XII. Magyar Számítógépes Nyelvészeti Konferencia

MSZNY 2016

Szerkesztette:

Tanács Attila
Varga Viktor
Vincze Veronika

Szeged, 2016. január 21-22.
http://rgai.inf.u-szeged.hu/mszny2016

ISBN: 978-963-306-450-4

Szerkesztette: Tanács Attila, Varga Viktor és Vincze Veronika
{tanacs, vinczev}@inf.u-szeged.hu
viktor.varga.1991@gmail.com

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

Nyomtatta: JATEPress
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2016. január

Előszó

2016. január 21-22-én immár tizenkettedik alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát. A konferencia fő célkitűzése a kezdetek óta mit sem változott: a rendezvény fő célja a nyelv- és beszédtechnológia területén végzett legújabb, illetve folyamatban levő kutatások eredményeinek ismertetése és megvitatása, mindemellett lehetőség nyílik különféle hallgatói projektek, illetve ipari alkalmazások bemutatására is.

Örömömre szolgál, hogy a hagyományokat követve a konferencia idén is nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferenciafelhívásra idén is nagy számban beérkezett tudományos cikkek közül a programbizottság 25 előadást, 8 poszter-, illetve 4 laptopos bemutatót fogadott el. Újdonságot jelent, hogy egyes témákat mind az előadások, mind pedig a laptopos bemutatók között is megtalálunk, ezzel is lehetőséget adva a kutatási témák minél szélesebb körű bemutatására. A programban a magyar számítógépes nyelvészet rendkívül széles skálájáról találhatunk előadásokat a számítógépes morfológiától kezdve a beszédtechnológián át a szaknyelvi szövegek számítógépes feldolgozásáig. Mindemellett a magyar nyelvtechnológiai műhelyek együttműködésében megvalósuló, egy egységes magyar előfeldolgozó lánc kifejlesztését célzó INFRA projektnek is külön szekciót szentelünk.

Nagy örömet jelent számomra az is, hogy Pléh Csaba, az MTA rendes tagja elfogadta meghívásunkat, és plenáris előadása is gyarapítja a konferencia résztvevőinek szakmai ismereteit.

Ahogy az már hagyománnyá vált, idén is tervezzük a „Legjobb Ifjú Kutatói Díj” odaítélését, mellyel a fiatal korosztály tagjait kívánjuk ösztönözni arra, hogy kiemelkedő eredményekkel járuljanak hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz.

A konferencia sikeréhez a Neumann János Számítógép-tudományi Társaság szíves anyagi támogatása is hozzájárul, illetőleg a konferencia fogadása a MeltWater R&D nagylelkű támogatásával valósul meg. A rendezőbizottság nevében ezúton is szeretném kifejezni hálás köszönetünket mindkét támogatónknak.

Szeretnék köszönetet mondani a programbizottságnak: Vámos Tibor programbizottsági elnöknek, valamint Alberti Gábor, Kornai András, Németh Géza, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság és a kötetszerkesztők munkáját is.

Csirik János, a rendezőbizottság elnöke

Szeged, 2016. január

Tartalomjegyzék

IV. Szemantika, szentimentelemzés

V. Szaknyelv, speciális nyelvhasználat

VI. Szintaxis

VII. Poszterek

VIII. Laptopos bemutatók

IX. Angol nyelvű absztraktok

I. Lexikon, fordítás

Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat A neurálishálózat-alapú szemantikai beágyazási modelleket létrehozó algoritmusok a diszrtibúciós szemantika egy viszonylag új, de egyre népszerűbb alkalmazási területe. A szavakhoz vagy kifejezésekhez rendelt folytonos reprezentációk azok jelentését jól reprezentálják angol nyelvű tanítóanyagok esetén. Cikkünkben arra vonatkozó vizsgálatokat mutatunk be, hogy magyar nyelvre mennyire használhatóak ezek a modellek, illetve egy konkrét kategorizációs feladatban is kiértékeljük ezek hatékonyságát.

1. Bevezetés

A szavak reprezentációjának meghatározása a nyelvtechnológiai alkalmazások számára alapvető feladat. A kérdés az, hogy milyen reprezentáció az, ami a szavak jelentését, vagy azok morfoszintaktikai, szintaktikai viselkedését is meg tudja határozni. Angol nyelvre egyre népszerűbb a kézzel gyártott szimbolikus és nyers szövegből tanulható ritka diszkrét reprezentációk helyett a folytonos vektorreprezentációk alkalmazása, melyek hatékonyságát a neurális hálózatokra alapuló implementációk használatával több tanulmány is alátámasztotta [5,8,2]. Ezekben a kísérletekben és alkalmazásokban azonban a leírt módszereket általában egy a magyarhoz képest jóval kevesebb szóalakváltozattal operáló, kötött szórendű és egyszerű szószerkezeteket használó nyelvre alkalmazzák.

Cikkünk célja a folytonos reprezentációt implementáló modellek használhatóságának és hatékonyságának vizsgálata magyar nyelvre.

Vizsgálatunk motivációja azonban kettős. Egyik célunk a módszer szemantikai érzékenységének felderítése, azaz, hogy mennyire alkalmas arra, hogy magyar nyelvű korpuszon tanítva a szavakat a szemantikai térben konzisztensen helyezze el. Másrészt pedig egy konkrét alkalmazás támogatása is a célok között szerepelt: egy morfológiai elemző adatbázisának kiegészítése olyan szemantikai jegyekkel, amelyek hatással vannak a szavak morfológiai, helyesírási, illetve szintaktikai viselkedésére. Ilyenek például a színek, anyagnevek, népnevek, nyelvek, foglalkozások, stb. Ezek kézzel való összegyűjtése és az adatbázishoz való hozzáadása igen idő- és munkaigényes feladat, ezért ennek a feladatnak az automatizálása szintén céljaink között szerepelt, kísérleteink egy része ezeknek a szemantikai csoportoknak a létrehozására ad módszert.

2. Folytonos disztribúciós szemantikai modellek

A disztribúciós szemantika lényege, hogy a szavak jelentése szorosan összefügg azzal, hogy milyen kontextusban használjuk őket. A hagyományos disztribúciós szemantikai modellek létrehozásakor az egyes szavak előre meghatározott méretű környezetét az azokban előforduló szavak nagy korpuszból számított előfordulási statisztikái alapján határozzuk meg.

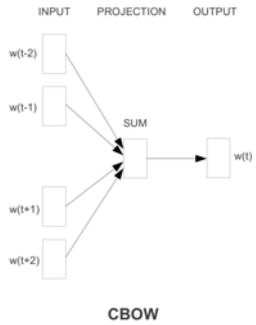
Ezzel szemben a nyelvtechnológiai kutatások egyik kurrens módszere a folytonos vektoros reprezentációk alkalmazása (*word embedding*), melyek nyers szöveges korpuszból szemantikai információk kinyerésére alkalmazhatók. Ebben a rendszerben a lexikai elemek egy valós vektortér egyes pontjai, melyek konzisztensen helyezkednek el az adott térben, azaz, az egymáshoz szemantikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. Mindemellett, a vektoralgebrai műveletek is alkalmazhatók ebben a térben, tehát két elem szemantikai hasonlósága a két vektor távolságaként meghatározható, illetve a lexikai elemek pozícióját reprezentáló vektorok összege, azok jelentésbeli összegét határozzák meg [8,6]. A módszer hátránya csupán az, hogy önmagában nem képes a poliszémia, illetve homonímia kezelésére, tehát egy többjelentésű lexikai elemhez is csupán egyetlen jelentésvektort rendel, azonban a szakirodalomban erre a problémára is találunk sikerrel alkalmazott módszereket [1,3,10].

Ennek a modellnek a tanítása során is az egyes szavak fix méretű környezetét vesszük figyelembe, az ezekből álló vektor azonban egy neurális hálózat bemenete. A környezetet reprezentáló vektorok összegét használja a hálózat arra, hogy megjósolja a célszót. A tanítás során a hiba visszaterjesztésével és ennek megfelelően a környezetet reprezentáló vektorok frissítésével jön létre a célszót helyesen megjósoló súlyvektor, ami a neurális hálózat megfelelő rétegéből közvetlenül kinyerhető. Mivel a hasonló szavak hasonló környezetben fordulnak elő, ezért a szövegkörnyezetre optimalizált vektorok a hasonló jelentésű szavak esetén hasonlóak lesznek. Az erre a feladatra felépített neurális hálózat a CBOW (*continous bag-of-words*) modellt implementálja, ami az 1. ábrán látható. Egy másik lehetőség az ún. skip-gram modell alkalmazása, amikor a hálózat bemenete a célszó, az optimalizálás célja pedig e szó környezetének megjósolása.

3. Kísérletek

A kísérleteinkben használt modelleket a `word2vec`[3] eszközzel hoztuk létre, ami mind a CBOW, mind a skip-gram modellek implementációját tartalmazza és a lexikai elemeket reprezentáló vektorok közvetlenül kinyerhetőek belőle. Mivel a két modell közül a CBOW modell betanítása hatékonyabb nagy tanítókorpuszok esetén, ezért mindegyik tanítás során ezt alkalmaztuk. Tanítóanyagként pedig egy majdnem 4 milliárd szavas magyar nyelvű webkorpuszt használtunk. Minden modell esetén 300 dimenziós vektorokat definiáltunk a lexikai elemek

[3] https://code.google.com/p/word2vec/



1. ábra. A CBOW (*continous bag-of-words*) modell

reprezentálására és 5 token sugarú mintavételezési ablakot a szövegkörnyezet kinyerésére.

3.1. Nyers szövegen tanított modell

Először egy a korpusz nyers változatán tanított modellt hoztunk létre (`SURF`), ami a szavak felszíni alakját reprezentáló vektorokat határozott meg, így az azonos tőhöz tartozó különböző ragozott alakok külön pozícióba kerültek a szemantikai térben. Ez a modell tehát a különböző morfológiai analógiák felderítésére használható. Például a *jó – rossz* és a *jobb – rosszabb* szópárok hasonlósága sokkal erősebb, mintha az azonos tő szerint hasonlítjuk őket össze (*jó – jobb*, illetve *rossz – rosszabb*). Ez a modell tehát jól reprezentálja a szemantikai és szintaktikai hasonlóságot. Néhány további példa az ebben a modellben az egy-egy szóhoz legközelebb álló szavakra a 1. táblázatban látható. A példákban a szavak melletti számok a korpuszbeli előfordulások számát adják meg.

3.2. Előfeldolgozott szövegen tanított modell

A másik modellben a korpusz szófaji egyértelműsített változatát használtuk oly módon, hogy a szavak lemmáját tartottuk meg, melyek után, külön tokenként szerepeltek a morfológiai elemző által generált címkék `ANA`. Mivel ezek a címkék az aktuális szó környezetében megmaradtak, ezért az általuk reprezentált szintaktikai információ továbbra is szerepet kapott az egyes szavakat reprezentáló vektorok létrehozásában, azonban a modell csak lemmákat tartalmazott, így robusztusabb modell jött létre az adatritkaság csökkenése miatt. A 2. táblázat néhány példát tartalmaz az ezzel a modellel kapott hasonlósági listákra. Látható, hogy a modell rangsorolása jól működik a szavak gyakoriságától függetlenül, hiszen a nagyon gyakori szavak nem előzik meg a szemantikailag jobban hasonló kifejezéseket.

