amelyek összefogják a különböző „szálakat", és legtöbbször kimenetekként működnek a hunmorph által kezelhető szótárba

Bibliográfia

1. Kálmán László, Orsós Anna: Beás nyelvtan: Alsóbb nyelvi szintek. Kézirat. MTA NYTI Elméleti Nyelvészeti Osztálya (2004)
2. Trón Viktor: HunLex – morfológiai szótárkezelő rendszer. In Alexin Zoltán, Csendes Dóra (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Informatikai Tanszékcsoport (2004) 177-182
3. Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, Dániel Varga: Hunmorph: Open Source Word Analysis. In: Proceedings of ACL 2005 Workshop on Software At the 43rd Annual Meeting of the ACL (2005)

Szótövező eljárások hatása magyar szövegek automatikus kategorizálásánál

Tikk Domonkos[1], Törcsvári Attila[2], Biró György[3], Bánsághi Zoltán[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
tikk@tmit.bme.hu
[2] Arcanum Development Ltd.
H-1117 Budapest, Baranyai u. 10. I/1
attila@arcanum.com
[3] TextMiner Bt.
H-1029 Budapest, Gyulai P. u. 37.
george.biro@gmail.com

Kivonat: Cikkünkben a szövegbányászat területén jellemzően alkalmazott vektortér-modell reprezentáció egyik fontos kérdését, a dimenzióredukciót tárgyaljuk. Ezen belül különböző szótövező eljárások hatását vizsgáljuk több szempontból. Egyrészt azt tekintjük át, milyen összefüggés van az alkalmazott szótövező és a szótár mérete között. Másrészt az egyik szövegbányászati alapfeladat, az osztályozás esetén azt tanulmányozzuk, hogy az egyes szótövezők alkalmazása milyen minőségi következménnyel jár. A vizsgálat során a HunStem szótövezőt, a szópárlista alapú szótövezőt, és egy általunk javasolt ún. óvatos szótövező eljárást hasonlítunk össze. Tesztjeink során a HITEC automatikus osztályozó programcsomagot használtuk.

1 Bevezetés

Szövegbányászati (text mining) és szöveges adatokon végzett információvisszakeresési (information retrieval) módszereknél leggyakrabban *vektortér-modellt* használnak a szövegek reprezentációjára, ahol a vektortér dimenziója megegyezik a vizsgált korpuszban előforduló különböző terminusok (szavak, kifejezések, n-grammok) számával, azaz a *szótár* méretével. Ez már kis méretű, azaz néhány megabájtos korpuszok esetén is igen nagyméretű lehet (ha csak szavak szerepelnek a szótárban akkor is lehet akár 100.000-es nagyságrendű), és a korpusz méretével igaz csökkenő mértékben tovább növekszik. Különösen igaz ez a magyar nyelvű korpuszokra, hiszen a nyelv toldalékoló jellege miatt egy terminus igen sok (akár több tucat) formában fordulhat elő. A nagyméretű szótár mind a tárigény, mind az előfeldolgozás és üzemszerű működés időigénye szempontjából hátrányos, ezért a szótár méretének csökkentése nagy jelentőségű feladat, amit minta-felismerési terminológiával a dimenzió redukálásának is neveznek.

A dimenzió redukálásának egyik leghatékonyabb módja szótövező alkalmazása, azaz amikor valamely szó toldalékolt (pontosabban többnyire csak a ragozott) előfordulásait a szótővel mint kanonikus alakkal helyettesítünk a reprezentációban. Angol nyelvű szövegek esetén ez az eljárás általában 50-65%-kal csökkenti a szótár méretét.

Cikkünkben egy igen gyakori szövegbányászati alkalmazás, a szövegosztályozás esetén vizsgáljuk meg a szótövezés hatását a szótár méretére és az osztályozás hatékonyságára vonatkozóan. A munkánk során a HITEC hierarchikus szövegosztályozót[82] és az [origo]-ról letöltött mintegy 18 ezer dokumentumot, valamint ezen portál kategóriarendszerét használtuk a tesztjeink elvégzésére.

2 Szótövező eljárások

Az általunk kifejlesztett HITEC szövegosztályozó programcsomag lehetőséget nyújt különböző nyelvű szövegek kezelésére és tetszőleges szótövező eljárás integrálására. A különböző nyelvek paramétereit (pl. karakterkészlet, kisbetű-nagybetű párosítás, funkciószavak listája, opcionálisan szótár megadása) egy XML formátumú nyelvdefiníciós állományban lehet megadni. Ugyanitt lehetőség van a szótövező szabályainak, illetve kivételeknek megadására is, de valamely komplett szótövező eljárást külső függvényként meghívva is aktiválni lehet a programból.

Tesztjeink során a szótövezés nélküli feldolgozást az alábbi szótövezők segítségével végzett feldolgozással hasonlítottuk össze:

- Hunstem – a Hunmorph statisztikai alapú ingyenesen elérhető programcsomag szótövező eljárása;
- Timid stemmer – egy általunk fejlesztett néhány szabály és egy korpuszfüggő *lexikon*[83] segítségével bármely nyelvre adaptálható „óvatos” szótövező algoritmus (részletes leírást ld. a 2.1 szakaszban);
- Szópárlista alapú szótövező – adott egy szavak ragozott és kanonikus alakjainak párjait tartalmazó lista, ennek segítségével egy egyszerű eljárás a ragozott alakot kicseréli a kanonikus alakra. A módszer hátránya, hogy a nagyméretű listát a memóriában kell tárolni a feldolgozás során.

Sajnos az összehasonlítás során anyagi okok miatt nem volt lehetőségünk a hazai piacon jelenlévő másik magyar termék, a Morphologic Kft. által fejlesztett HelyesLEM[84] szoftver kipróbálására.

2.1 Az óvatos szótövező

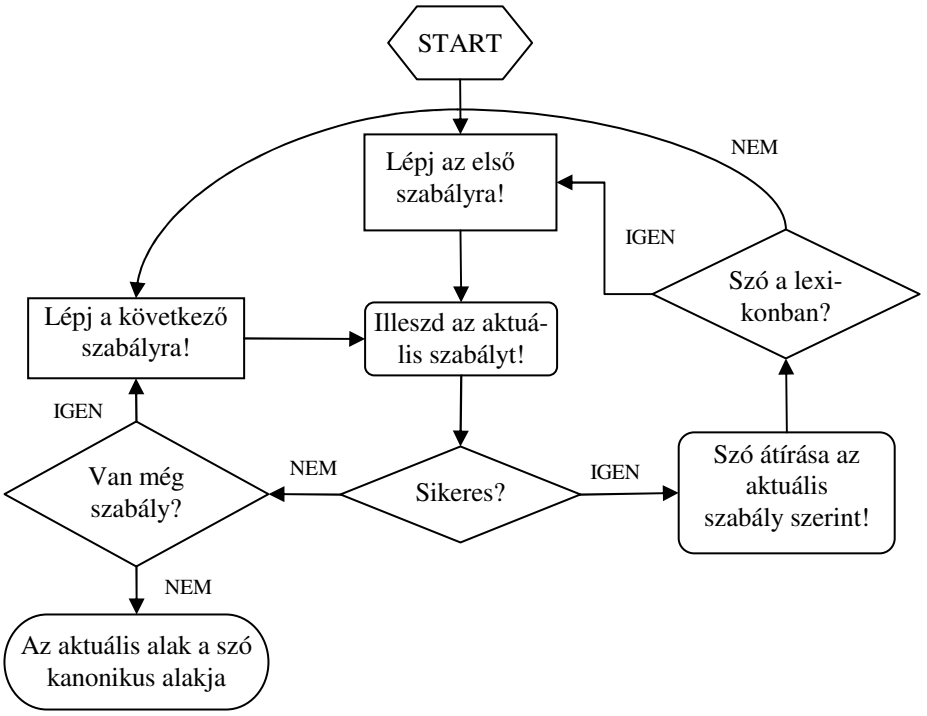
Az óvatos szótövező (timid stemmer) működési elve a következő. Legyen adott egy kiindulási lexikon, és a toldalékok levágását megadó átírószabályok sorozata. Az eljárás sorrendben illeszteni próbálja az átírószabályokat a vizsgált szóra, és amennyiben sikerül, megvizsgálja, hogy az így kapott átírt szó szerepel-e a lexikonban. Amennyiben igen, alkalmazzuk az átírószabályt és az így kapott szóra rekurzívan

[82] http://www.textminer.hu

[83] Ezt az elnevezést használjuk, hogy ne keveredjen a korpusz szavaiból álló szótár fogalommal.

[84] http://morphologic.hu/h_hlem.htm

próbáljuk a lehető legrövidebb kanonikus alakot megtalálni. Amennyiben nem, további illesztésekkel kísérletezik. Az átírószabályok sorrendje egyben prioritást is jelent. Ha az illesztés egyik lehetséges átírószabály esetén sem sikeres, akkor a szó az eredeti alakjában bekerül a lexikonba (ld. 1. ábra).



1. Ábra. Az óvatos szótövező folyamatábrája

Ezen a módon a szöveg maga gazdagítja a lexikont, lehetőséget nyújtva speciális szókészletű szövegek pontosabb szótövezésére. Az átírószabály lehet rag, illetve toldaléklevágó (postfix) szabály, illetve prefixátíró szabály is (pl. igekötők, és a felsőfok jelének a detektálására), valamint a hajlító nyelvek esetén szükséges reguláris kifejezések alkalmazása is. A lexikonban lehetőség van helyettesítő jelek (wildcard) használatára, valamint kivételek és szinonimák megadására is (ami egy alternatív lehetőség a dimenzió redukcióra).

Az óvatos szótövezőnek előnye, hogy nincs hozzá feltétlenül szükség kiinduló lexikonra, hiszen azt a korpusz alapján is létre tudja hozni. Természetesen egy átfogó kiinduló lexikon (illetve egyéb nyelvspecifikus adatok, pl. elhagyandó szavak listájának megadása) növeli a szótövezés hatékonyságát. További előnye, hogy némi nyelvismerettel bárki próbálkozhat átírószabályok megadására, amire a HITEC környezetében egyszerű XML elemek megadásával lehetőség van. Ebből következik, hogy bármely nyelvre könnyen adaptálható.

2.2 Szótövező eljárások összehasonlítása

Az alábbi táblázatban röviden összehasonlítjuk a vizsgált szótövező eljárásokat.