1. táblázat. Példák a nyers szövegből kinyert modellek alapján kapott hasonló kifejezésekre. A zárójeles számok a korpuszbeli előfordulások számát mutatják.

kenyerek	pirosas	egerekkel	fiaik	megeszi
kiflik(349)	lilás(2476)	patkányokkal(524)	lányaik(593)	eszi(12615)
zsemlék(283)	rózsaszínes(1638)	férgekkel(513)	leányaik(251)	megenné(563)
lepények(202)	barnás(6463)	majmokkal(606)	férjeik(759)	elfogyasztja(1129)
pogácsák(539)	sárgás(7365)	hangyákkal(343)	gyermekeik(12028)	megeszik(6433)
pékáruk(771)	zöldes(5215)	nyulakkal(366)	feleségeik(638)	Megeszi(189)
péksütemények(997)	fehéres(2517)	legyekkel(252)	gyerekeik(5806)	megette(7868)
sonkák(613)	vöröses(5496)	rágcsálókkal(259)	asszonyaik(458)	megrágja(477)
tészták(2466)	feketés(1157)	hüllőkkel(241)	gyermekei(31241)	megeheti(287)
kalácsok(277)	narancssárgás(429)	pókokkal(436)	fiak(1523)	bekapja(977)
kekszek(1046)	sárgászöld(723)	bogarakkal(425)	unokái(3528)	lenyeli(1862)

2. táblázat. Példák a tövesített és elemzett szövegből kinyert modellek alapján kapott hasonló kifejezésekre. A zárójeles számok a korpuszbeli előfordulások számát adják meg.

kenyér	eszik	csavargó	csónak	franciakulcs
hús(136814)	iszik(244247)	koldus(15793)	tutaj(3950)	feszítővas(846)
kalács(10658)	főz(120634)	zsivány(3497)	ladik(3895)	csípőfogó(345)
rizs(31678)	csinál(1194585)	haramia(2024)	motorcsónak(4079)	csavarkulcs(473)
zsemle(6690)	megeszik(68347)	vadember(2497)	hajó(238807)	kisbalta(491)
pogácsa(11066)	fogyaszt(160724)	csirkefogó(2019)	kenu(6649)	konyhakés(1501)
sajt(46660)	etet(43539)	szatír(1649)	kocsi(283438)	pajszer(567)
kifli(9715)	zabál(13699)	útonálló(1942)	gumicsónak(1033)	partvis(648)
krumpli(37271)	megiszik(31002)	bandita(6334)	mentőcsónak(2511)	villáskulcs(764)
búzakenyér(306)	eszeget(3928)	suhanc(4144)	dereglye(962)	erővágó(360)
tej(113911)	alszik(359268)	vándor(14070)	sikló(4394)	péklapát(475)

3.3. Helyesírási hibák és nem sztenderd szóalakok

A modell vizsgálata során fény derült arra is, hogy a jelentésben hasonló szavak között megjelentek a különböző elírt változatok is. Ezek adták az ötletet arra, hogy olyan szóalakokhoz tartozó listákat is lekérdezzünk, melyek eleve hibásak. Ebben az esetben olyan szóalakokat kaptunk eredményül, melyek ugyanolyan vagy hasonló jellegű helyesírási hibát tartalmaznak, vagy amiket a lemmatizáló ugyanúgy rontott el, ugyanakkor ezekben a listákban is érvényesül a szemantikai rangsor. A 3. táblázat első két oszlopa ilyen példákat tartalmaz. A rendszernek ez a képessége jól hasznosítható hibák felderítésére és javítására, illetve egy adott nyelvtechnológiai feladat hibatűrővé tételére azáltal, hogy a számára ismeretlen szavakat is egy ismert szóhoz való hasonlósága révén kezelhetővé tesszük.

Mivel a tanítókorpusz a webről gyűjtött szövegekből áll, ezért sok nem sztenderd vagy szleng szóalak is előfordul benne. A modell ezekre is jól működik, ami szintén jól hasznosítható a csupán sztenderd szóalakokat ismerő szövegfeldolgozó

eszközök támogatása során. A 3. táblázat utolsó két oszlopa ilyen kifejezésekre kapott eredményeket tartalmaz.

3. táblázat. Példák a rendszer által a hibásan lemmatizált (első oszlop) és a hibásan írt (második oszlop) szavakhoz visszaadott hasonló kifejezésekre, illetve nem sztenderd szóalakokra (utolsó két oszlop).

pufidzsek	angolúl	mittomén	hehehe
rövidnac(43)	magyarúl(486)	mittudomén(2969)	hihihi(1203)
napszemcs(37)	németűl(132)	mifene(2455)	hahaha(3822)
szemcs(37)	francziául(25)	mittoménmi(412)	höhö(1827)
szmöty(45)	angolol(27)	mittudoménmi(441)	brr(1212)
zacs(170)	írül(95)	nemtommi(469)	muhaha(1498)
suzuk(131)	mindenrõl(422)	neadjisten(1741)	heh(1603)
sap(374)	minderrõl(129)	blablabla(2590)	Muhaha(879)
törcs(11)	ilyenröl(58)	stbstb(1739)	muhahaha(428)
kispolszk(41)	Amirõl(143)	bla-bla-bla(711)	hajaj(1579)
sokmindenk(58)	olyasmirôl(38)	jahh(466)	höhöhö(361)

3.4. Analógiavizsgálatok

A beágyazási modellek kiértékelésének egyik módszere az angol nyelvű modellek esetén az analógiatesztek elvégzése [7]. Ezeknél a teszteknél egy szópárosból és egy tesztszóból indulnak ki. A rendszer feladata annak a szónak a megtalálása, ami tesztszóhoz az eredeti szópáros közötti relációnak megfelelően viszonyul. Például a *férfi – nő* páros és a *király* tesztszó esetén a várt eredmény a *királynő*. Elvégeztünk ugyan néhány ilyen tesztet, azonban mivel a többértelmű szavakhoz egy reprezentációs vektor tartozik, ezért a szópárok közötti relációkat kevésbé sikerült jól modellezni. Az előbbi példában a *nő* szó igei és főnévi jelentései keverednek, ezért a *férfi* és a *nő* szavak közötti távolság nem pontosan felel meg a *király* és a *királynő* közötti távolságnak (aminek oka a *király* szó többértelműsége is). Így csupán elvétve találtunk olyan analógiapéldákat, melyek helyes eredményt adtak. Ilyen volt például a *hó – tél* páros és a *nap* tesztszó esetén eredményül kapott *nyár*. Részletes kiértékelést azonban ebben a feladatban nem végeztünk, hiszen előbb a jelentés-egyértelműsítés problémakörének megoldását tartjuk kritikus fontosságúnak.

3.5. Szemantikai csoportok kinyerése

A fenti modelleket szemantikai csoportok kinyerésére használtuk fel. Mivel a cél ebben a részfeladatban a kifejezések szemantikai besorolása volt, ezért ehhez csak az `ANA` modellt (tehát a lemmákat tartalmazót) használtuk. Minden szemantikai csoporthoz meghatároztunk egy kezdő szót, ami az adott csoportba tartozik.

Ehhez a szóhoz meghatároztuk a 200 leghasonlóbb szót a létrehozott modellből, majd ennek a listának a 200. eleméhez szintén lekérdeztük a 200 leghasonlóbb szót és ezt a lépést ismételtük legfeljebb 10 alkalommal. Az így létrejött max. 2000 elemű listában ellenőriztük, hogy melyik indikátorszó nem járult hozzá a korábbiakhoz képest új elemekkel, ezeket a szavakat töröltük a lekérdezések közül, majd újra lefuttattuk az algoritmust. Így minden szemantikai csoporthoz, a csoportba tartozó egyetlen kiindulási szó meghatározása után több száz vagy akár ezer, az azonos csoportba tartozó kifejezést nyertünk ki automatikusan. Úgy találtuk, hogy bizonyos (szűkebb) szemantikai mezőkben a 200 szavankénti lekérdezés túl sok zajt eredményezett, például amikor kifejezetten ruhaanyagok gyűjtése volt a cél. Ekkor az egyszerre lekérdezett kvantum 50 eleműre csökkentésével kaptunk viszonylag jól használható eredményt.

4. Eredmények

Az eredmények vizsgálatát több módszerrel végeztük. A szemantikai kategorizációs feladatban kézzel számoltuk meg az eredményül kapott listában a helyes és nem helyes szavak arányát. Ahhoz azonban, hogy a kézzel történő ellenőrzést hatékonyabban tudjuk végezni, egy klaszterezést is alkalmaztunk az eredménylistára, illetve az eredménylistában szereplő szavak sokdimenziós reprezentációját leképeztük egy kétdimenziós térbe, ahol a klaszterezés eredményével együtt jelenítettük meg a szavakat, jól áttekinthető vizuális megjelenítéssel támogatva az ellenőrzést.

4.1. Klaszterezés

A lexikai elemek klaszterezéséhez hierarchikus klaszterezést alkalmaztunk, melynek bemenete a csoportosítandó szavakat tartalmazó listán szereplő lexikai elemekhez tartozó szemantikai vektor, a klaszterezés során pedig a vektorok távolságát Ward [11] módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre. Célunk azonban nem egy bináris faként ábrázolt teljes hierarchia meghatározása volt, hanem a fogalmak elkülönülő csoportjainak meghatározása, azaz a kapott dendrogram egyes kompakt részfái. A klaszterezés és a részfák kivágására szolgáló módszer részleteit [9]-ben közöltük. A 4. táblázatban néhány eredményül kapott klaszterre láthatunk példát egy-egy szemantikai kategórián belül. Jól látható, hogy az egy klaszterbe sorolt kifejezések egymáshoz szorosabban kapcsolódnak az adott kategórián belül is. Természetesen, az algoritmus lehetőséget biztosít a klaszterezés kifinomultságának állítására, így akár nagyobb, vagy még kisebb csoportosítás is könnyen kinyerhető. A példák között a foglalkozások között kiemelendő a különböző katonai rangok rövidített alakjainak csoportja, illetve a nyelvek esetén a magyar nyelvjárásokat összegyűjtő csoport. Külön klaszterekbe gyűltek össze az adott feladat szempontjából ugyan szemantikailag releváns, de önmagában nem tökéletes megoldások is, például a

nyelveknél azok a földrajzi nevek, amelyek egy-egy nyelvváltozat jelzői, de önmagukban nem nyelvnevek, a nyelvpárok, illetve a kifejezetten tévesen a listán feltűnő elemek, például színpárok. Ez meglehetős mértékben megkönnyíti a generált listák kézi ellenőrzését is, mert a nyilvánvalóan hibás csoportok gyorsan kiszűrhetők.