Szótövező	Lexikon bővíthetősége	Szabályok bővíthetősége	Automatikus lexikonépítés	Adaptálás más nyelvekre	Átlagos hatékonyság	Futási idő	Memóriaigény
HunStem	Nehéz	Nehéz	Nincs	Nem	Kiváló	Kicsi	Kicsi
Óvatos	Könnyű	Könnyű	Van	Igen	Jó	Kicsi	Közepes
Szópárlista	Nehézkes	Nincs	Nincs	Igen	Gyenge	Min.	Nagy

Természetesen a leghatékonyabb a magyar nyelvre specializált HunStem eljárás, és külön ki kell emelni, hogy itt a felhasználónak semmilyen nyelvészeti képzettséggel nem kell rendelkeznie. Hátránya azonban, hogy nehézkesen módosítható, hiszen nincsen erre alkalmas felülete, és ezért nehéz valamely szakterület korpuszára, pl. genetikai, vagy kémiai szövegekre alkalmazni. A másik két módszernek fő előnye a rugalmasság, viszont a felhasználónak nagyobb energiát kell befektetnie egy használható verzió létrehozásába. A szótövezők memóriaigénye fordítottan arányos az egyszerűségükkel: a legkisebb tárigénnyel a Hunstem módszer bír. A másik két eljárás esetén a lexikon számottevő memóriatöbbletet jelent, főleg a szópárlista, amelynek mérete 64 MB, több, mint 2,7 millió szópárt tartalmaz, és ami a program futása során kb. 350 MB memóriát igényel.

2.3 HITEC dimenziócsökkentő paraméterei

A HITEC két alapvető dimenziócsökkentő paraméterrel rendelkezik, amely az osztályozási feladatra vonatkozó alábbi intuitív megállapításokat használják ki

- A nagyon alacsony gyakoriságú szavak elhagyhatók, mert nem befolyásolják jelentékenyen az osztályozás minőségét. Küszöbérték: $\min(d_1)$
- Az olyan szavak, amely a dokumentumok (és így persze a kategóriák nagy részében előfordul) nem bírnak megkülönböztető jelleggel a kategóriák között, és így elhagyhatók. Küszöbérték: $\max(d_2)$

A két paraméter optimális értéke természetesen függ a korpusz méretétől és a dokumentumok sokféleségétől, azonban már viszonylag kicsiny d_1 érték esetén jelentősen csökken a szótár mérete.

3 Eredmények

A cikkünkben vizsgált korpusz 130219 az [origo] portálról letöltött dokumentumot tartalmaz. A lemmatizálás után 215423 különböző szót tartalmaz. A szópárlista alapú szótövező ezen szavaknak csak 55%-át ismeri, így amennyiben az ismeretlen szava-

kat változatlan formában hagyjuk, akkor 122027 szó, amennyiben ezeket elhagyjuk, akkor csak 40034 marad a szótárban.

Vizsgálatainkban megmutatjuk, hogy magyar nyelvű szövegek esetén bármely szótövező alkalmazása jelentősen csökkenti a szótár méretét, ez akár 80% feletti eredményt is adhat.

Az osztályozás hatékonysága azonban jóval kisebb mértékben függ a szótár méretétől: a legegyszerűbb szótövező is közel olyan hatékony, mint a Hunstem eljárás.

4 Köszönetnyilvánítás

Tikk Domonkost az MTA Bolyai János kutatói ösztöndíja támogatta.

A Magyar Referencia Beszédadatbázis és alkalmazása orvosi diktálórendszerek kifejlesztéséhez

Vicsi Klára[1], Kocsor András[2],Tóth László[2], Velkei Szabolcs[1],
Szaszák György[1], Teleki Csaba[1], Bánhalmi András[2] és Paczolay Dénes[2]

[1] BME Távközlési és Médiainformatikai Tanszék,
Beszédakusztikai Kutatólaboratórium, 1111 Budapest, Sztoczek u. 2.
{vicsi, velkei, szaszak, teleki}@tmit.bme.hu

[2] MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport,
6720 Szeged, Aradi vértanúk tere 1.
{kocsor, tothl, banhalmi, pdenes}@inf.u-szeged.hu

Kivonat: Poszterünk bemutatja a Magyar Referencia Beszédadatbázist, továbbá az erre épülve párhuzamosan fejlesztett két orvosi diktálórendszer jelenlegi szerkezetét és képességeit.

1 A Magyar Referencia Beszédadatbázis

A *Magyar Referencia Beszédadatbázis*t *(MRBA)* a BME TMIT Beszédakusztikai Laboratóriuma és a szegedi SZTE Informatikai Tanszékcsoport együttműködésben hozta létre [1]. A cél egy olyan irodai, otthoni környezetben olvasott folyamatos szöveget tartalmazó beszédadatbázis megalkotása és akusztikai, nyelvi feldolgozása volt, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére.

Az adatbázis szöveganyagát úgy terveztük meg, hogy lehetőséget adjon különböző típusú beszédfelismerők betanítására és kiértékelésére. Ezek közül a legnagyobb kihívást a folyamatos beszédet felismerő diktáló rendszerek jelentik, amelyeknél a felismerés szónál kisebb felismerési egységek (beszédhangok, difón, trifón egységek) modellezésén alapul. Ezek betanításához olyan folyamatos szöveg összeállítására van szükség, amelyben ezek az elemek elegendően sokszor fordulnak elő, mindamellett a szöveganyag lehetőleg minél rövidebb. Az MRBA szöveganyagának összeállításához újságcikkek szövegét használtuk fel, az adatbázisba bekerülő mondatokat úgy válogatva össze, hogy a leggyakoribb di- és trifónok megfelelő mennyiségben álljanak rendelkezésre. A mondatok mellett fonetikailag gazdag szavakat is kiválasztottunk, az esetlegesen hiányzó vagy nem kellő számban előforduló beszédhangok példányszámának növelése érdekében. Így egy adatközlő 12 különböző mondatot és 12 különböző, a mondatoktól független szót olvas fel, összességében pedig 332 adatközlő hanganyaga került az adatbázisba.

A beszédadatbázis felvételeit különböző helyszíneken: zajos, kevésbé zajos irodai helyiségekben, laborokban, otthonokban rögzítettük. A felvételeknél szinkronban két

különböző rendszerrel dolgoztunk. Az egyik az ún. *referenciarendszer*, amelyben mindig ugyanazt a jó minőségű mikrofont, hangkártyát és laptopot használtunk. A másik rendszer, az ún. *variált rendszer* esetében különböző, jobb, kevésbé jó mikrofonokat, hangkártyákat, PC-ket használtunk, a lehető legnagyobb variáltsággal. A régiók, dialektusok és generációk lefedése céljából a felvételeket Magyarország 4 különböző tájegységében lévő városban rögzítettük: Budapesten, Szegeden, Győrben és Miskolcon, lehetőség szerint különböző életkorú és nemű beszélőket választva.

A felvételek mindegyikét annotáltuk, ami azt jelenti, hogy minden hangfájl mellé egy címkefájlt készítettünk, amely különféle információkat tartalmaz a hangfájl paramétereivel és tartalmával kapcsolatban: az elhangzott szöveg ortografikus lejegyzését, hibás kiejtést, nem érthető szavakat, szótöredékeket, a beszélő nem beszédből származó hangjait, környezeti zajokat, stb. Az adatbázis közel egyharmadán, azaz 100 beszélő anyagán manuálisan fonetikai szintű szegmentálást és címkézést is végeztünk, a fonetikai szegmentumok címkézéséhez a SAMPA nemzetközi kódtáblát használva.

2 Orvosi diktálórendszerekről általában

Az automatikus beszédfelismerési technológia jelenleg még nem képes az általános célú folyamatos diktálás tökéletes megoldására, viszont elfogadható pontosságot tud nyújtani olyan feladatok esetében, ahol a szókincs és a nyelvtani felépítés korlátozott. Így lehetővé teheti az ún. beszédalapú dokumentálást olyan szakmák esetében, amelyek szakszöveg-jellegű dokumentációt igényelnek. Kitűnő példa erre az orvosi vizsgálati eredmények rögzítése, amely folyamat felgyorsítása különösen nagy jelentőséggel bír. Ilyen diktálórendszerek a világnyelvekre már léteznek, viszont kisebb és speciális nyelvi tulajdonságokkal rendelkező nyelvekre egyelőre nagyon kevés orvosi diktálószoftver látott ezidáig napvilágot, amely többek között a nyelvi sajátosságokon túl a magas fejlesztési költségeknek tudható be. Az MRBA adatbázisra alapozva mind a BME TMIT Beszédakusztikai Laboratóriuma, az MTA-SZTE Mesterséges Intelligencia Kutatócsoportja belefogott egy orvosi diktálórendszer kifejlesztésébe. A két csoport részben eltérő részfeladatokat tűzött ki maga elé (endoszkópos leletek diktálása illetve pajzsmirigy-scintigráfiás leletek diktálása) és részben eltérő technológiákat alkalmaznak, de természetesen eredményeiket folyamatosan egyeztetik, ami lehetővé teszi a tapasztalatok kicserélését és a technológiák összehasonlítását.

3 Endoszkópos leletek diktálása

Az endoszkópiai leletek gépi beszédfelismerésére és karakteres lejegyzésére képes rendszert a BME TMIT Beszédakusztikai Laboratóriuma készítette el. A laboratóriumban kifejlesztésre került egy Windows XP alatt működő beszédfelismerő fejlesztői környezet, amely alkalmas különböző középszótáras 1000-10000 szavas szövegek betanítására és felismerésére. A felismerő a statisztikai alapon működő HMM akusztikai fonémamodellekkel [2], valamint a statisztikai alapú bi-gram nyelvi modellel működik, nemlineáris simítást használva [3]. Az akusztikai modelleket az MRBA beszédadatbázissal tanítottuk. A nyelvi betanításhoz a budapesti SOTE II. sz. Belgyógyászati Klinikájától (2700 lelet) és a szegedi Orvostudományi Egyetemről (6365

lelet) gyűjtött korábbi leletanyag korpuszt használtuk. Ezen szövegkorpusz alapján elkészítettük el a teljes szóalakszótárat, amely 14331 szót tartalmaz, a kiejtési szótárat és ezek téma szerint osztott kisebb szótárait, valamint a korpusz alapján morfémaszótárat is készítettünk, amelynek nagysága 6824 morfémaelem.
A felismerő optimális működését az akusztikai [4] és nyelvi modellek változtatásával állítottuk be. Lényegében a nyelvi modellhez n-gram modelleket használtunk, de az egyik megoldásban a hagyományos szóalakok az alkotó elemek, a másik megoldásban viszont a morfémák.
Külön súlyt fektettünk a valós idejű felismerés elérésére: a dinamikus címzésen és az akusztikai modellek indirekt megközelítésén túl memóriaelérési optimalizáció, valamint nyalábolt keresésnél (Beam Search) változó terű nyaláb alkalmazásával.