4. táblázat. Klaszterekbe rendezett kifejezések a négy vizsgált szemantikai csoport esetén

Foglalkozások
író költő író drámaszerző prózaíró novellista színműíró regényíró drámaíró
ökológus entomológus zoológus biológus evolúcióbiológus etológus
hidegburkoló tapétázó mázoló szobafestő festő-mázoló szobafestő-mázoló bútorasztalos
tehénpásztor kecskepásztor birkapásztor fejőnő marhahajcsár tehenész marhapásztor
őrm ftörm zls alezr vőrgy szkv ezds hdgy őrgy szds fhdgy
Nyelvek
kuwaiti szaudi szaúdi kuvaiti jordán szaúd-arábiai jordániai
lengyel cseh bolgár litván román szlovák szlovén horvát szerb
osztrák-német német-osztrák elzászi dél-tiroli flamand
bánsági háromszéki gömöri széki gyimesi felföldi sárközi
Anyagnevek
feketeszén kőszén barnaszén lignit feketekőszén barnakőszén
fluorit rutil apatit aragonit kvarc kalcit földpát magnetit limonit
konyhasó kálium-klorid nátriumklorid nátrium-klorid
Textilek
selyemszatén béléselyem düsesz shantung
posztó szűrposztó abaposztó őzbőr teveszőr kendervászon házivászon háziszőttes
csipke bársony selyem kelme brokát selyemszövet tafota damaszt batiszt

4.2. Vizualizáció

Mivel a fogalmakat reprezentáló vektorok egy szemantikai térben helyezik el az egyes lexikai elemeket, ezért gyakran alkalmazott módszer ennek a szerveződésnek a vizualizációja. Ehhez a sokdimenziós vektorokat egy kétdimenziós térbe képeztük le a t-sne algoritmus alkalmazásával [4]. A módszer lényege, hogy a szavak sokdimenziós térben való páronkénti távolságának megfelelő eloszlást közelítve helyezi el azokat a kétdimenziós térben, megtartva tehát az elemek közötti távolságok eredeti arányát. Így könnyen áttekinthetővé válik a szavak szerveződése, a jelentésbeli különbségek jól követhetőek és felmérhetőek.

A vizualizáció során a klaszterezés eredményeit is megjelenítettük, a különböző klaszterbe került szavakat különböző színnel jelenítve meg. Az így létrejött ábrán jól követhetővé váltak a klaszterek közötti távolságok is.

2. ábra. A foglalkozásnevek t-sne vizualizációjának egy kis részlete. A kifejezések egymástól való távolsága az ábrán arányos a jelentésbeli hasonlóságukkal.

4.3. Kvantitatív kiértékelés

A klaszterezésnek köszönhetően tehát könnyebbé vált az egyes szemantikai kategóriákba sorolt szavak ellenőrzése, hiszen az egy klaszterbe sorolt szavak csoportosan ellenőrizhetőek voltak, mivel a részfák kivágása során úgy állítottuk be a paramétereket, hogy inkább kisebb, de tömör és koherens csoportok jöjjenek létre. Az így kapott eredményeket a 5. táblázat foglalja össze.

5. táblázat. A szemantikai kategorizáció eredményei

	nyelvek	foglalkozások	anyagnevek	textilek
Helyes	755	2387	1139	120
Hibás	98	134	162	114
Átmeneti	391	37	54	0
Összesen	1244	2558	1355	234

Négy szemantikai kategóriára végeztük el a méréseket: nyelvek, foglalkozások, anyagnevek és ezen belül a textiltípusokra. A kapott lista szavait (illetve a klaszterezés eredményeként kapott klasztereket, amikor ezek homogének voltak) három kategóriába soroltuk: helyes, hibás, illetve létrehoztunk egy átmeneti kategóriát, amelybe a feladat szempontjából nem egészen hibás, de nem is tökéletes megoldások kerültek (például a foglalkozásnevek között az egyéb embert, emberszerű lényt vagy foglalkozástól különböző emberi szerepet jelölő szavak, pl. *srác, öregasszony, hölgy, albérlő, élettárs, kobold*). Látható, hogy a textilek kategóriáját kivéve mindegyik esetben igen nagy volt a helyes szavak aránya (a textilek esetében pedig nagyon egyszerű volt a találatok kézi ellenőrzése, mert gyakor-

latilag az összes téves találat külön klaszterekbe gyűlt össze, amelyek kizárólag ruhaanyagokból készült cikkeket: ruhadarabokat, lábbeliket, lakástextiltermékeket tartalmaztak). A 10 indikátor szó alapján 755 nyelv, 2387 foglalkozás és 1139 anyagnév gyűlt össze, amik igen jó eredménynek számítanak ahhoz képest, ha ezeket a listákat kézzel kéne összeállítani. Sok esetben az átmeneti jelölést kapott szavak is helyesek lehetnek egy-egy feladatban, most azonban a legszigorúbb értékelést alkalmaztuk, ezért nem jelöltük őket elfogadottnak.

5. Részletes hibaelemzés

A négy kategória közül az egyikre (nyelvek) részletes hibaelemzést is készítettünk. Az egyes szavak helyességének, illetve a nem nyelvként szereplő nevek hibatípusának megítélésekor az eredeti célt tartottuk szem előtt, azaz a morfológiai adatbázis szemantikai jegyekkel való bővítését. Így, ebben az esetben több szóalakot is elfogadhatónak tekintettünk.

A 6. táblázat a különböző nyelvkategóriák disztribúcióját tartalmazza, melyek a következők:

Az első csoport nyelveket, nyelvtípusokat tartalmaz.

- Sztenderd nyelvek: egy nyelv hivatalos neve, a helyesírási szabálynak megfelelő alakban.
- Kitalált nyelv: egy irodalmi alkotás szerzője által kitalált nyelv neve.
- Egy nyelvcsoport vagy nyelvcsalád neve: pl. *uráli*
- Népcsoport neve, de nem nyelv: pl. *zsidó*. Ezeket a kifejezéseket a köznyelvben gyakran használják úgy, mintha nyelvek lennének (pl. *zsidó nyelv, zsidóul*).
- Egy írásrendszer neve: pl. *dévanágari*, *cirill*. A nyelvtani konstrukciók, amikben ezek szerepelnek hasonlóan viselkednek a nyelvekkel használt konstrukciókhoz.
- Nyelvtípus: pl. *kreol*, *patois*, *pidzsin* (az ilyen típusú nyelvek összetett nevének utolsó része)

A második csoportba nyelvek attribútumait sorolhatjuk:

- Földrajzi helyet jelölő tulajdonság: egy nyelv, dialektus vagy nyelvcsoport jelzője, ami önmagában nem használható a nyelv neveként, pl. *iraki* (arab), *mezopotámiai* (nyelvek)'
- Más (nem földrajzi) attribútumok: *rabbinkus* (héber)'

A harmadik csoportba helyesírási változatokat, szinonimákat és elírt változatokat soroltunk:

- Szinonimák: egy nyelv alternatív (pl. régies) megnevezése, pl. *tót – szlovák* , *hellén – görög*.
- Helyesírási változatok (nyelv, nyelvcsoport vagy dialektus esetén): archaikus alakok, fonetikai variánsok vagy latin helyesírás szerinti alakok, pl. *franczia*, *bulgár*, *szittya*, *scytha*

– Súlyosabb elírások: egy nyelv, dialektus vagy nyelvcsoport nevében hiányzó, fölösleges, vagy felcserélt betűk

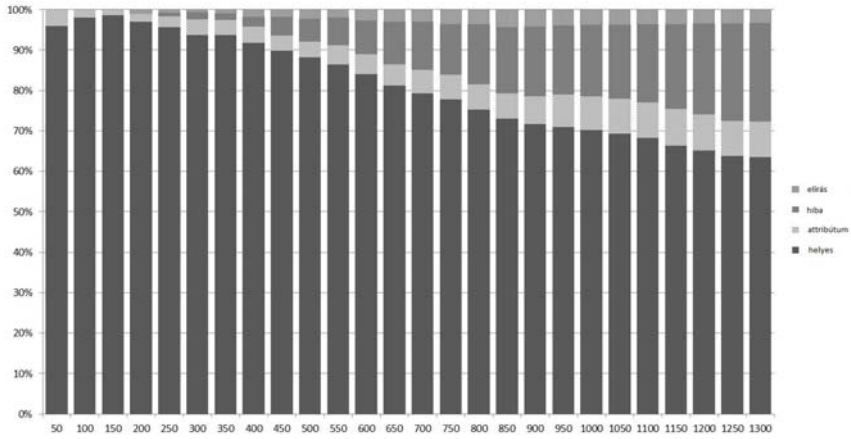
Az ebbe a három csoportba tartozó szóalakok a morfológiai elemző adatbázisának bővítése szempontjából nyelvnek tekinthetők. Ezek a közel 1300 szónak a 74,96%-át teszik ki. A többi 25,04% nem nyelvmegnevezés. Ide soroltuk például azokat a nyelvpárokat (pl. *magyar-angol*), ahol a nyelvpár nem egy nyelvcsoportot jelöl, viszont az olyan párokat, mint pl. a *bajor-osztrák*, ahol a két nyelv együtt alkot egy dialektust, azokat nyelvként fogadtuk el.

6. táblázat. A nyelvekre készített részletes hibaelemzés eredménye. A százalékértékek az 1244 elemű listából számított arányok.

típus	**példa**	**pontosság**
sztenderd nyelv	*joruba*	39,83%
kitalált nyelv	*újbeszél*	1,11%
dialektus neve	*Cockney*	5,33%
nyelvcsoport vagy nyelvcsalád neve	*uráli*	4,37%
népcsoport, de nem nyelv	*zsidó*	1,03%
írásrendszer	*cirill*	0,72%
nyelvtípus	*kreol*	0,32%
írásváltozat	*scytha*	10,25%
szinonima	*hellén*	2,07%
elírás	*ngol*	3,42%
földrajzi jelző	*iraki*	8,51%
más jelző	*rabbinkus*	0,40%
		74,96%
nem nyelv, nyelvpár	*magyar-angol*	25,04%

A 3. ábra a módszer pontosságának alakulását mutatja az automatikusan kinyert nyelvnévlista hosszának függvényében. Látható, hogy a lista elején sokkal kevesebb hiba található, míg ha az eredeti indikátorszavaktól egyre távolabb kerülünk a szemantikai térben, úgy kerül be egyre több zaj a kinyert listába. Az ábra jól illusztrálja a word2vec algoritmusban implementált hasonlóságszámítás hatékonyságát is, ami alapján ez a rangsorolás létrejön.

A módszer által adott lista fedésének becslése jóval nehezebb feladat, mint a pontosság meghatározása, mivel magyar nyelven nem találtunk a nyelveket, nyelvcsaládokat és nyelvcsoportokat tartalmazó teljes listát. (Ha létezne ilyen, akkor ezt használhattuk volna az eredeti feladatban is.) Ugyanez igaz a többi szemantikai kategóriára (foglalkozások, anyagnevek, stb.), ráadásul a bemutatott módszer tetszőleges szemantikai csoport kinyerésére alkalmazható.