4 Pajzsmirigy-scintigráfiás leletek diktálása

Szegeden kifejlesztettünk egy magyar nyelv automatikus felismerésére alkalmas magmodult, amelyre különböző, speciális feladatokhoz igazított diktálórendszerek építhetőek. A magmodul tartalmazza az un. akusztikai modellt, amely alkalmas a magyar nyelv beszédhangkészletének felismerésére és reprezentatív módon történő modellezésére. A modell felépítésére két egymástól relevánsan eltérő megközelítést alkalmaztunk. Az egyik a beszédfelismerésben közismert és gyakran alkalmazott Rejtett Markov Modell, a másik pedig a Szegeden kifejlesztett újszerű sztochasztikus szegmentális megközelítés. Mindkét modell betanításához és teszteléséhez az MRBA adatbázist használtuk fel.

A rendszerhez jelenleg egy olyan nyelvi modellt fejlesztünk, amely pajzsmirigy-scintigráfiás leletek diktálását teszi lehetővé. A nyelvi modellt 9231 írott pajzsmirigy lelet és több mint 2500 szóalak alapján építettünk fel. A nyelvi modellezésre többféle technológiát kipróbáltunk. A legegyszerűbb ezek közül az ún. szó N-gram modell. Ez megadja, hogy milyen valószínű egy adott szó az N-1 darab előtte álló szó ismeretében. Az N-gram modell kiszámításakor az ún. előrehozott N-gramm kiértékelési technológiát használjuk, amelynek segítségével a keresés során a hipotézisek száma lecsökkenthető, így a felismerés gyorsabbá tehető.

A magyar nyelv szabad szórendűsége miatt az N-gram technika nem olyan hatásos, mint pl. az angol esetében, ezért a hosszú távú kapcsolatok leírásához más modellekre is szükség van. Egy ilyen lehetőség az MSD-kód (morfoszintaktikai kód) alapú szabályok alkalmazása. Az MSD kódos leírásnál a szavak jelentése eltűnik, csak a szavak mondattani szerepe marad meg, így az MSD-kódon alapuló nyelvtanok segítségével modellezhető a mondatok felépítése.

Mind az MSD-kódokon alapuló nyelvtanok, mind a szó-N-grammok esetén komoly gondot okoz a memóriaigény, illetve az modellek pontos betanítása/kialakítása. Egy lehetséges megoldás az, hogy az osztályokra készítünk egy nagyobb (4-, vagy 5-gramm) és szavakra egy kisebb (2-, vagy 3-gramm) szótár alapú nyelvtant. A nyelvtannak az osztály-N-gramm része szintaktikai szabályokat, míg a szó-N-gramm része inkább szemantikai szabályokat szolgáltat. Jelenleg ez a kombinált megoldás tűnik a legígéretesebbnek nyelvi modelljeink közül.

Folyamatos beszéd felismerésekor további problémát jelent a hasonulás, ami akusztikailag megváltoztathatja a szavak végét vagy elejét. A hasonulás kezelését bonyolítja az is, hogy nem tudjuk, hogy a beszélő tart-e szünetet a szavak között vagy

sem, ezért ezeknek a hasonulást leíró szabályoknak, mint alternatíváknak kell megjelenniük a nyelvtanban. A hasonulást leíró szabályok összetettsége és nagy száma miatt a felismerés során a hasonulás megfelelő sebességgel való kezelése speciális problémaként jelentkezik.

Jelenleg a rendszerünk 95% körüli szó-szintű találati pontosság elérésére képes (a konkrét értékek természetesen függenek a teszt-adatbázistól és a használt nyelvi modelltől).

Bibliográfia

1. Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben, II. Magyar Számítógépes Nyelvészet Konferencia, 2004.
2. Claudio Becchetti, Lucio Prina Ricotti. Speech Recogntion,Theory and C++ implementation. Fondazione Ugo Bordoni, Rome, 1999. ISBN 0-471-97730-6
3. Ney, H., Essen, U., Kneser, R. On Structuring Probabilistic Dependencies in Stochastic Language Modeling. Computer Speech and Language, 8:1-38. oldal
4. Velkei Szabolcs, Vicsi Klára: Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából, MSZNY 2004.
5. Kocsor, A., Bánhalmi, A., Paczolay, D., Csirik, J., Pávics, L.: The OASIS Speech Recognition System for Dictating Medical Reports, Annual Congress of the European Association of Nuclear Medicine, EANM'05, 15-19 October, Istanbul, 2005.
6. Kocsor, A., Bánhalmi, A., Paczolay, D.: Informatikai és matematikai módszerek egy pajzsmirigy scintigráfiás leletek diktálására alkalmas rendszerben, IV. Alkalmazott Informatika Konferencia, AIK 2005, Május 27, Kaposvár, 2005.

Akusztikus fonetikai adatbázis-kezelő nyelvészeknek és nyelvészhallgatóknak

Zsigri Gyula[1], Paczolay Dénes[2], Sejtes Györgyi[1] és Kocsor András[2]

[1]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
H-6722 Szeged, Egyetem u. 2.
{zsigri, sejtes}@hung.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
H-6720 Szeged, Aradi vértanúk tere 1.
{pdenes, kocsor}@inf.u-szeged.hu

Kivonat: Poszterünk bemutatja, hogy egy fonetikailag feldolgozott beszédadatbázis információinak strukturált kinyerésére alkalmas program, hogyan segítheti újszerű fonetikai ismeretek kidolgozását, továbbá nyelvészhallgatók fonetikaoktatásának támogatását.

1 Bevezetés

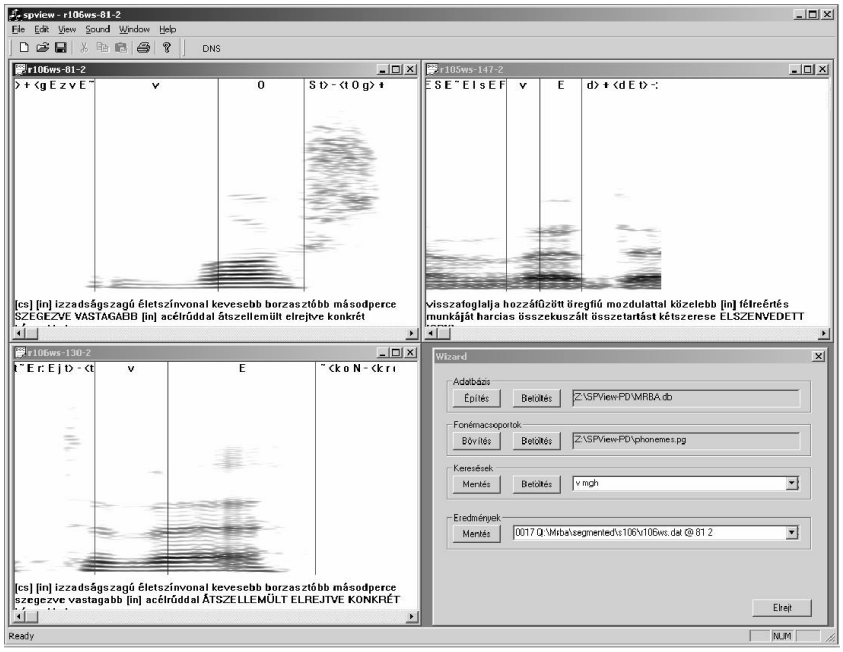
A nyelvészek többsége a legutóbbi időkig csak másodkézből jutott hozzá akusztikus fonetikai adatokhoz. Olyan műszerekhez, amelyekkel a felvett hanganyagot hangszínképpé lehetett alakítani, egészen mostanáig csak a fonetikai laboratóriumokban dolgozó eszközfonetikusok férhettek hozzá. Ennek máig is ható következménye, hogy a fonológusok a legtöbbször csak artikulációs mozzanatokból absztrahált megkülönböztető jegyeket (vagy újabban elemeket) használnak a munkáikban [2]. Vannak azonban olyan fonológiai kijelentések, amelyeket akusztikai adatokkal erősíteni vagy gyengíteni lehet. Ilyen például az a nagyon meggyőző, de mérésekkel eddig még alá nem támasztott hipotézis, hogy a /v/ fonémának a szótag eleji allofónja zengőhang, a szótag végi allofónja viszont zörejhang. Ha ez igaz, akkor jól magyarázza azt, hogy miért csak a szótag végi /v/ vesz részt a zöngésségi hasonulásban (*hí*/v/*ta* → *hí*[f]*ta,* de *á*[t]*visz, *á*[d]*visz*).

2 Akusztikai adatbázis-kezelés az Spview szoftverrel

A Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportjában kifejlesztett, Spview munkanevű akusztikai adatbázis-kezelő program segítségével könnyen, gyorsan és nagy számban előkereshetjük a munkánkhoz szükséges hangkapcsolatok hangszínképét (ld. 1. ábra). Az előbbi példánál maradva, a szótag eleji /v/-re v + magánhangzó kapcsolatként kereshetünk, a szó végi /v/-re pedig v + szavak közötti szünetként.

A program jelenleg a Magyar Referencia Beszédadatbázist (MRBA) használja fonetikai tudástárként [3]. Az adatbázis tartalmaz 345 db olyan wav formátumú hangfájlt, melyekhez tartozik fonetikus átirat és helyesírás szerinti lejegyzés is. Az összes beszédhangpéldányok száma 99346 db. A fonetikus átirat a SAMPA kódrendszer

felhasználásával készült, annyi módosítással, hogy a zárhangok esetén külön szegmentumként jelöltük be a záralkotást, a zár tartamát és a zár felpattanását. Így keresni tudunk például a homorgán nazálisok előtti fel nem pattanó zárhangokra.



1. ábra. A SpView program működés közben.

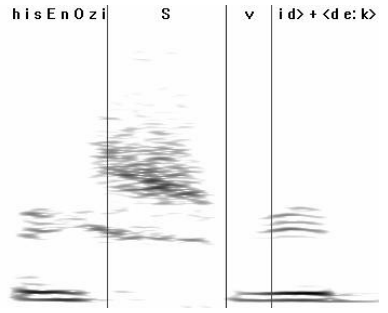
A hangkapcsolatokra SAMPA-szimbólumokkal, hangosztályokra vagy hangkapcsolatokra utaló, előre definiált rövidítésekkel és a gyakorlati helyesírás alapján is kereshetünk. A program egy SAMPA kiterjesztésű szöveges fájlból szerzi azt az információt, hogy mi tekintendő SAMPA-szimbólumnak, és egy .pg (= phone group) kiterjesztésű szöveges fájlból olvassa ki a hangosztályokra és hangkapcsolatokra utaló rövidítéseket, majd ezek alapján dönti el, hogy hogyan kell keresnie.

A fonológusokat már régóta foglalkoztatja az, hogy két mássalhangzó, a *h* és a *v* miért viselkedik felemásan a zöngésségi hasonulásban.

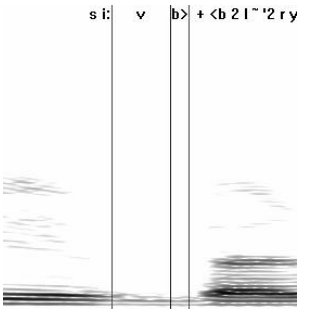
A magyar zöngésségi hasonulás szabálya így fogalmazható meg: a zöngés zörejhangok (b, d, g; v, z, zs; dz, dzs, gy) zöngétlenek lesznek (p, t, k; f, sz, s; c, cs, ty) zöngétlen zörejhangok előtt, és a zöngétlen zörejhangok zöngések lesznek zöngés zörejhangok előtt. A zengőhangok (szonoránsok) nem vesznek részt a zöngésségi hasonulásban: a *szoknya* szóban a zöngés *ny* nem zöngésíti az előtte levő zöngétlen *k*-t, és az *ajtó* szóban sem zöngétleníti a zöngétlen *t* az előtte levő zöngés *j*-t. A *v* és *h* felemásan viselkedik. A *v* zöngétlenedik (*hívta* → *hífta*), de nem zöngésít (*átvisz* ↛ *ádvisz*). A *h* pedig zöngétlenít (*egyház* → *etyház*), de nem zöngésedik (*pechből* [-x:b-]). A *h* és a *v* felemás viselkedésére különféle magyarázatok születtek [2]. Ezek között olyanok is vannak, amelyek mérésekkel ellenőrizhető állításokat tartalmaztak. A *v* felemás viselkedését [1] azzal magyarázza, hogy szerinte a *v* zengőhangok előtt

(a magánhangzók is zengőhangok) zengőhangként viselkedik, zörejhangok előtt vagy szó végén pedig zörejhangként. Abból, hogy egy hang zengőhangként viselkedik, nem feltétlenül következik, hogy az is, de érdemes megnézni, hogy az-e.

Az alábbi két ábra közül az első egy magánhangzó előtti *v* hangszínképét mutatja, a második pedig egy mássalhangzó előttiét. A magánhangzó előtt *v* hangszínképében jól látszanak a zengőhangokra jellemző formánsok, a zörejhang előttiéből viszont hiányoznak:



2. ábra. Egy magánhangzó előtti *v*.

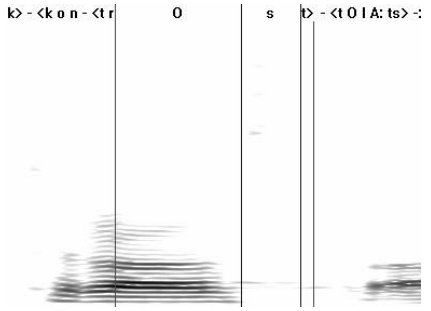


3. ábra. Egy mássalhangzó előtti *v*.

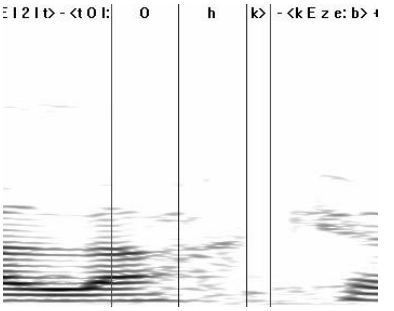
E szerint a két ábra szerint a *v*-nek nemcsak a viselkedése felemás, hanem a realizációi is mások.

Arra, hogy a *h* miért nem zöngésedik zöngés zörejhangok előtt (a *h*-nak a zengőhangok közötti zöngésedése, pl. *lehet, pelyhes,* más folyamat), tipológiai és fonotaktikai magyarázat is van. A tipológiai magyarázat azon alapul, hogy a világ nyelveiben viszonylag ritka a γ (a veláris [x]-nak a zöngés párja), nincs a magyarban sem, és ilyen ritka, nemkívánatos hangot a zöngésségi hasonulás sem hozhat létre [2]. A fonotaktikai magyarázat szerint ugyanazért nem zöngésedik a *h* a *potrohból* szóban, amiért az [fsk] kapcsolat sem zöngésedik a *Szverdlovszkban* szóban, vagy amiért a [θ] sem zöngésedik a „Jártál-e már a dél-angliai Bathban?" mondatban (Zsigri 1998). A *Bath* végén olyan mássalhangzó van, ami a magyarban nincs, a *Szverdlovszk* végén pedig úgy követik egymás a mássalhangzók, ahogy a magyarban sohasem. Az ilyen, idegen hangra vagy a magyarban fonotaktikailag rossz formájú hangkapcsolatra végződő szavak idegen testként épülnek be a szövegbe, mint a sejtekbe a zárványok. Formájuk független a környezetüktől, ennélfogva nem is vesznek részt a környezetfüggő folyamatokban. A szótag végi *h* nem idegen ugyan, de úgy viselkedik, mintha az lenne. Fonotaktikailag ugyanúgy rossz formájú, mint a szó végi [fsk]. A tipológiai magyarázat hívei szerint ez nem igaz, szerintük a szótag végi *h* ugyanolyan jó, mint bármilyen szótag végi mássalhangzó. Nem tartják komoly érvnek azt sem, hogy a *technika* szót [x] helyett sokan [k]-val ejtik, mondván, hogy ez egyedi eset, annak ellenére, hogy a *xilofon* eleji [ks] kapcsolat [s]-re egyszerűsödését viszont döntő érvként ismerik el amellett, hogy a szó eleji [ks] kapcsolat a magyarban rossz formájú. A *h* azonban nemcsak nem zöngésedik, hanem a mássalhangzók melletti rövidülésnek is ellenáll. A magyarban, ha egy hosszú mássalhangzót szünet nélkül egy másik mássalhangzó követ, a hosszú mássalhangzó lerövidül: *sakk* [k:], de *sakktábla* [kt]. A szótag végi *h* ilyenkor is hosszú maradhat: *pechből, potrohból*. Az

adatbázisban a *pechből, potrohból* szavak nem szerepelnek, de az alábbi két ábrából is jól látszik, hogy a szótag végi [h] frikatív zöreje jóval korábban kezdődik, és tovább tart, mint a zörejhang előtti [s] élesebb, de rövidebb zöreje:



4. ábra zörejhang előtti [s]



5. ábra szótag végi [h]

A *h* rövidülésének az elmaradása a fonotaktikai alapú magyarázatot erősíti.

A Spiew program nemcsak a nyelvészek munkáját könnyíti meg, hanem a fonetikatanításban is jól használható. A programot használó diákok passzív befogadás helyett interaktív módon ismerkedhetnek meg a beszédhangok hangszínképével. Az SPView programmal hangosztályokat lehet definiálni, és hangkapcsolatokra lehet keresni. A találatokat az SPView-val egyenként is meg lehet jeleníttetni, és menteni is lehet a találatokat. Mentéskor a program minden egyes találatnak lementi a hangszínképét, aláírja a keresett hangkapcsolatot tartalmazó szövegrészletet helyesírási formában és SAMPA-ba átiratban is.

Bibliográfia

1. Jakobson, Roman. Die Verteilung der stimmhaften und stimmlosen Geräuschlaute im Russischen. In Margarete Woltner & Herbert Bräuer (eds.) Festschrift für Max Vasmer zum 70. Geburstag. Berlin: Harrasowitz. pp. 199–202. 1956
2. Siptár Péter, Törkenczy Miklós: The phonology of Hungarian. Oxford: Oxford University Press, 2000.
3. Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben, II. Magyar Számítógépes Nyelvészet Konferencia, 2004.

IX. Laptopos bemutatók

„Szemfüles” - Hallási megkülönböztető képesség fejlesztő szoftver hallássérült gyerekek részére

Magyar Viktor [1], Sikné dr. Lányi Cecília[1], dr. Váry Ágnes[2]

[1] Veszprémi Egyetem,
Műszaki Informatikai Kar, Képfeldolgozás és Neuroszámítógépek Tanszék
magyarviktor@gmail.com
[2] Veszprémi Egyetem,
Műszaki Informatikai Kar, Képfeldolgozás és Neuroszámítógépek Tanszék
lanyi@almos.vein.hu
[3] Dr. Török Béla Óvoda, Általános Iskola, Speciális Szakiskola, Egységes Gyógypedagógiai Módszertani Intézmény, Diákotthon és Gyermekotthon
varyagnes@yahoo.com

Kivonat: A „Szemfüles” szoftver hallássérült gyermekek számára készült játékos készségfejlesztő multimédiás program. Segítségével a felhasználók a magasfrekvenciás, ún. sziszegő hangok („sz”, „z” , „c” „zs”, „s”, „cs”) megkülönböztetését gyakorolhatják. A programban használt szókincsanyagot a mindennapi élet számtalan területéről összegyűjtve úgy állítottuk össze, hogy azok a hallássérült gyermekek által leginkább összekevert hangokat tartalmazzák. A szoftver szavak illusztrálására több száz képet tartalmaz. Így a hallási megkülönböztető képesség fejlesztése mellett a diákok szókincsének bővítése is lehetővé válik. Az elkészült programot jelenleg a Dr. Török Béla Óvoda, Általános Iskola, Speciális Szakiskola, Módszertani Intézmény, Diákotthon és Gyermekotthon általános iskolás korú hallássérült tanulói használják, segítségével hatékony tanulás, készségfejlesztés folytatható.