3. ábra. A módszer pontossága az automatikusan kinyert lista hosszának függvényében. A *helyes* szavak azok, amiket nyelvnek fogadtunk el, az *attribútumok*, amiket nyelvek jelzőinek, az *elírások* olyan nyelvnevek, nyelvcsoportok, nyelvcsaládok, stb., amikben kisebb elírás szerepel, a *hiba* kategóriába pedig azok a szavak tartoznak, amik a fentiek közül egyik kategóriába sem tartoznak.

6. Konklúzió

Cikkünkben bemutattuk, hogy az egyre népszerűbb, neurális hálózatok betanításán alapuló szemantikai beágyazási modellek magyar nyelvre is jó eredménnyel működnek kellő méretű és elemzett tanítóanyag alkalmazása esetén. Néhány általános kísérlet elvégzése mellett a létrejött szóreprezentációk egy konkrét feladatra való felhasználhatóságát is megvizsgáltuk. Ennek során célunk többek között egy meglévő morfológiai elemző lexikonában a morfológiai, szintaktikai, szemantikai szempontból releváns kategóriainformáció gazdagítása, illetve ellenőrzése. Mivel a modell alkalmasnak bizonyult arra, hogy szavakhoz azokhoz valamilyen szempontból hasonló szavakat rendeljen, ezért az egy kategóriába (foglalkozások, nyelvek, anyagnevek) tartozó szavak automatikusan kinyerhetőek. Továbbá, a modellek folytonosságából adódóan a hasonlóság mértéke tetszőlegesen állítható, így a kategorizálás különböző absztrakciós szinteken valósítható meg. Az eredményekben megmutattuk, hogy számos olyan szót tudtunk a megfelelő kategóriacímkével ellátni, melyre kézi gyűjtés esetén csak nagyon sok további munka árán lett volna lehetőség. Ugyancsak alkalmasnak bizonyult a módszer a különböző annotációs és egyéb korpuszhibák kimutatására és osztályozására is.

Hivatkozások

1. Banea, C., Chen, D., Mihalcea, R., Cardie, C., Wiebe, J.: Simcompass: Using deep learning word embeddings to assess cross-level similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 560–565.

Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014), http://www.aclweb.org/anthology/S14-2098
2. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 238–247. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-1023
3. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: Learning sense embeddings for word and relational similarity. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 95–105. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-1010
4. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne (2008)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119 (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality
8. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://www.aclweb.org/anthology/N13-1090
9. Siklósi, B., Novák, A.: Közeli rokonunk, az autó. In: Tanács, A., Varga, V., Vincze, V. (eds.) XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 27–36. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2016)
10. Trask, A., Michalak, P., Liu, J.: sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. CoRR abs/1511.06388 (2015), http://arxiv.org/abs/1511.06388
11. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963), http://www.jstor.org/stable/2282967

Building Definition Graphs using Monolingual Dictionaries of Hungarian

Gábor Recski[1], Attila Bolevácz[1], Gábor Borbély[2]

[1] Research Institute for Linguistics
Hungarian Academy of Sciences
recski@mokk.bme.hu, attila.bolevacz@protonmail.hu

[2] Department of Algebra
Budapest University of Technology
borbely@math.bme.hu

1 Introduction

We adapt to Hungarian core functionalitites of the `4lang` library [12], which builds `4lang` -style semantic representations [7] from raw text using an external dependency parser as proxy, and processes definitions of monolingual dictionaries to build definition graphs for concepts not defined in the hand-written `4lang` dictionary [8]. In Section 2 we provide a short overview of the `4lang` formalism, Section 3 describes the architecture of the `text_to_4lang` and `dict_to_4lang` systems. We describe in detail the steps taken to adapt our system to Hungarian in Section 4. The new tool is evaluated in Section 5. The new components presented in this paper are part of the latest version of the `4lang` library, which is available under an MIT license from http://www.github.com/kornai/4lang.

2 The `4lang` representation

`4lang` is both a formalism for representing meaning via directed graphs of concepts and also the name of a manually built lexicon of such representations for ca. 2700 words[3]. A formal presentation of the system is given in [7], the theoretical principles underlying `4lang` are presented in [5], we shall provide a short overview only.

`4lang` meaning representations are directed graphs of concepts with three types of edges. Nodes of `4lang` graphs correspond to *concepts*. `4lang` concepts are not words, nor do they have any grammatical attributes such as part-of-speech (category), number, tense, mood, voice, etc. For example, `4lang` representations make no distinction between the meaning of *freeze* (N), *freeze* (V), *freezing*, or *frozen*. Therefore, the mapping between words of some language and the language-independent set of `4lang` concepts is a many-to-one relation. In particular, many concepts will be defined by a single link to another concept

[3] https://github.com/kornai/4lang/blob/master/4lang

that is its hypernym or synonym, e.g. `above` $\xrightarrow{0}$ `up` or `grasp` $\xrightarrow{0}$ `catch`. Encyclopaedic information is omitted, e.g. `Canada`, `Denmark`, and `Egypt` are all defined as `country`, their definitions also containing an indication that an external resource (we use Wikipedia for this) may contain more information. In general, definitions are limited to what can be considered the shared knowledge of competent speakers - e.g. the definition of `water` contains the information that it is a colourless, tasteless, odourless liquid, but not that it is made up of hydrogen and oxigen.

The most common connection in `4lang` graphs is the 0-edge, which represents attribution: `dog` $\xrightarrow{0}$ `friendly`, the `IS_A` relation (synonymy and hypernymy): `dog` $\xrightarrow{0}$ `animal`, and unary predication: `dog` $\xrightarrow{0}$ `bark`. Edge types 1 and 2 connect binary predicates to their arguments, e.g. `cat` $\xleftarrow{1}$ `catch` $\xrightarrow{2}$ `mouse`). There are no ternary or higher arity predicates, see [6]. The formalism used in the `4lang` dictionary explicitly marks binary (transitive) elements – by using UPPERCASE printnames. The tools presented in this paper make no use of this distinction, any concept can have outgoing 1- and 2-edges. However, we will retain the uppercase marking for those binary elements that do not correspond to any word in a given phrase or sentence. The `4lang` tools described here also enforce a slight modification to the formalism: the 0-relation shall hold between a subject and predicate regardless of whether the predicate has another argument, so that e.g. the `4lang` representations for *John eats* and *John eats a muffin* shall share the subgraph `John` $\xrightarrow{0}$ `eat`. The `4lang` dictionary contains manually specified definition graphs for ca. 2700 concepts, a typical definition in the dictionary can be seen in Figure 1. `4lang` contains words for each concept in four languages: English, Hungarian, Polish, and Latin.

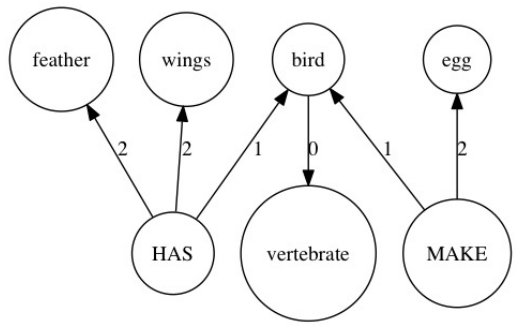


Fig. 1. 4lang definition of `bird`.

3 Architecture

The core tools in the `4lang` library include the `dep_to_4lang` module for processing the output of a dependency parser and building `4lang` representations by mapping dependencies to graph edges, the `text_to_4lang` module for using this functionality for mapping raw text to `4lang` graphs, and the `dict_to_4lang` module for processing monolingual dictionaries to acquire definition graphs for words not manually defined in the `4lang` dictionary. We now give a brief overview of these systems before presenting the modifications that enable us to run them on Hungarian data in Section 4.

The `dep_to_4lang` module implements a mapping from dependency triplets output by a syntactic parser to subgraphs over `4lang` concepts corresponding to content words in the sentence. Words are lemmatized using the `hunmorph` morphological analyzer [13], concept nodes are created for lemmas of each content word that takes part in a dependency relation that `dep_to_4lang` processes. The output of the dependency parser is first postprocessed by a separate, language-specific module that recognizes some patterns of dependencies and adds new triplets based on them that can later be used to create the correct `4lang` subgraphs. The mapping itself enforces two types of rules: some dependencies trigger an edge between two nodes, e.g. for a relation `dobj(x, y)` the edge `y` $\xrightarrow{2}$ `x` is added. Other relations will result in a binary node being added to the graph, e.g. the triplet `tmod(x, y)` will trigger `x` $\xleftarrow{1}$ `AT` $\xrightarrow{2}$ `y` (for a description of all Stanford dependency types see [2], for the full mapping for English see [12]). When processing raw English text using the `text_to_4lang` module, the Stanford Coreference Resolution system is run in addition to the Stanford Dependency parser and pairs of nodes in the resulting `4lang` graph are unified accordingly. The `dict_to_4lang` module for processing dictionary definitions contains parsers for various monolingual dictionaries of English, and also runs a preprocessor for each datasource that transforms the definitions in order to make them easier to parse and more informative; e.g. the pattern `someone who` will be removed from the beginning of Longman definitions, reducing parser errors considerably, but without losing any relevant information: the pattern also triggers the addition of the edge $\xrightarrow{0}$ `person` to the definition graph. Finally, the root node of each definition, which nearly always corresponds to a hypernym of the headword, is unified with the headword's node.

4 Modifications for Hungarian

In order to adapt the `text_to_4lang` and `dict_to_4lang` pipelines to Hungarian, we used the NLP library `magyarlanc` for dependency parsing and implemented a mapping to `4lang` graphs that is sensitive to the output of morphological analysis – to account for the rich morphology of Hungarian encoding many relations that a dependency parse cannot capture. We describe the output of `magyarlanc` and the straightforward components of our mapping in Section 4.1. In Section 4.2 we discuss the use of morphological analysis in our pipeline, and

in Section 4.3 we present some arbitrary postprocessing steps similar to those already implemented for English.

We shall also use our modifications to run the `dict_to_4lang` pipeline on two explanatory dictionaries of Hungarian: volumes 3 and 4 of the *Magyar Nyelv Nagyszótára* (`NSzt`), containing nearly 5000 headwords starting with the letter *b* [4][4], and over 120 000 entries of the complete *Magyar Értelmező Kéziszótár* (`EKsz`) [10], which has previously been used for NLP research [9]. Preprocessing of definitions involved replacing abbreviations in definitions, e.g. replacing *vmi* with *valami* 'something' or *Mo.* with *Magyarország* 'Hungary', performed by the `eksz_parser` and `nszt_parser` modules.

4.1 Dependencies

The `magyarlanc` library[5] [15] contains a suite of standard NLP tools for Hungarian, which allows us, just like in the case of the Stanford Parser, to perform tokenization, morphological analysis, and dependency parsing using a single tool. The dependency parser component of `magyarlanc` is a modified version of the Bohnet parser [1] trained on the Szeged Dependency Treebank [14]. The output of `magyarlanc` contains a much smaller set of dependencies than that of the Stanford Parser. Parses of the ca. 4700 entries of the NSzT data contain nearly 60,000 individual dependencies, 97% of which are covered by the 10 most frequent dependency types. The dependencies `att, mode,` and `pred`, all of which express some form of unary predication, can be mapped to the 0-edge. `subj` and `obj` are treated in the same fashion as the Stanford dependencies `nsubj` and `dobj`. The dependencies `from, tfrom, locy, tlocy, to,` and `tto` encode the relationship to the predicate of adverbs and postpositional phrases answering the questions 'from where?', 'from when?', 'where?', 'when?', 'where to?', and 'until when?', respectively, hence they are mapped to the binary relations `FROM, since, AT, TO,` and `until` (see Table 1).

4.2 Morphology

In Hungarian the relationship between a verb and its NP argument is often encoded by marking the noun phrase for one of 21 distinct cases – in English, these relations would typically be expressed by prepositional phrases. While the Stanford Parser maps prepositions to dependencies and the sentence *John climbed under the table* yields the dependency `prep_under(table, climb)`, the Hungarian parser does not transfer the morphological information to the dependencies, all arguments other than subjects and direct objects will be in the `OBL` relation with the verb. Therefore we updated the `dep_to_4lang` architecture to allow our mappings from dependencies to `4lang` subgraphs to be sensitive to the morphological analysis of the two words between which the dependency holds. The

[4] The author gratefully acknowledges editor-in-chief Nóra Ittzés for making an electronic copy available.

[5] http://www.inf.u-szeged.hu/rgai/magyarlanc

Table 1. Mapping from `magyarlanc` dependency relations to `4lang` subgraphs

Dependency	Edge
att mode pred	$w_1 \xrightarrow{0} w_2$
subj	$w_1 \xrightarrow{1} w_2$
obj	$w_1 \xrightarrow{2} w_2$
from	$w_1 \xleftarrow{1} \texttt{FROM} \xrightarrow{2} w_2$
tfrom	$w_1 \xleftarrow{1} \texttt{since} \xrightarrow{2} w_2$
locy tlocy	$w_1 \xleftarrow{1} \texttt{AT} \xrightarrow{2} w_2$
to	$w_1 \xleftarrow{1} \texttt{TO} \xrightarrow{2} w_2$
tto	$w_1 \xleftarrow{1} \texttt{until} \xrightarrow{2} w_2$

resulting system maps the phrase *a késemért jöttem* the knife-POSS-PERS1-CAU come-PAST-PERS1 'I came for my knife' to `FOR(come, knife)` based on the morphological analysis of *késem* performed by `magyarlanc` based on the `morphdb.hu` database [13].

While this method yields many useful subgraphs, it also often leaves uncovered the true semantic relationship between verb and argument, since nominal cases can have various interpretations that are connected to their 'primary' function only remotely, or not at all. The semantics of Hungarian suffixes *-nak/-nek* (dative case) or *-ban/-ben* (inessive case) exhibit great variation – not unlike that of the English prepositions *for* and *in*, and the 'default' semantic relations `FOR` and `IN` are merely one of several factors that must be considered when interpreting a particular phrase. Nevertheless, our mapping from nominal cases to binary relations can serve as a strong baseline, just like interpreting English *for* and *in* as `FOR` and `IN` via the Stanford dependencies `prep_for` and `prep_in`. The full mapping from nominal cases of `OBL` arguments to `4lang` binaries is shown in Table 2.

4.3 Postprocessing

In the Szeged Dependency Treebank, and consequently, in the output of `magyarlanc`, copular sentences will contain the dependency relation `pred`. Hungarian only requires a copular verb in these constructions when a tense other than the present or a mood other than the indicative needs to be marked (cf. Figure 3). While the first example is analyzed as `subj(Ervin, álmos)`, all remaining sentences will be assigned the dependencies `subj(Ervin, volt)` and `pred(volt, álmos)`. The same copular structures allow the predicate to be a noun phrase

Table 2. Mapping nominal cases of `OBL` dependants to `4lang` subgraphs

Case	Suffix	Subgraph
sublative superessive	*-ra/-re* *-on/-en/-ön*	$w_1 \xleftarrow{1} \texttt{ON} \xrightarrow{2} w_2$
inessive illative	*-ban/-ben* *-ba/-be*	$w_1 \xleftarrow{1} \texttt{IN} \xrightarrow{2} w_2$
temporal adessivel	*-kor* *-nál/nél*	$w_1 \xleftarrow{1} \texttt{AT} \xrightarrow{2} w_2$
elative ablative delative	*-ból/-ből* *-tól/-től* *-ról/-ről*	$w_1 \xleftarrow{1} \texttt{FROM} \xrightarrow{2} w_2$
allative terminative	*-hoz/-hez/-höz* *-ig*	$w_1 \xleftarrow{1} \texttt{TO} \xrightarrow{2} w_2$
causative	*-ért*	$w_1 \xleftarrow{1} \texttt{FOR} \xrightarrow{2} w_2$
instrumental	*-val/-vel*	$w_1 \xleftarrow{1} \texttt{INSTRUMENT} \xrightarrow{2} w_2$

(e.g. *Ervin tűzoltó* 'Ervin is a firefighter'). In each of these cases we'd like to eventually obtain the `4lang` edge `Ervin` $\xrightarrow{0}$ `sleepy` (`Ervin` $\xrightarrow{0}$ `firefighter`), which could be achieved in several ways: we might want to detect whether the nominal predicate is a noun or an adjective and add the `att` and `subj` dependencies accordingly. Both of these solutions would result in a considerable increase in the complexity of the `dep_to_4lang` system and neither would simplify its input: the simplest examples (such as (1) in Figure 3) would still be treated differently from all others. With these considerations in mind we took the simpler approach of mapping all pairs of the form `nsubj(x, c)` and `pred(c, y)` (such that `c` is a copular verb) to the relation `subj(x, y)`, which can then be processed by the same rule that handles the simplest copulars (as well as verbal predicates and their subjects.)

Unlike the Stanford Parser, `magyarlanc` does not propagate dependencies across coordinated elements. Therefore we introduced a simple postprocessing step where we collect words of the sentence governing a `coord` dependency, then find for each the words accessible via `coord` or `conj` dependencies (the latter connects coordinating conjunctions such as *és* 'and' to the coordinated elements). Finally, we unify the dependency relations of all coordinated elements[6].

[6] This step introduces erroneous edges in a small fraction of cases: when a sentence contains two or more clauses that are not connected by any conjunction – i.e. no connection is indicated between them – a `coord` relation is added by `magyarlanc` to connect the two dependency trees at their root nodes.

Table 3. Hungarian copular sentences

(1) *Ervin álmos* Ervin sleepy 'Ervin is sleepy'
(2) *Ervin nem álmos* Ervin not sleepy 'Ervin is not sleepy'
(3) *Ervin álmos volt* Ervin sleepy was 'Ervin was sleepy'
(4) *Ervin nem volt álmos* Ervin not was sleepy 'Ervin was not sleepy'

5 Evaluation

5.1 `text_to_4lang`

To evaluate the `text_to_4lang` pipeline we chose 20 random sentences and checked the output manually. The source of our sample is the Hungarian Webcorpus [3], to obtain a random sample we ran the GNU utility `shuf` on a sequence of files containing one sentence on each line. We shall start by providing some rough numbers regarding the average quality of the 20 `4lang` graphs, then proceed to discuss some of the most typical issues, citing examples from our sample. 10 of the 20 graphs were correct `4lang` representations, or had only minor errors. An example of a correct transformation can be seen in Figure 3. Of the remaining graphs, 4 were mostly correct but had major errors, e.g. 1-2 content words in the sentence had no corresponding node, or several erroneous edges were present in the graph. The remaining 6 graphs had many major issues and can be considered mostly useless.

When investigating the processes that created the more problematic graphs, nearly all errors seem to be caused by sentences with multiple clauses. When a clause is introduced by a conjunction such as *hogy* 'that' or *ha* 'if', the dependency trees of each graph are connected via these conjunctions only, i.e. the parser does not assign dependencies that hold between words from different clauses. While we are able to build good quality subgraphs from each clause, further steps are required to establish the semantic relationship between them based on the type of conjunction involved – a process that requires case-by-case treatment. An example from our sample is the sentence in Figure 2; here a conditional clause is introduced by a phrase that roughly translates to 'We'd be glad if...'. Even if we disregard the fact that a full analysis of how this phrase affects the semantics of the sentence would require some model of the speaker's desires – clearly beyond our systems current capabilities – we could still interpret the sentence literally by imposing some rule for conditional sentences, e.g. that given

a structure of the form A if B, the `CAUSE` relation is to hold between the root nodes of B and A. Such arbitrary rules could be introduced for several types of conjunctions in the future. A further, smaller issue is caused by the general lack of personal pronouns in sentences: Hungarian is a *pro-drop* language: if a verb is inflected for person, pronouns need not be present to indicate the subject of the verb, e.g. *Eszem.* 'eat-1SG' is the standard way of saying 'I'm eating' as opposed to *?Én eszem* 'I eat-1G' which is only used in special contexts where emphasis is necessary. Currently this means that `4lang` graphs built from these sentences will have no information about who is doing the `eat`ing, but in the future these cases can be handled by a mechanism that adds a pronoun subject to the graph based on the morphological analysis of the verb. Finally, the lowest quality graphs are caused by very long sentences containing several clauses and causing the parser to make multiple errors.