1. Bevezetés

A hallássérülés közvetlen következménye lehet a nyelv, a beszéd elsajátításának zavara vagy a beszéd hiánya. A beszédzavarok különböző beszédfunkciók területén jelentkezhetnek úgymint az artikulációban, a választékos szókincs kialakulásának hiányában vagy grammatikai hibák előfordulásában.[1-2] Az általunk kifejlesztett szoftver elsősorban az első két terület által képviselt nyelvi nehézségekkel küszködő általános iskolás korú gyermekek számára készült. A programban használt szókincsanyagot a mindennapi élet számtalan területéről összegyűjtve úgy állítottam össze, hogy azok a hallássérült gyermekek által leginkább összekevert hangokat tartalmazzák. Így a szókincsfejlesztés mellett lehetővé válik, a magasfrekvenciás hangok megkülönböztető készségének fejlesztése. A szoftver főmenüjéből a diákok öt különböző feladattípus közül választhatnak. Az elkészült programot jelenleg a Dr. Török Béla Óvoda, Általános Iskola, Speciális Szakiskola, Módszertani Intézmény, Diákotthon és

Gyermekotthon általános iskolás korú hallássérült tanulói, segítségével hatékony tanulás, készségfejlesztés folytatható.

2. Háttér

A speciálpedagógiai elméleti háttér lényege, hogy a hallássérült gyermekek beszédfejlesztésének alapfeltétele a hallási figyelem, hallási emlékezet, hallási differenciáló- és diszkrimináló képesség fejlesztése. E képességek nélkül elképzelhetetlen a beszédhangok helyes ejtésének kimunkálása, a későbbiekben pedig a helyesírás, az olvasás nehézségektől mentes fejlődése. Ma már köztudott, hogy ezen auditív képességek fejlesztése nélkül a halló gyermekek számára is igen nagy nehézséget jelent az iskolakezdés. Tehát a program nemcsak hallássérültek, de halló gyermekek logopédiai, diszlexia-prevenciós terápiája során is remekül alkalmazható.
A szoftver a" sziszegő hangok"(„sz", „z", „c", „s", „zs", „cs") köré csoportosítja a feladatokat, de természetesen erre az analógiára bármely más beszédhang gyakorlását célzó feladatok összeállíthatók. Hogy miért pont ezekre a hangokra esett a választás, annak a következő a magyarázata: Ezek azok a beszédhangok, amelyek formánsai a legmagasabb frekvenciákon, 3000 és 8000 Hz között hallhatóak. Vagyis a hallássérült gyermek (főleg az idegi eredetű hallászcsökkenéssel élők) számára – a legoptimálisabban beállított hallókészülékes erősítés mellett is- lehetetlen vagy torz az érzékelés.[3] Nem véletlen az sem, hogy a halló kisgyermekek körében is ezeknek a hangoknak a megkülönböztetésével illetve helyes ejtésével van a legtöbb probléma.

3. A szoftver bemutatása

A programot elindítva először a névbekérés ablak látható. Itt a felhasználónak meg kell adnia a nevét. Ennek később az eredményrögzítés funkció tárgyalásakor lesz jelentősége. A név megadása után a főmenübe jutunk, amely a 1. ábrán látható. Itt lehet választani az öt feladatlap közül. A bal alsó sarokban található a kilépés gomb. A bal felső sarokban két hangerőszabályzó található. Mivel a szoftver hallássérült gyerekeknek készül, mindenképp szem előtt tartottuk, hogy a hanganyagot a megfelelő hangerőn hallgathassák. A hangjegy szimbólumra kattintva ki- illetve bekapcsolhatjuk az alapzenét. Ezen ikon alatt található csúszka segítségével pedig a zene hangereje állítható be.
A szoftver rendelkezik hanganyaggal is. Amikor a feladatok megoldása során a felhasználó rákattint valamelyik képre, akkor a beállításoktól függően a számítógép kimondja a képen lévő tárgynak a nevét. A mikrofonra kattintva ez a funkció kapcsolható be vagy ki. Ezen szimbólum alatt található csúszka segítségével pedig a szavak bemondásának hangereje állítható.
Miután a főmenüben található feladatlapok közül a felhasználó, vagy a vele foglalkozó szakember kiválasztotta, hogy melyik feladattal akar foglalkozni, újabb választási lehetőség következik: el kell dönteni, hogy melyik hangot vagy hangokat akarja gyakorolni. A megfelelő hang hívójelére kattintva egy pipa jelenik meg, ami azt jelenti, hogy ez a betű kiválasztásra került. A kiválasztott hangra való újbóli

kattintással a kijelölés megszüntethető. A gyakorolandó hangok száma függ az egyes feladatlapoktól is.

1. ábra: A szoftver főmenüje

4. Feladatlapok

A Szemfüles szoftver öt különböző feladatlapot tartalmaz, mely külső szemlélő számára első ránézésre egyszerű játékprogramnak tűnhet, azonban ezek a feladatlapok úgy vannak összeállítva és elkészítve, hogy azok elősegítsék a magasfrekvenciás hangok megkülönböztető képességének fejlesztését.

Az első feladatlapon a felhasználónak el kell döntenie, hogy az adott szóban az előzőleg gyakorlásra beállított két hang közül melyik hang szerepel. Választ a felső sorban található a két hang hívójelét ábrázoló képek közül a megfelelőre kattintva adhat. A programban való könnyebb navigálást elősegítendő, a képernyő bal oldalán navigációs gombokat találunk. Ezek segítségével a felhasználó a feladat végrehajtását megszakíthatja, visszatérhet a betűválasztáshoz más gyakorolandó betűket választhat, vagy visszatérhet a főmenübe is, ahonnan másik feladatba kezdhet. Ha a feladat megoldása sikeres volt, lehetőség van a feladatot újra megoldani más, véletlenszerűen kisorsolt képekkel.

A második feladatlap egy speciális, ún. Kép-szókép memóriajáték. Ez annyiban különbözik a hagyományos értelemben vett memóriajátéktól, hogy míg abban két azonos ábrát kell összepárosítani, addig ebben a típusban, mint ahogy az a nevéből is adódik egy képet, és a képen látható tárgy nevét kell egyidejűleg felfordítani

A harmadik feladatban egy mondat alapján kell kitalálni, hogy a képernyő alján két sorban megjelenő képek közül a mondatban melyikre is gondoltunk. A feladatban szereplő mondatok szöveges fájlban kerülnek tárolásra, így azok mélyebb programozási ismeretek nélkül is testre szabhatók, módosíthatók.

A negyedik feladat megoldása során a lent megjelenő képek közül kell választani, hogy az a gyakorolt hang a szóban hol helyezkedik el. Ezután a képernyő tetején látható kígyó megfelelő részén elhelyezett hívójelekre (az elején, a közepén, vagy a végén) kattintva lehet válaszolni. Helyes válasz esetén a kijelölt kép mérete csökken,

kerete sárgára változik, majd a kép alatt megjelenik a képen látható tárgy neve. Rossz válasz esetén a kijelölés megszűnik. Ezután a felhasználónak újra kell próbálkoznia az elrontott részfeladat megoldásával.
Az ötödik felad egy speciális kakukktojás feladat. A játékhoz kisorsolt képek nevét kiejtve mindegyikben ugyanaz a „problémás hang" szerepel, kivéve a kakukktojásban.

5. Eredményrögzítés

A program működés közben az eredményeket nem értékeli, nem pontozza. Ezért igyekeztem minden jó válasz esetén valamilyen animációval mosolygós fejjel jutalmazni a felhasználókat. Annak érdekében, hogy a gyerekek teljesítményét értékelni lehessen, elláttam a programot egy eredményrögzítő funkcióval. Ez a felhasználó számára teljes egészében rejtett. A program elindításkor bekéri a felhasználó nevét, hisz valamilyen módon meg kell különböztetni a programot használókat. A feladatok megoldása során szöveges állományban rögzítésre kerül a feladatot megoldó személy és a feladatlap neve. A szoftver elmenti továbbá azt is, hogy a felhasználó melyik hangokat választotta ki gyakorlásra, valamint a feladat megoldásának idejét. Ezt azért tartottam fontosnak, mert így a gyerekekkel foglalkozó szakember nyomon követheti képességeik fejlődésének mértékét. A szövegfájlban tárolásra kerül az is, hogy a megoldás menete során melyik képre kattintott a felhasználó, mit ábrázolt a kép, és a részfeladatra milyen választ adott, valamint mennyit gondolkodott válaszadás előtt. Ha a megoldás rossz volt, akkor az annak megfelelő sor végén látható az, hogy milyen választ adott a jó helyett. Így az eredményekből kiderül, hogy az adott diáknak melyik hanggal van a legtöbb problémája, melyik az, amelyiket legnehezebben hallja, vagy melyik az a kettő, amelyeket a legnehezebben tudja egymástól megkülönböztetni

```
A tanuló neve: Horváth Endre
Mondd ki a szavakat, melyik hangot hallod?
Feladat: sz és cs hangok megkülönböztetése.
A feladat megoldásának időpontja: 2005. október 26.
Feladat: csibe Gondolkodási idő: 1 sec. Válasz: cs -- Jó válasz
Feladat: gyümölcs Gondolkodási idő: 2 sec. Válasz: sz - Rossz
                                                    válasz (cs helyett sz)
Feladat: gyümölcs Gondolkodási idő: 3 sec. Válasz: cs -- Jó válasz
```

6. Bővíthetőség

A programot és a hozzátartozó könyvtárszerkezetet úgy terveztem meg, hogy a szoftver által használt képanyag később dinamikusan bővíthető, testreszabható, változtatható legyen. A szoftver a feladatlapokhoz véletlenszerűen választja ki a képeket. Azon feladatok esetében, amelyekhez nem kell kiválasztani a hangokat, először kisorsol egy betűt. Ezt követően véletlenszerűen kiválasztásra kerül az, hogy

az adott betű a feladatként kijelölendő szóban hol szerepeljen. A következő lépésben beolvassa a program a megfelelő mappából a képek.txt szöveges fájlt. Ezen fájlok formátuma kötött Az első sorban az adott hanggal kezdődő szavak vannak, a harmadik sorban erre a hangra végződő, míg a második sorban ezt a hangot a belsejükben tartalmazó szavak találhatók. Attól függően, hogy az előző lépésben az adott mássalhangzó helyét kiválasztó sorsolás milyen eredményt adott, az algoritmus megvizsgálja, hogy a szövegfájl megfelelő sorában hány szó található, majd ezen szavak közül kerül kisorsolásra az, amelyik a feladatlapra felkerül.
Mivel a képek nevei szöveges állományban kerülnek tárolásra, és a képek sorsolása is ezen fájlok alapján történik, lehetővé válik, hogy a szövegfájlok módosításával változtatható legyen a szoftver által használt képanyag. Ezen módosítások mélyebb programozási ismeretek nélkül elvégezhetők, anélkül, hogy a program forráskódját módosítani kelljen. Ezen lehetőség segítségével a gyerekekkel foglalkozó szakemberek személyre szabott képanyaggal ellátott feladatlapokat készíthetnek. A harmadik feladatlap mondatai szintén szöveges állományból kerülnek kisorsolásra. Ebből adódóan a pedagógus saját feladatait is beilleszthetti az eredeti feladatok közé.