Örülnénk,	*ha*	*a*	*konzultációs*	*központok*
rejoice-COND-1PL	if	the	consultation-ATT	center-PL
közötti	*kilométerek*	*nem*	*jelentenének*	
between-ATT	kilometer-PL	not	mean-COND-3PL	
az	*emberek*	*közötti*	*távolságot.*	
the	person-PL	between-ATT	distance-ACC	
'We'd be glad if the kilometers between consultation centers did not mean distance between people'				

Fig. 2. Subordinating conjunction

5.2 `dict_to_4lang`

We also conducted manual error analysis on the output of the `dict_to_4lang` pipeline, in this case choosing 20 random words from the EKsz dictionary[7]. The graphs built by `dict_to_4lang` were of very good quality, with only 3 out of 20 containing major errors. This is partly due to the fact that `NSzt` contains many very simple definitions, e.g. 4 of the 20 headwords in our random sample contained a (more common) synonym as its definition. All 3 significant errors are caused by the same pattern: the analysis of possessive constructions by `magyarlanc` involve assigning the `att` dependency to hold between the possessor and the possessed, e.g. the definition of `piff-puff` (see Figure 4) will receive the dependencies `att(hang, kifejezés)` and `att(lövöldözés, hang)`, resulting in the incorrect `4lang` graph in Figure 5

[7] the 20 words, selected once again using `shuf`, are the following: *állomásparancsnok, beköt, biplán, bugás, egyidejűleg, font, főmufti, hajkötő, indikál, lejön, munkásőr, nagyanyó, nemtelen, összehajtogat, piff-puff, szét, tipográfus, túlkiabálás, vakolat, zajszint*

1995	*telén*	*vidrafelmérést*	*végeztünk*
1995	winter-POSS-SUP	otter-survey-ACC	conduct-PST-1PL
az	*országos*	*akció*	*keretében.*
the	country-ATT	action	frame-POSS-INE

'In the winter of 1995 we conducted an otter-survey as part of our national campaign'

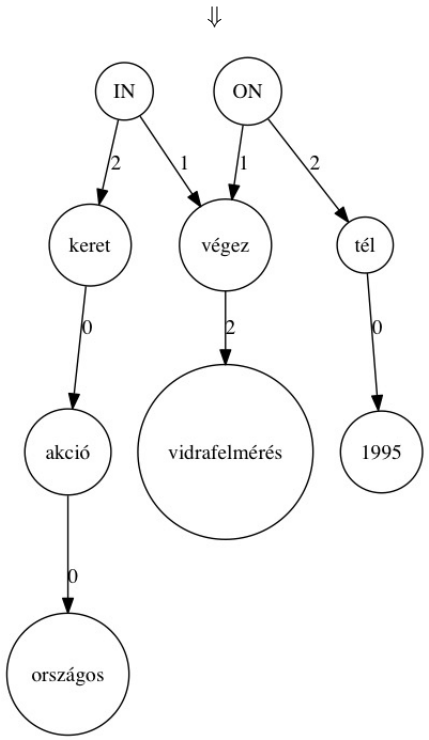


Fig. 3. Example of perfect `dep_to_4lang` transformation

instead of the expected one in Figure 6. `kifejezés` $\xrightarrow{0}$ `hang` $\xrightarrow{0}$`lövöldözés` instead of `kifejezés` $\xleftarrow{2}$ `HAS` $\xrightarrow{1}$ `hang` $\xleftarrow{2}$ `HAS` $\xrightarrow{1}$ `lövöldözés`. These constructions cannot be handled even by taking morphological analysis into account, since possessors are not usually marked (although in some structures they receive the dative suffix *-nak/-nek*, e.g. in embedded possessives like our current example (*hangjának* 'sound-POSS-DAT' is marked by the dative suffix as the possessor of *kifejezésére*). Unless possessive constructions can be identified by `magyarlanc`, we shall require an independent parsing mechanism in the future. The structure of Hungarian noun phrases can be efficiently parsed using the system described in [11], the grammar used there may in the future be incorporated into a `4lang` -internal parser, plans for which are outlined in [12].

Lövöldözés vagy ütlegelés hangjának kifejezésére
Shooting or thrashing sound-POSS-DAT expression-POSS-DAT
'Used to express the sound of shooting or thrashing'

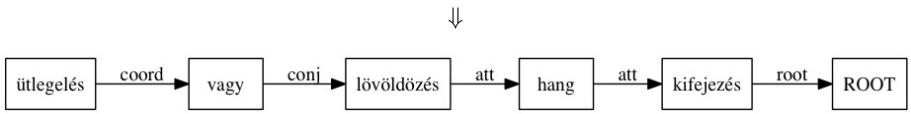


Fig. 4. Dependency parse of the `EKsz` definition of the (onomatopoeic) term `piff-puff`

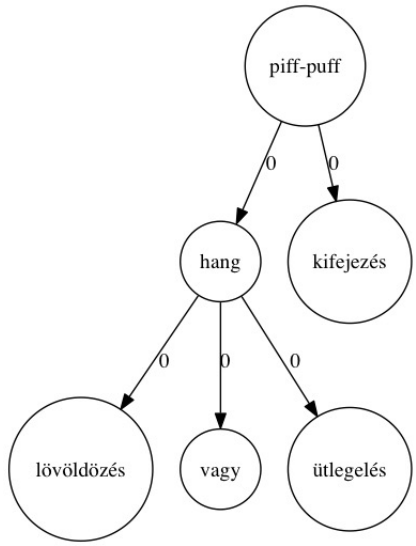


Fig. 5. Incorrect graph for `piff-puff`

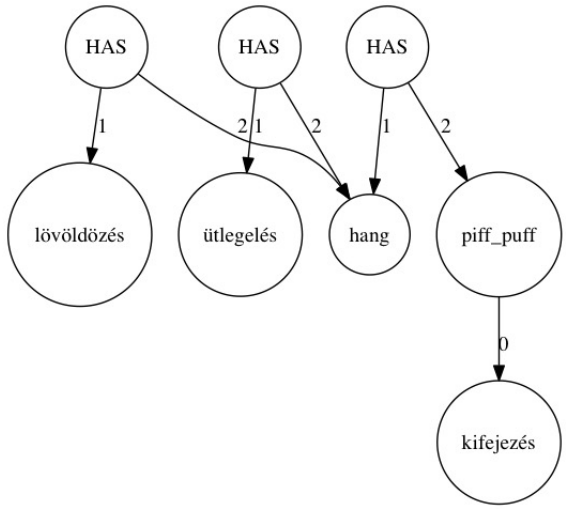


Fig. 6. Expected graph for `piff-puff`

References

1. Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August 2010. Coling 2010 Organizing Committee.
2. Marie-Catherine DeMarneffe, William MacCartney, and Christopher Manning. Generating typed dependency parses from phrase structure parses. In *Proc. LREC*, volume 6, pages 449–454, Genoa, Italy, 2006.
3. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*, pages 203–210, 2004.
4. Nóra Ittzés, editor. *A magyar nyelv nagyszótára III-IV*. Akadémiai Kiadó, 2011.
5. András Kornai. The algebra of lexical semantics. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *Proceedings of the 11th Mathematics of Language Workshop*, LNAI 6149, pages 174–199. Springer, 2010.
6. András Kornai. Eliminating ditransitives. In Ph. de Groote and M-J Nederhof, editors, *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*, LNCS 7395, pages 243–261. Springer, 2012.
7. András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado, June 2015. Association for Computational Linguistics.
8. András Kornai and Márton Makrai. A 4lang fogalmi szótár. In Attila Tanács and Veronika Vincze, editors, *IX. Magyar Számitógépes Nyelvészeti Konferencia*, pages 62–70, 2013.

9. Márton Miháltz. *Semantic resources and their applications in Hungarian natural language processing*. PhD thesis, Pázmány Péter Catholic University, 2010.
10. Ferenc Pusztai, editor. *Magyar értelmező kéziszótár*. Akadémiai Kiadó, 2003.
11. Gábor Recski. Hungarian noun phrase extraction using rule-based and hybrid methods. *Acta Cybernetica*, 21:461–479, 2014.
12. Gábor Recski. *Computational methods in semantics*. PhD thesis, Eötvös Loránd University, Budapest, 2016.
13. Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In Martin Jansche, editor, *Proceedings of the ACL 2005 Software Workshop*, pages 77–85. ACL, Ann Arbor, 2005.
14. Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. Hungarian dependency treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, 2010.
15. János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *Proceedings of RANLP*, pages 763–771, 2013.

Közeli rokonunk, az autó

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat Számos nyelvtechnológiai probléma megoldására sikerrel alkalmazhatóak a különböző statisztikai módszerek. A fogalmak szemantikai reprezentációja esetén azonban még mindig sokszor az előre, kézzel létrehozott lexikai erőforrásokra támaszkodunk. Cikkünkben azt mutatjuk be, hogy a disztribúciós szemantika modelljét alkalmazva, szöveges korpuszból hogyan nyerhetők ki releváns fogalmi csoportok automatikus módszerekkel. Az algoritmust több különböző doménből származó magyar nyelvű részkorpuszra alkalmaztuk. Az eredmények bebizonyították, hogy a módszer alkalmas az általános értelemben kapcsolódó fogalmak csoportosítása mellett a lazább és asszociációs relációk felismerésére is, illetve jól kezeli a különböző domének közötti hangsúlybeli különbségeket is.

1. Bevezetés

A *big data* és a gyakorlatilag végtelen kapacitású számítógépek korában a nyelvtechnológiai alkalmazások is egyre inkább a statisztikai módszerekhez nyúlnak. Ennek ellenére, a világismeret és a szemantikai relációk ábrázolása általában még mindig kézzel készült lexikai erőforrások (pl. WordNet[4,12]) és doménspecifikus tezauruszok segítségével történik.

Zhang [19] tanulmányában kimutatta, hogy gyakran a fogalmak ilyen mesterséges rendszerezése és a kognitív emberi tudásreprezentáció között jelentős különbség van. Ezért a tudásábrázolás létrehozásakor érdemes lehet a szöveges adatokból kiindulni, ahelyett, hogy ezeket próbálnánk egy előredefiniált fogalomrendszerhez való illeszkedésre kényszeríteni. Ráadásul, kisebb nyelvek esetén, mint a magyar, jóval kevesebb és kisebb lefedettségű lexikai erőforrás áll rendelkezésre, mint a nagyobb nyelveknél [9].

Jelen cikkünkben olyan kísérletekről számolunk be, amelyek azonos fogalmak különböző jellegű szövegekben való használatát vizsgálják. Olyan statisztikai módszereket alkalmaztunk, melyek magyar szövegekből kinyert fogalmakat és kifejezéseket szemantikai csoportokba sorolnak az adott doménen belül való disztribúciós viselkedésük alapján.

A WordNet jellegű erőforrásokban a fogalmakat szinonimahalmazok (synsetek) reprezentálják, melyek az azonos jelentésű szavakat foglalják össze. Ezek között a halmazok között állhatnak fenn explicit relációk. Disztribúciós modellek használatával ezeket a relációkat nem tudjuk automatikusan azonosítani, és

az automatikusan létrejövő szóhalmazok is tartalmaznak oda nem illeszkedő szóalakokat. Ugyanakkor az eredmények azt mutatják, hogy a modell valóban összetartozó kifejezéseket ismer fel, csupán az összetartozás szemantikai típusa tér el akár egy csoporton belül. A kinyert hasonlóságok jellege paradigmatikus, azaz a hasonló kifejezések egymással felcserélhetőek, de ez épp úgy igaz lehet szinonimákra, hipernimákra, hiponimákra és akár antonimákra is. Ezek megkülönböztetésére nem alkalmas a módszer. Ennek ellenére a létrejövő fogalmi csoportok relevánsak. Sőt, azok az asszociációs relációk, melyek a modell alkalmazásával felismerhetők, gyakran hiányoznak a klasszikus ontológiákból. Továbbá, mivel az alkalmazott módszerek statisztikai alapúak, nem felügyelt módszerekkel tanuló algoritmusokkal jönnek létre, ezért könnyen adaptálhatók tetszőleges doménre és nyelvre, ami a kézzel készített statikus erőforrásokról nem mondható el.

2. Kapcsolódó munkák

Számos módszer létezik szöveges korpuszokból hasonló szavak csoportjainak kinyerésére. A statisztikai módszerek általában a disztribúciós szemantika elméletét használják ki, azaz a szavak jelentését a környezetükben való előfordulásukhoz kötik. A különbség az egyes módszerek között leginkább a környezet definíciója. Egy lehetséges ábrázolás az egyes szavakhoz a függőségi relációk mentén kapcsolódó szavak használata környezetként, melyre több példát találunk a szakirodalomban (néhány ezek közül: [8], [7], [15] és [13]). Ezek alapja azonban egy jó minőségű függőségi elemző, vagy egy kézzel elemzett szöveges korpusz, amik közül gyakran egyik sem érhető el bizonyos nyelvekre. Más megvalósítások együttes előfordulások gyakorisági értékeiből hozzák létre a vektortérmodelleket és ezeket valamilyen vektorhasonlósági mérték alapján hasonlítják össze [16,3]. Napjaink egyre elterjedtebb módszere pedig a neurális hálózatok alkalmazása, amik egy kellően nagy szöveges korpuszból tanult folytonos vektorreprezentációt rendelnek az egyes szavakhoz, illetve a szavak jelentéséhez [10,11].

A jelen cikkben bemutatott megoldás abban különbözik a fentiektől, hogy a klaszterezés alapjául szolgáló, az egyes szavakat és kifejezéseket reprezentáló vektorok létrehozása során nem támaszkodtunk a szófaji egyértelműsítésnél mélyebb nyelvtani elemzésre, viszont nem is csupán a szavak együttes előfordulásainak statisztikáját vettük figyelembe.

3. Módszer

Az algoritmus három fő lépésből áll. Először a többszavas kifejezéseket összevonjuk a korpuszban. Ezeket aztán a későbbi lépések egyetlen egységként kezelik. A második lépés során létrehozzuk a disztribúciós modellt, azaz minden szópárhoz kiszámoljuk a disztribúciós hasonlóságuk mértékét. A harmadik lépésben ezeket a számértékeket használva jellemzőkként, minden kifejezéshez egy jellemzővektort hozunk létre, melyeket végül hierarchikus klaszterezéssel rendszerbe szervezünk. Ebből aztán tetszőleges sűrűséggel emelhetünk ki összetartozó kifejezéseket tartalmazó klasztereket.

3.1. Többszavas kifejezések

Mind az általános, mind a doménspecifikus nyelvhasználatban előfordulnak olyan kifejezések, amik több szóval írnak le egyetlen fogalmat. Függetlenül attól, hogy ezek szerkezete mennyire kompozicionális, önálló egységként kezelhetjük őket. Munkánk során a c-value algoritmus [6] módosított változatát használtuk a többszavas kifejezések azonosítására. Ennek bemeneteként szófaji egyértelműsítésen átesett szöveget használtunk, kimenetként pedig a többszavas kifejezések listáját kapjuk, a hozzájuk rendelt c-value értékek szerint sorba rendezve. Minél előrébb szerepel tehát egy kifejezés ebben a listában (azaz minél magasabb c-value értéket kapott), annál inkább tekinthető valódi többszavas kifejezésnek.

Az algoritmus menete a következő. Először kigyűjtjük a korpuszból az összes lehetséges n-grammot, ahol n értéke 1 és k közé esik (k tetszőlegesen választható, esetünkben $k = 20$). Ezután egy nyelvi szűrőt és egy stopword szűrőt alkalmaztunk, majd a szűrés után megmaradt n-grammokhoz meghatároztuk a korpuszbeli gyakoriságukat. Végül, a leghosszabbtól a legrövidebb n-grammok felé haladva kiszámítottuk a hozzájuk tartozó c-value értéket. Az algoritmus részletei megtalálhatók [17]-ben és [6]-ben.

A c-value érték meghatározása a korpuszból számított statisztikákra alapul, ezért nincs szükség külső lexikai erőforrások használatára a többszavas kifejezések megállapításához. Azonban a nyelvi szűrő kézzel definiált nyelvspecifikus szabályokat tartalmaz, annak biztosítására, hogy a kinyert kifejezések helyes kifejezések legyenek. Esetünkben ez a magyar főnévi szerkezetekre korlátozta a számításba jövő kifejezéseket, mivel kísérleteink során csak nominális fogalmakkal foglalkoztunk. Mivel a későbbi lépések során továbbra is szükségünk volt a szófaji címkékre, ezért az eredményként kapott összevont kifejezések a kifejezés fejének szófaji címkéjét kapták, megtartva ezáltal a szintaktikai szerepüket az adott kontextusban.

3.2. Disztribúciós hasonlóság

A releváns kifejezések csoportosításához szükség van egy hasonlósági metrikára is, ami két kifejezés jelentésbeli távolságát határozza meg. Erre szintén olyan nem felügyelt módszert alkalmaztunk, amely a hasonlóságokat nem egy külső erőforrás, ontológia alapján határozza meg, hanem a kifejezések korpuszbeli előfordulásai, az adott korpuszban való használatuk alapján.

A disztribúciós szemantika lényege, hogy a szemantikailag hasonló szavak hasonló környezetben fordulnak elő [5]. Tehát két szó jelentésének hasonlósága meghatározható a környezetük hasonlósága alapján. A szavak környezetét olyan jellemzőhalmazokkal reprezentáltuk, ahol minden jellemző egy relációból (r) és az adott reláció által meghatározott szóból (w') áll [8]. Ezek a relációk más alkalmazásokban általában függőségi relációk, mi azonban a függőségi elemző alkalmazásától most eltekintettünk. Carrol és tsai. [1] csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt. Mivel a mi esetünkben a morfológiai elemzés is rendelkezésre állt, ezért a következő jellemzőket vettük figyelembe:

- prev_1: a szót megelőző szó lemmája
- prev_w: a szó előtt 2-4 távolságon belül eső szavak lemmái
- next_1: a rákövetkező szó lemmája
- next_w: a szó után 2-4 távolságon belül eső szavak lemmái
- pos: a szó szófaja
- prev_pos: a szót megelőző szó szófaja
- next_pos: a szót követő szó szófaja

Szavak alatt pedig a lemmatizált szóalakot értjük a relációk mindkét oldalán.

Minden egyes jellemzőhöz meghatároztuk a korpuszbeli gyakoriságát. Ezekből a gyakoriságokból határozható meg a (w, r, w') hármas információtartalma $(I(w, r, w'))$ maximum likelihood becsléssel a következő képlettel:

$$I(w, r, w') = log \frac{||w, r, w'|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||}$$

Mivel $||w, r, w'||$ a (w, r, w') hármas korpuszbeli gyakoriságának felel meg, ezért ha a hármas bármelyik tagja $*$, akkor a hármas többi tagjára illeszkedő összes hármas gyakoriságának az összegével számolunk. Például a $||*, next\_1, ember||$ megfelel az olyan szavak gyakoriságának az összege, amit az *ember* szó követ.

Ezután a két szó (w_1 és w_2) közötti hasonlóságot a következő metrikával számoltuk [8] alapján:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

ahol $T(w)$ azoknak az (r, w') pároknak a halmaza, ahol az $I(w, r, w')$ pozitív.

Bár a modell a korpuszban szereplő összes szóra alkalmazható, érdemes szófajonkénti modelleket építeni. Munkánk során csupán főnevekkel és nominális kifejezésekkel végeztünk kísérleteket, melyek legalább ötször előfordulnak a felhasznált korpuszban.

3.3. Hierarchikus klaszterezés

A szavak és kifejezések páronkénti hasonlóságából kiindulva fogalmi hierarchiát határozhatunk meg. Ehhez a leggyakoribb kifejezések és szavak csoportján agglomeratív klaszterezést hajtottunk végre. A klaszterező algoritmus megválasztásakor [14] érvelését vettük figyelembe, miszerint az írott szövegek kifinomult változatossága miatt a használt fogalmak csoportjainak száma előre nem megjósolható. Egy hierarchikus szerveződés azonban alkalmas arra, hogy az aktuális szövegre jellemző önálló, kompakt fogalmi csoportokat előre megfogalmazott általánosítás helyett az aktuális eredmény alapján nyerjük ki minden egyes szöveg esetén.

A legtöbb vektortéralapú módszer a fogalmak csoportosításakor azok együttes előfordulását veszi figyelembe, az egyes kifejezéseket leíró vektorok ilyen jellemzőket tartalmaznak. Ezek a megközelítések azonban alkalmatlanok arra, hogy olyan fogalmak között is felfedezzék a hasonlóságot, amik sosem fordulnak elő együtt, annak ellenére, hogy gyakran éppen az ilyen szavak azok, amik használatukat tekintve hasonlóak. Ezért az egyes kifejezéseket a többi kifejezéshez való hasonlóságukból álló jellemzővektorokkal ábrázoltuk. Így az egy kifejezéshez tartozó $c(w)$ vektor c_i eleme $SIM(w, w_i)$. Az egyes kifejezésekhez így létrehozott jellemzővektorokat klaszteresztük, ahol a klaszterek távolságát Ward ([18]) módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre.

Célunk azonban nem egy bináris faként ábrázolt teljes hierarchia meghatározása volt, hanem a fogalmak elkülönülő csoportjainak meghatározása, azaz a kapott dendrogram egyes kompakt részfái. Ezeket úgy kaphatjuk meg, ha a vágási pontokat a klaszterezés szintjei között lévő nagy ugrásoknál határozzuk meg. Formálisan ez úgy határozható meg, hogy a teljes fában lévő minden egyes részfát összekötő link magasságát összehasonlítjuk az alatta lévő szomszédos linkek magasságával egy adott mélységig. Ha ezek különbsége nagyobb, mint egy előre meghatározott küszöbérték (azaz a link inkonzisztens), akkor a vizsgált csomópont egy vágási pont. A teljes fából tehát az így meghatározott pontok alatti részfák levelei (azaz a szavak és kifejezések) egy csoportot alkotnak. Ezeknek a csoportoknak a sűrűsége a linkinkonzisztencia-küszöbérték változtatásával dinamikusan állítható.

4. Kísérletek

Kísérleteink során a Szeged Korpusz [2] 11 részkorpuszát használtuk a 1. táblázat szerint.

1. táblázat. A kísérletek során használt részkorpuszok, azok jellege és mérete

Név	**domén**	**Méret (token)**
10elb	diákfogalmazás	126841
8oelb	diákfogalmazás	92625
1984	szépirodalom	96843
utas	szépirodalom	75932
pfred	szépirodalom	60651
gazdtar	jog	153430
szerzj	jog	100153
newsml	rövid üzleti hírek	211742
mh	újság	49162
np	újság	74479
win2000	számítástechnika	66242

Minden egyes részkorpuszra egyesével alkalmaztuk a fenti algoritmust, azaz összevontuk az adott szövegtípusra jellemző többszavas kifejezéseket, ezután ezek hasonlóságát meghatároztuk, majd létrehoztuk a szemantikus csoportokat. Bár a Szeged Korpusz egyes részeinek mérete nem túl nagy, ezért az ezeken tanított statisztikai modellek kevésbé megbízhatóak, az összevont korpusszal is végeztünk kísérleteket, azonban ekkor sokkal kevésbé koherens csoportokat kaptunk eredményül.