Összegzés

A Szemfüles készségfejlesztő szoftver elsősorban idegi eredetű hallászcsökkenéssel élő általános iskolás korú diákok számára készült, a magasfrekvenciás sziszegő hangok megkülönböztetésének gyakoroltatására. Segítségével játékos keretek közt hatékony készségfejlesztés végezhető. Mélyebb informatikai ismeretek nélkül bővíthető és testreszabható a program képanyaga. A felhasználók eredményeit szövegfájlba rögzíti, így hatékony eszköz lehet a logopédusok kezében, akik diákjaik teljesítményét ezúton is nyomon tudják követni.

Bibliográfia

1 Pataki László: Hallássérülés – Hallási fogyatékosság, Gyógypedagógiai alapismeretek, ELTE tankönyv, Szerzői kiadás 183-195. oldal
2 Farkas Miklós, Perlusz Andrea: A hallássérült gyermekek óvodai és iskolai nevelése és oktatása, Gyógypedagógiai alapismeretek, ELTE tankönyv, Szerzői kiadás 507-533. oldal
3 Dr. Bodó Gabriella, Dr. Vízkelety Tibor: Halláscsökkenés és kezelési lehetőségei http://www.medlist.com/HIPPOCRATES/VI/1/055main.htm

Az OpenOffice.org irodai program nyelvi eszközei

Németh László

BME – Média Oktató és Kutató Központ
1111 Budapest, Stoczek u. 2.
Nemeth@MOKK.BME.hu

Kivonat: Az OpenOffice.org irodai programcsomag Lingucomponent modulja és az OpenOffice.org natív nyelvi fejlesztései nyelvi eszközöket, nagyobbrészt helyesírási szótárakat biztosítanak jelenleg mintegy 70-80 nyelvhez. Az OpenOffice.org irodai programcsomag képességeinek növekedésével, például az új Thesaurus komponens vagy a Hunmorph morfológiai elemző beépítésével a helyesírási szótárak mellett a szinonimaszótárak és morfológiai elemzésre is használható erőforrások fejlesztése is elkezdődött. Az OpenOffice.org növekvő népszerűségére, a nyílt forráskódú fejlesztési modellre, valamint a nyelvi erőforrások elkészítésének és karbantartásának automatizálására alapozva lehetővé válhat a nyelvi eszközök és erőforrások folyamatos fejlesztése.

1. Bevezetés

Az OpenOffice.org [8] a legjelentősebb nyílt forráskódú irodai programcsomag. Az OpenOffice.org fejlesztése sok különböző projektre különül el. A Lingucomponent projekt foglalkozik a számítógépes szövegbevitelt és szövegfeldolgozást segítő nyelvtechnológiai fejlesztésekkel. A Lingucomponent projekt célkitűzése, hogy versenyképes nyelvi eszközöket (szó- és mondatszintű helyesírás-ellenőrző, elválasztó program, szinonimaszótár, stb.) biztosítson minél több nyelvhez [1]. A feladat tehát kettős: az alkalmazások és a nyelvi erőforrások fejlesztését is jelenti. Mintegy 70-80 nyelvhez érhető el OpenOffice.org nyelvi erőforrás. Ezek az erőforrások legalább a helyesírás-ellenőrzéshez használt egynyelvű szótárat artalmazzák, legtöbbször a nyelv morfológiáját tükröző *affixumtömörítéssel* [6].

A magyar nyelv támogatása lényegesen javulni fog az OpenOffice.org 2.0-s változatában, mivel az irodai programcsomag helyesírás-ellenőrzőjét felváltja a magyar fejlesztésű Hunspell [4]. A Hunspell támogatja a bonyolult morfológiát és összetettszó-kezelést, valamint az Unicode karakterkódolást is, így nemcsak a magyar, hanem sok más nyelv kezelését is lehetővé teszi. A Hunspell-lel tucatnyi ázsiai és afrikai Aspell erőforrás válik elérhetővé az OpenOffice.org számára. Már az OpenOffice.org integráció befejeződése előtt elkezdődött az új Hunspell erőforrások fejlesztése (például a nepáli, moszi, akan nyelvekhez), illetve a meglévő erőforrások javítása (a finn a morfológia, a német és az afrikaans az összetettszó-kezelés terén használja ki a Hunspell képességeit).

2. OpenOffice.org

Az OpenOffice.org irodai programcsomag a Microsoft Office kiváló ingyenes alternatívájaként ismert hazánkban. Az OpenOffice.org kódbázisára épül számos egyéb kereskedelmi irodai programcsomag is, mint a Sun StarOffice vagy az EuroOffice (korábban MagyarOffice), illetve része az IBM Workplace programcsomagjának is.

Az OpenOffice.org jelentősége azonban túlmutat az alternatíva és a kódbázis szerepen [5]: a platformfüggetlenségnek, a nyitott szabványok támogatásának, a nyílt forráskódnak és fejlesztési modellnek, továbbá a natív nyelvi projektek támogatásának köszönhetően jóval szélesebb közönséget céloz meg, mint az egyes kereskedelmi programváltozatok.

3. Natív nyelvi projektek

Körülbelül 60 natív nyelvi projektje van az OpenOffice.org-nak, pár ezer közreműködővel. További negyven natív nyelvi projekt létrejöttére számít a közösség egy éven belül.

4. Lingucomponent projekt

A Lingucomponent projekthez tartozó jelenlegi OpenOffice.org alkalmazások: elválasztó komponens (AltLinux), szószintű helyesírás-ellenőrző (MySpell, a 2.0.1-es változattól Hunspell) és szinonimaszótár (OpenOffice.org Thesaurus). Az elválasztó komponens D. E. Knuth nevezetes szedőprogramja, a TeX elválasztási algoritmusán alapul. A Hunspell program az Ispell helyesírás-ellenőrző családba tartozik, így képes az Ispell, Aspell és MySpell szótárak használatára.

Az alkalmazásokat érintő jelentősebb tervezett fejlesztések: morfológiai elemző illesztése az elválasztó modulhoz és a szinonimaszótárhoz, a helyesírás-ellenőrző összetettszó-kezelésének általánosítása és a kiejtési hasonlóságon alapuló javaslattevés implementálása, mondatszintű ellenőrzés megvalósítása.

Több más fejlesztés is kapcsolódik közvetve a Lingucomponent projekthez. Ilyen az OpenThesaurus projekt [7], amely egy nyílt keretrendszer Wordnettel szinkronizált szinonimaszótárak fejlesztéséhez (már 5-6 nyelv esetében használják OpenOffice.org erőforrások létrehozására), valamint az An Crúbadán projekt [10], amely már 144 nyelv webkorpuszát gyűjti folyamatosan az Internetről. A webkorpuszok megfelelő kiindulási alapot jelentenek az OpenOffice.org nyelvi erőforrások készítéséhez.

Míg nyelvtechnológiai vonatkozásban a hiányzó alkalmazások és képességek megtervezése és kifejlesztése a feladat, a nyelvi erőforrások esetében a natív nyelvi projektek munkájának összehangolása, minőségbiztosítása, a nyelvészeti munkát egységes keretbe foglaló fejlesztői környezet kialakítása a cél, ami jelentősen megkönnyíti a több száz nyelvierőforrás-fejlesztő munkáját.

A Lingucomponent projekt eredményei nem csak az irodai programcsomag felhasználói, hanem a kutatói közösség számára is fontosak, így a jövőben számítunk az alapkutatásban érintettek (nyelvészek, számítógépes nyelvészek, szoftverergonómusok) fokozott részvételére.

Magyar vonatkozású fejlesztések

Bár az elválasztó modul morfológiailag nem elemzi még a szavakat, a Huhyphn elválasztási szabálygyűjtemény [2] tartalmazza az eddigi legteljesebb magyar elválasztási kivételszótárat.

A magyar Hunspell erőforrás [3] új, unicode-os változata régi hiányosságot pótol: a helyesírási szabályzat előírja az idegen ékezetes latin betűk használatát az idegen szavakban, de ezt eddig nem támogatták a magyar helyesírás-ellenőrzők. Az Unicode és a bonyolult szóösszetételek (például földrajzi nevek) támogatása fontos előrelépést jelent a különféle szakszövegek gondozásában.

A magyar morfológiai elemző [12] a toldalékolt szavak kezeléséhez és a mondatszintű elemzéshez nélkülözhetetlen. Az OpenOffice.org 2.0.1-es változata a magyar morfológiai elemzőt tövezésre, és a szinonimák toldalékolt alakjainak előállítására fogja használni.

Az OpenOffice.org magyar szinonimaszótárának OpenThesaurus alapú fejlesztése elkezdődött a tervezés szintjén. A közösségi fejlesztési modell csak aktív résztvevők megléte esetén működőképes, de a német OpenThesaurus [9] és a magyar Wikipédia [13] sikere biztató a magyar OpenThesaurus jövőjére nézve is.

Bibliográfia

1. Lingucomponent projekt: http://lingucomponent.openoffice.org
2. Nagy Bence: Huhyphn – Magyar elválasztás TeX-hez, Scribushoz, OpenOffice.org-hoz, 2003, http://www.tug.org/tex-archive/language/hungarian/**huhyphn**.pdf
3. Németh László: Magyar Ispell – Válasz a Helyes-e?-re, IV. GNU/Linux szakmai konferencia, LME, Budapest, 2002, 99–107. o., http://mek.oszk.hu/01200/01240/
4. Németh László: A Szószablya fejlesztés, 2003, V. GNU/Linux szakmai konferencia, LME, Budapest, 2003, 103–108. o., http://mek.oszk.hu/02200/02230/
5. Miguel de Icaza: The Global Importance of OpenOffice.org (nyitóbeszéd és prezentáció), OpenOffice.org konf., Berlin, 2004, http://marketing.openoffice.org/conference/thursday.html
6. Geoff Kuennings: International Ispell: http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html
7. Daniel Naber: OpenThesaurus: Building a Thesaurus with a Web Community, 2004, http://www.openthesaurus.de/download/openthesaurus.pdf
8. OpenOffice.org: http://www.openoffice.org
9. OpenThesaurus: http://www.openthesaurus.de
10. Kevin P. Scannel: An Crúbadán, Corpus building for minority languages http://borel.slu.edu/crubadan/
12. Trón, V, Németh, L, Halácsy, P, Kornai, A, Gyepesi, G, and Varga, D. Hunmorph: open source word analysis, In: Proceeding of ACL. ACL. (2005)
13. Wikipédia: http://hu.openoffice.org

Automatikus zárt ë-jelölő program

Novák Attila és Endrédy István
MorphoLogic Kft. 1126 Budapest Orbánhegyi út 5.,
{novak,endredy}@morphologic.hu

A magyar helyesírás nem jelöli a nyílt *e* és az egyes nyelvváltozatokban még élő félzárt *ë* fonéma különbségét, mivel az érvényes helyesírási norma kialakításának alapja egy olyan nyelvjárás volt, amelyben az *ë* fonéma már nem létezett. A mai budapesti köznyelvben sem szerepel ez a fonéma, és a magyar beszélők többségének nyelvi kompetenciájának nem része a nyílt *e* és a félzárt *ë* megkülönböztetése sem a beszédprodukció, sem a beszédértés szintjén.

Az *ë*-ző nyelvváltozatokat anyanyelvként használó magyar beszélők egy része szükségét érzi, hogy ezt a fonémát írásban is megkülönböztesse. Az *ë*-k írásbeli jelölését szorgalmazó legismertebb személyiség Kodály Zoltán volt. Jelenleg is létezik egy alapítvány[85], amely *ë*-jelölt szövegeket, illetve kiejtési szótárakat és szójegyzékeket ad ki. Ők kérték fel a MorphoLogicot arra, hogy készítsünk egy olyan eszközt számukra, amellyel *ë*-jelölést nem tartalmazó szövegeket lényegében automatikusan (egy utólagos félmanuális korrektúra lehetőségével) át lehet alakítani *ë*-jelölt szövegekké.

Az eszköz alapját egy olyan szóalaktani leírás képezi, amelyet a standard magyar köznyelv morfológiaielemző-adatbázisának kiegészítésével hoztunk létre. A magyar magánhangzó-rendszer ismeretében a toldalékok rendszerének megfelelő módosítását MorphoLogic Humor elemzőprogramjához készített nyelviadatbázis-kezelő keretrendszer segítségével (Novák, 2003 [1]) nem volt nehéz elvégezni. Ugyanakkor a tövek *e* hangjainak jelölését, illetve az elöl képzett harmóniájú nyitótövek azonosítását nyelvi kompetencia hiányában nem mi, hanem az alapítvány munkatársai, Buvári Márta és Mészáros András végezték.

Az *ë* fonémát is tartalmazó kibővített adatbázis alapján készített módosított szóalaktani elemzőprogram képes az *ë*-jelölt szövegek elemzésére, igény esetén készíthető helyesírás-ellenőrző is ehhez a nyelvváltozathoz. Az adatbázis további módosításával hoztuk létre azt az eszközt, amely a szabályos magyar helyesírással írt szövegeket átalakítja olyanra, amelyben jelölve van a két *e* hang közti különbség.

A program többértelmű szavak esetében a legvalószínűbb változatot választja, de a döntése minden egyes többértelmű szó esetében egy a jobb egérgomb megnyomására feltűnő kontextusmenü használatával nagyon könnyen felülbírálható. A jelöltek sorrendezése statisztikán, illetve kézzel hangolt jelöltségi sorrendeken alapul. Az alábbi három tényezőt vesszük csökkenő súlyozással figyelembe:

–ë-jelölt szövegkorpuszból nyert szóalak-gyakorisági statisztikát,
–az egyes tövekhez rendelt jelöltségi sorrendet és
–az egyes toldalékmorféma-sorozatokhoz rendelt jelöltségi sorrendet.

Az elemzések sorrendezéséhez használt statisztika az elemzőtől függetlenül változtatható, hangolható. Kontextuális tényezőket nem veszünk figyelembe a jelöltek rende-

[85] Bárczi Géza Kiejtési Alapítvány

zésénél, de így is általában nagyon kevés kézi utómunkára van szükség a szöveg végleges formára hozásához. Az utólagos kézi ellenőrzést segíti, hogy a program minden az *ë*-jelölés szempontjából többértelmű szót megjelöl, és külön jelöli a számára ismeretlen *e*-betűt tartalmazó szavakat is. A többértelmű szavak közötti választást a program olyan segédszavak megjelenítésével segíti, amelyek segítségével minden olyan magyarul tudó felhasználó is könnyen ellenőrizni tudja, hogy a gép választása az adott kontextusban helyes-e, illetve el tudja végezni az egyértelműsítést, aki az *ë*-ző nyelvváltozatot nem beszéli:

csënd (főnév) / csend (te azt)
szemetek (főnév) / szëmëtëk (főnév (birtokos alak))
illetékësëk (főnév) / illetékësek (melléknév)
finnek (olyannak (melléknév)) / finnëk (olyanok (melléknév))

Ez a vállalkozás példaértékű abból a szempontból, hogy hasonló módon esetleg más nyelvváltozatok leírására is lesz lehetőség.

Bibliográfia

1. Novák Attila.: Milyen a jó humor? Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), pp. 138–145, Szegedi Tudományegyetem, 2003.

Magyar nyelvű kérdő mondat elemző szoftver

Tikk Domonkos[1], Szidarovszky Ferenc P.[2], Kardkovács Zsolt Tivadar[1], Magyar Gábor[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
{tikk, kardkovacs, magyar}@tmit.bme.hu
[2] Szidarovszky Kft.
H-1392 Budapest, Pf. 283.
ferenc.szidarovszky@szidarovszky.com

Kivonat: Cikkünkben „A szavak hálójában" projekt keretében megvalósított kérdő mondatok elemzését végző programot ismertetjük. A projekt egyik célja egy olyan keresőszolgáltatás algoritmikus és architekturális feltételeinek megteremtése, amely lehetővé teszi, hogy természetes nyelvű magyar kérdésekkel internetes adatbázisok tartalmában az ún. mélyhálóban keressünk. Ezen cél elérésének egyik fontos része a kérdő mondatok szintaktikai és szemantikai feldolgozása. A bemutatott szoftver a nyelvtani szerkezetek felismerésén kívül minta- és szótáralapú névelem-detektáló feladatokat is elvégez, amelynek nagy jelentősége van a kérdéselemzés következő lépése, a kontextusfelismerés során.

1 Bevezetés

A projekt célkitűzése, hogy a felhasználók számára megkönnyítse az internetes adatbázisok tartalmában való keresést. Ezen adatbázisok tartalma, amelyet összességében *mélyhálónak* hívunk, sokszorosa az ún. felszíni világháló tartalmának, ráadásul a jellemzően relációs adatbázisokban tárolt információk pontosabbak, és hamarabb is frissülnek. A hagyományos keresőmotorok azonban nem tudják indexelni ezt az értékes információforrást, mivel a tartalmuk csak az adott internetes adatbázis felhasználói felületéről kezdeményezett keresések eredményeként, dinamikusan jelenik meg. A projektünk célja, hogy olyan keresőszolgáltatást nyújtson, amely a mélyhálós tartalomszolgáltatókkal együttműködve a szolgáltatott tartalmakat egy közös kereső platformon keresztül teszi elérhetővé és kereshetővé úgy, hogy a kereséseket *természetes magyar nyelvű mondatok* formájában lehessen megadni.

A projekt keretében megvalósuló prototípus-alkalmazás a jelenleg böngészővel közvetlenül el nem érhető, adatbázisban található tartalom egy részét kívánja elérhetővé tenni, amelyek a *könyv, film, labdarúgás* és *étterem* témakörébe esnek. Az alkalmazás csak olyan jellegű kérdésekre kísérel meg válaszolni, amelyre a válasz megtalálható a mélytartalmat szolgáltató partnerek adatbázisaiban. Ez természetesen

bizonyos megszorításokat jelent a kérdés típusára, jellegére és témájára vonatkozóan. Az alábbiakban vázlatosan bemutatjuk a szoftver működési elvét és funkcióit.

2 A program működése

A szoftver Windows platformra C++ nyelven íródott. A működéséhez különböző segédeszközöket (pl. morfológiai elemző), adattárakat használt fel (pl. névelem-tár).

A program bemenetként nem összetett, kérdőszóval kezdődő a magyar nyelvtan és helyesírás szabályainak megfelelő, tényszerű tartalomra vonatkozó kérdéseket elemez. Nem fogad el, illetve nem garantál jó elemzést eldöntendő, szubjektív, intencionális, kauzális kérdésekre. Az elemzés eredménye XML formátumú értelmezési *alternatívák* sorozata, valamint grafikusan megjelenített tokenizációs és elemzési fák. Ez utóbbiak láthatók cikkünk ábráin.

A felhasználónak lehetősége van bizonyos értelmezési opciók megadására, amelylyel elősegítheti a többértelmű szerkezetek egyértelműsítését. A felhasználó az alábbi értelmezési opciókkal segítheti a mondat helyes felismerését (ld. még ábrák).

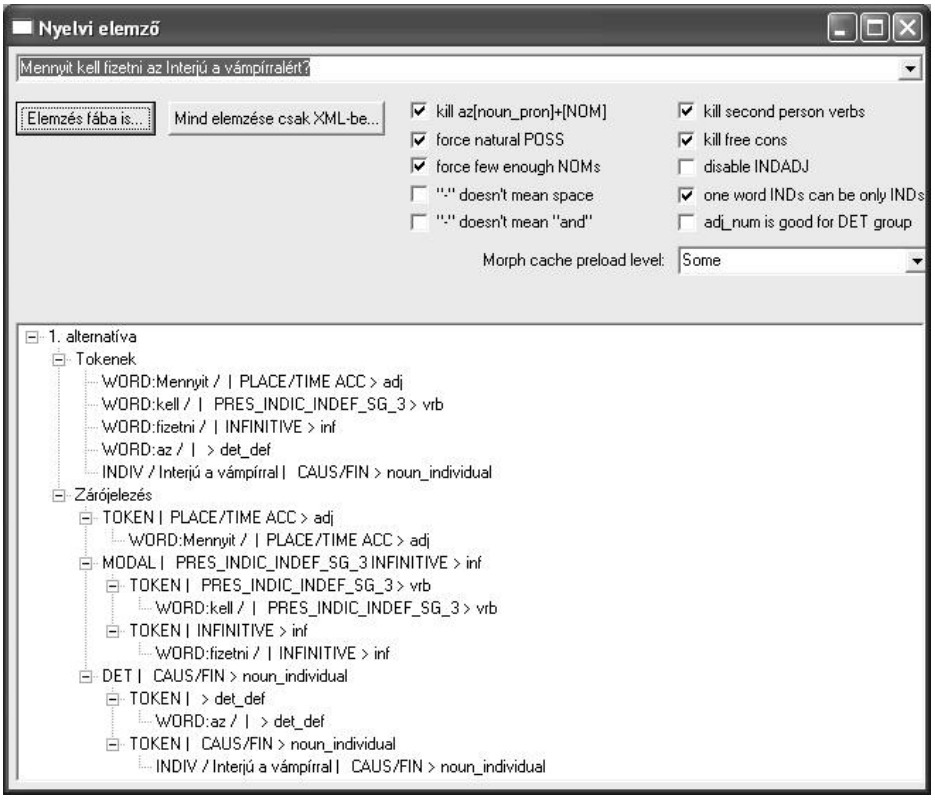
- *az* szó csak névelőt jelent (alapértelmezett); a morfológiai elemző által adott *főnév* elemzés általában hibás alternatívát generál.
- természetes birtokos sorrend preferálása (alapértelmezett); ha a birtokot közvetlenül megelőző token egyben lehetséges birtokos is, akkor más lehetséges birtokost nem keresünk.
- túl sok felső szintű alanyesetű főnév szűrése (alapértelmezett); ezzel az opcióval a 2-nél több valamilyen struktúrában nem szereplő alanyesetű főnevet tartalmazó mondatokat szűrjük ki.
- kötőjel értelmezése (2 opció); a kérdésben esetlegesen szereplő kötőjel karakterek értelmezésének megadása (szóköz; szóösszetétel; vagy „és”), amelyek hiányában az eredeti kérdésből több alternatívát készítünk.
- 2. személyű igéket tartalmazó alternatívák szűrése (alapértelmezett).
- szabad összekötőket tartalmazó alternatívák szűrése (alapértelmezett); a nem feldolgozott összekötők (és, vagy) tartalmazó mondatok eldobása.
- névelemek melléknévi szerepben (engedélyezett); *Shakespeare szonett* típusú kifejezések helyes felismerését lehetővé tevő opció.
- egy szavas névelemeket ne értelmezze csak névelemként (alapértelmezett); ha kikapcsoljuk, akkor a névelemként is szereplő közszavakat kétféleképpen tokenizálja.
- számnév elfogadása névelőként; az *egy* vagy más számnevek helyes értelmezése adható meg.

Az alternatívák elemzése lépések egymás utáni végrehajtásából áll. A lépések végrehajtása során keletkezhetnek szabálytalan alternatívák is, amelyeket a lépés végrehajtása után eldobunk.

2.1 Tokenizálás, névelemek keresése és felismerése

Az alternatívákat szavakra és vesszőkre tagoljuk. A szavak morfológiai elemzését a szoftverbe integrált Hunmorph[86] programmal végezzük. A szavak különböző morfológiai elemzései multiplikatíve alternatívákat generálnak. Például egy 5 szóból álló mondat esetén, ha minden szónak két alternatív morfológiai elemzése van, akkor 32 lehetséges mondatalternatívát generálunk.

A tokenizált alternatívákon először a szótár-alapú névelem felismerést végzünk (részleteket ld. [1]). Amennyiben valamely szó, vagy szósorozat névelemnek bizonyul, akkor a speciális „névelem" címkével látjuk el, amely függetlenül a benne szereplő szavak számától *egy* token lesz. Ha valamely token nem csak névelemként értelmezhető, akkor több alternatívát készítünk. Minta alapján az alábbi névelemeket ismerjük fel: postai címek; URL-k, e-mail címek; a pénzmennyiség; a dátumok/időpontok.



1. ábra: Példa többszavas névelem tokenizálására

Már a tokenizáció fázisában elvégzzük az alábbi szűréseket, amellyel csökkenthetjük a lehetséges alternatívák számát:

[86] http://mokk.bme.hu/eszkozok/hunmorph/

- összetett mondatok szűrése: ha több finite ige van a mondatban, az alternatívát eldobjuk. Ha nincs benne finite ige, akkor a „van” igét beszúrjuk.
- kérdőszó vizsgálat: ha az első szó nem megengedett kérdőszó, akkor az alternatívát eldobjuk

Az 1. ábrán egy többszavas névelem tokenizálását illusztráljuk.

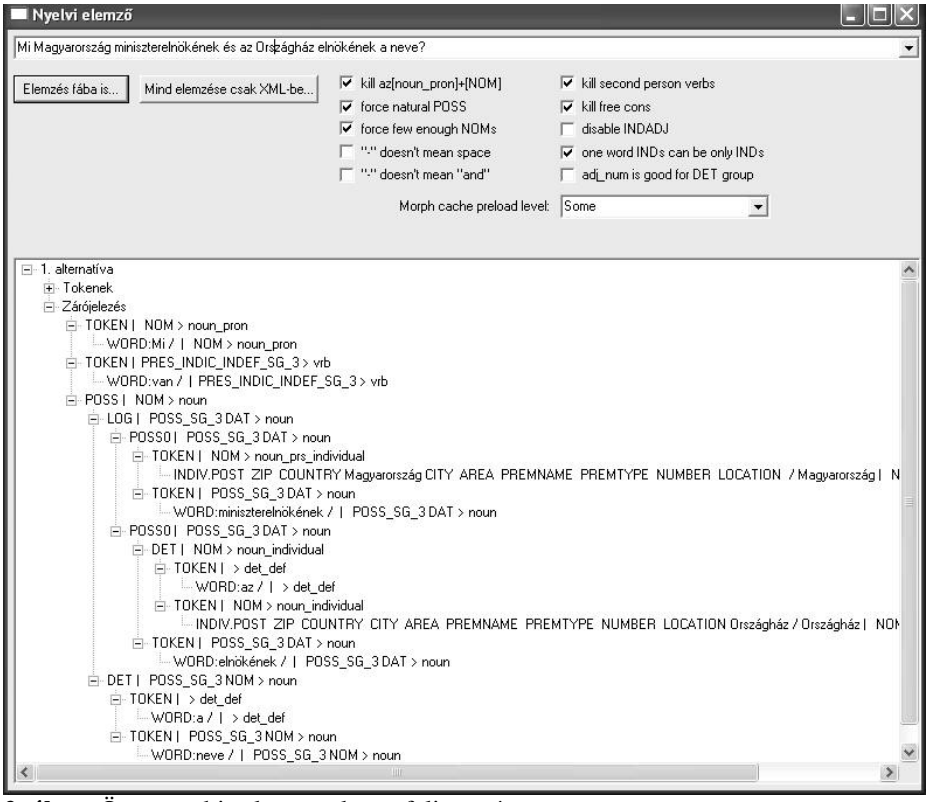
2.2 Zárójelezés

A zárójelezés célja, hogy az alternatívák tokenjeit egymásba ágyazott csoportokba ossza. A csoportok elemei szemantikai kapcsolatban lévő és így csoportban egységet képező (nem feltétlen szomszédos) tokenek.

Az alábbi csoportokat különböztetjük meg:

- igekötős szerkezetek; az igekötő [prv] címkével ellátott szavakhoz keresünk finite, nem utólagosan beillesztett igét. Elváló igekötők és az ige összekapcsolására alkalmas. Az ige nélküli igekötőt tartalmazó mondatalternatívákat eldobjuk.
- főnévi igeneves szerkezetek (ld. 1. ábra); [inf] címkéjű szavakhoz keresünk olyan segédigét vagy egyéb szót, amellyel a modalitást jellemző csoportot tudunk alkotni.
- főnévi csoport (jelzős szerkezetek), ahol a jelző melléknév vagy szótári névelem;
  - *-(j)ú,-(j)ű* ragú melléknévi szerkezetek felismerése; ekkor olyan szegmenseket keresünk, amelyek {[noun]+NOM}{[adj]_jÚ}{[noun]} alakúak, pl. *Mátrix című filmet.*
  - egyéb melléknévi szerkezetek felismerése; a zárójelezés feltétele, hogy valamely főnév [noun] előtti szó melléknév [adj], szótári névelem (ld. még opciók), vagy már korábban alkotott jelzős csoport legyen. Az eljárást rekurzívan végezzük.
- névelős szerkezetek; főnév, vagy főnévi csoport előtti névelőt [det] a csoporthoz kapcsoljuk.
- névutós szerkezetek; névutóból [post] és előtte álló főnévi csoportból képezzük.
- logikai összekapcsolások (és, vagy); a logikai kapcsoló két oldalán azonos morfológiai jellemzőkkel ellátott tokenekből (csoportokat) logikai csoportot képezünk. Azokat a felsorolásokat, ahol csak az utolsó két tag között szerepel kötőszó, a többi tag között pedig vessző, többszintű logikai csoportként ismerjük fel, ha a fenti feltételek fennállnak.
- birtokos szerkezetek; legfeljebb 3 elemű birtokos szerkezeteket ismerünk fel a morfológiai elemző által megadott birtokosra és birtokra jellemző toldalékok alapján (ld. 2. ábra.). A keresést a lehetséges birtokkal kezdjük, és ahhoz illesztjük a birtokosok láncát.
- értelmező jelzős szerkezetek; feltétele, hogy vessző előtt és után ugyanolyan morfológiai jellemzőjű csoport álljon. Ekkor két alternatívát gyárt, az egyikben csak az értelmezett kifejezés szerepel (vessző előtti rész), a másikban pedig csak az értelmezés (vessző utáni rész).

Ha valamely csoportképzés nem egyértelmű, akkor több alternatívát készítünk. A csoportok előfordulásainak keresését célszerűen kialakított, rögzített sorrendben végezzük. Vannak olyan csoportok, melyek keresése a rögzített sorrendben többször szerepel.



2. ábra: Összetett birtokos szerkezet felismerése

A program utolsó fázisa a szűrés, amikor a valamely szempontból hibás alternatívák eldobásra kerülnek. Az eldobás indokát az XML kimenetben feltüntetjük, ahol a teljes elemzési folyamatot is végig lehet követni.

3 Köszönetnyilvánítás

A cikk a Nemzeti Kutatási és Fejlesztési Pályázatok NKFP-0019/2002 jelű projektjének támogatásával készült.

Irodalomjegyzék

[1] D. Tikk et al: Ismert névelemek felismerése és morfológiai annotálása szabad szövegben, In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (2005), ebben a kötetben.

Névmutató