Az alkalmazott algoritmus másik fő paramétere, ami hatással van az eredményként kapott klaszterek minőségére, a dendrogram vágási pontjait meghatározó inkonzisztenciaérték. Ezt az egyes részkorpuszoknál külön-külön állítottuk be a megfelelő eredmény elérése érdekében. Ennek a beállítása azonban függ az eredményül kapott csoportok felhasználási módjától, illetve a további feldolgozással kapcsolatos elvárásoktól. Ha nagyobb fedést szeretnénk elérni, akkor a küszöbérték magasabbra állítható, így nagyobb, de kevésbé tömör csoportokat kapunk. Ha azonban inkább kisebb, de szorosabban kapcsolódó kifejezéseket tartalmazó csoportokra van szükség, akkor a magasabb pontosság eléréséhez alacsonyabb küszöbértéket használhatunk. Mivel jelen munkánk során végzett kísérleteink csupán a módszer alkalmazhatóságát vizsgálták, a csoportok további feldolgozása nem volt meghatározva, ezért a küszöbérték beállítása empirikusan, a pontosság és a fedés közötti egyensúlyra törekedve történt minden egyes részkorpuszra.

Módszerünk eredményességét több síkon vizsgáltuk. Az egyik szempont a módszer domének közötti különbségekre való érzékenysége. Ehhez az eredményül kapott klasztereket az egyes részkorpuszokra páronként összehasonlítva egy kereszthasonlóság-értéket határoztunk meg minden párhoz. A részkorpuszok minden egyes lehetséges párosításához összegyűjtöttük azokat a szavakat és kifejezéseket, amik mindkét korpuszban előfordultak. Az ezeket a kifejezéseket tartalmazó csoportokra megvizsgáltuk a két részkorpusz esetén a csoportok metszetét és különbségét. A kapott halmazok méretének kumulált arányát a következő képlettel határoztuk meg:

$$\sum_{w} \frac{\|clust_A \bigcap clust_B\|}{\|clust_A\| + \|clust_B\| - \|clust_A \bigcap clust_B\|}$$

ahol $w \in (text_A \bigcap text_B)$ és $clust_A$ és $clust_B$ a két vizsgált részkorpusz klaszterei, amikben a w kifejezés előfordul. A 2. táblázat tartalmazza az így kapott kereszthasonlóság érték szerint rendezett lista első és utolsó 5 elemét.

Ahogy az eredményekből látszik, az egymáshoz közelebb álló domének szerepelnek a lista elején, míg az egymástól távolabbi párok kerültek a rangsor végére. A párok közötti eltérés nem csak a bennük előforduló különböző szavakból fakad (a `mh` és a `newsml` pár metszetében szereplő szavak száma közel azonos a `10elb` és a `gazdtar` pár metszetében szereplő szavak számával, mégis az előbbi pár a lista elején szerepel, míg az utóbbi a végén), hanem az azonos szavak különböző használatából is. Ezekre a különbségekre nem derülhet fény, egy előre definiált erőforrás alapján, hiszen abban nem különböztetjük meg a különböző típusú szövegekben megjelenő kisebb jelentésbeli vagy hangsúlybeli különbségeket.

2. táblázat. A kereszthasonlóság (KH) vizsgálatának eredményei

Pár	**KH**	**domén**
10elb–8oelb	10.880	diákfogalmazás–diákfogalmazás
newsml–np	4.586	hír–újság
mh–newsml	4.500	újság–hír
mh–np	3.672	újság–újság
gazdtar–szerzj	3.513	jog–jog
10elb–np	2.223	diákfogalmazás–újság
1984–pfred	2.179	szépirodalom–szépirodalom
8oelb–utas	2.057	diákfogalmazás–szépirodalom
10elb–newsml	1.917	diákfogalmazás–hír
...	...	...
8oelb–szerzj	0.321	diákfogalmazás–jog
mh–pfred	0.321	hír–szépirodalom
pfred–szerzj	0.222	szépirodalom–jog
gazdtar–utas	0.192	jog–szépirodalom
10elb–gazdtar	0.182	diákfogalmazás–jog
pfred–win2000	0.154	szépirodalom–számítástechnika
8oelb–win2000	0.000	diákfogalmazás–számítástechnika
gazdtar–pfred	0.000	jog–szépirodalom
utas–win2000	0.000	szépirodalom–számítástechnika

5. Eredmények

Az eredményül kapott fogalmi klaszterek különböző szempont szerinti csoportosításokat eredményeztek. Néhány csoportban általános vagy klasszikus értelemben kapcsolódó kifejezések gyűltek össze, mint például testrészek (ezek elsősorban szépirodalmi szövegekben jelentek meg, ahol az egyes szereplők leírása részletesebb), napok és hónapok nevei (elsősorban a hír és a diákfogalmazás részkorpuszokban) vagy pénznemek (a gazdasági és hírkorpuszokban). Bár ezek az általános csoportok akár előre is meghatározhatók, nincs garancia arra, hogy nem jelenik meg egy olyan kifejezés egy adott szövegben, ami eredetileg nem lenne benne az előre definiált listákban, így ezeket is érdemesebb az adott szövegből kinyerni. Továbbá a kinyert csoportok nem tartalmaznak olyan szavakat és kifejezéseket, amik az adott szövegben nem szerepelnek, így az eltárolandó eredmény mérete sem haladja meg azt, amire feltétlenül szükség van.

A létrejött csoportok egy másik típusa valamilyen nyelvtani szempont szerinti rendeződés alapján jött létre, mint például a funkcióigés szerkezetek főnévi magját alkotó elemek.

A harmadik fő típusba pedig olyan csoportok sorolhatók, amikben a szavak valamilyen tágabb értelemben kapcsolódnak, leginkább az adott részkorpuszra jellemző használatuk alapján. Néhány ilyen példát láthatunk a 3. táblázatban.

Ahogy a példákon is látszik, az alkalmazott algoritmus sokszor valamilyen asszociációs kapcsolatban álló kifejezéseket csoportosított össze, különösen a diákfogalmazás és a szépirodalmi részkorpuszok esetén. Például a *erdő, falu, város,*

3. táblázat. Néhány példa az eredményül kapott csoportokra az egyes részkorpuszokból

Text	**cluster**
gazdtar	*vezető tisztségviselő, könyvvizsgáló, személy, igazgatóság, ügyvezető, igazgató*
gazdtar	*társasági szerződés, alapító okirat, alapszabály*
10elb	*erdő, falu, város, ház, diszkó, part*
10elb	*cucc, táska, csomag, holmi*
1984	*ujj, test, arc, szem, fej, kar, kéz, tömeg, agy, száj, láb*
1984	*férfi, asszony, pillanat, hang, telekép, lány, ember, pont, Mr., éves kor*
1984	*lázadás, szokás, remény, napló, hit, dátum*
8oelb	*öröm, élmény, irány, nyaralás, történet, délután*
newsml	*költség, kiadás, díj, adósság, befektetés, eszköz*
newsml	*fél, egész, arány, időszak*
szerzj win	*fejezet, cikk, pont, törvény, §, bekezdés* *NTFS, állományrendszer, helyfoglalási egység, adat, lemez, logikai lemez, kötet, merevlemez, fizikai lemez*

ház, diszkó, part csoportban a kifejezések páronkénti kapcsolata nem feltétlenül megjósolható (pl. az *erdő* és *diszkó* pár esetén), de ismerve a részkorpuszt (diákok által írt szövegek), illetve a csoportba sorolt többi szót, már könnyen belátható, hogy a csoportosításnak van értelme, a kifejezések valóban kapcsolódnak egymáshoz. Egy másik jellemzője az alkalmazott algoritmusnak, hogy könnyen alkalmazkodik a doménspecifikus, vagy akár teljesen egyedi szóhasználathoz is. Például a diákfogalmazásokra jellemző szleng is megfelelően csoportosítható. Ezeket a szóalakokat szinte lehetetlen egy előre definiált kategóriarendszerbe besorolni, hiszen nagyon gyorsan jelennek meg, vagy tűnnek el a nyelvből, esetleg átalakul a jelentésük. Egy másik példa a szépirodalmi szövegekből alkotott csoportosítások esetén látható, különösen George Orwell *1984* című regénye esetén. Ez a korpusz rengeteg sajátos szóalakot tartalmaz, amik csupán a szerző által kitalált, a valóságban nem létező, vagy nem az ebben a műben használt értelemben használt kifejezések, az alkalmazott algoritmus azonban ezeket is helyesen tudta csoportosítani, a ténylegesen létező szavakkal együtt az általánostól esetlegesen eltérő, éppen megfelelő jelentésük szerint (pl. *lázadás, szokás, remény, napló, hit, dátum*).

Az eredmények vizsgálata azonban nem csak az egyes részkorpuszok esetén érdekes, hanem a létrejött csoportosítások metszetét és különbségeit is érdemes elemezni. Például az *autó* szó több részkorpuszban is a családtagokat leíró csoportba került besorolásra. Szigorúan szemantikai szempontból ennek a relációnak

nincs értelme, ugyanakkor a valóságban gyakran tényleg létező jelenség az autóra mint családtagra való utalás. A diákfogalmazások esetén pedig még a *bicikli* szó is ebbe a csoportba került, ami hasonlóan magyarázható. Megfigyelhetőek továbbá a különböző domének közötti apró eltolódások is a szóhasználatot illetően. Például a 8. osztályos diákok által írt fogalmazásokban a *szülő* és a *barát* szavak még egy csoportba kerültek, azonban a tizedikes diákok által írt fogalmazásokban ez a két szó már elválik, ami jól tükrözi a gyerek-szülő viszony eltolódását ennél a korosztálynál.

6. Konklúzió

Jelen cikkünkben olyan kísérletekről számoltunk be, amelyek azonos fogalmak különböző jellegű szövegekben való használatát vizsgálják. Ehhez eszközül a disztribúciós szemantika egy modelljét alkalmaztuk. A többszavas kifejezések meghatározása után minden szót/kifejezést a többi szóhoz való hasonlóságát tartalmazó vektorral ábrázoltunk (ahol a páronkénti hasonlóság számítása a kölcsönös információtartalmon alapult [8]). Az így kapott vektorokat pedig hierarchikus klaszterezéssel tömör, koherens csoportokba osztályoztuk. Az eredményül kapott csoportok tehát olyan kifejezéseket és szavakat tartalmaznak, amelyek használatuk szempontjából hasonlóak.

A fenti algoritmust a Szeged Korpusz [2] egyes részkorpuszaira külön-külön alkalmaztuk. Az eredmények elemzésekor pedig azt vizsgáltuk, hogy ugyanazon kifejezések disztribúciós viselkedése hogyan változik különböző domének esetén. Így olyan kifinomult különbségekre is fény derült, melyek semmilyen formális ontológiában vagy fogalmi rendszerben nem ábrázolhatóak.

A módszerünk ellenőrzéseként definiáltunk egy olyan metrikát, ami a különböző doménekből létrejött csoportok közötti átfedés mértékét vizsgálja. Ezzel kimutattuk, hogy a hasonló jellegű (gazdasági-jogi, sajtónyelvi, szépirodalmi, iskolai) szövegekből épített fogalmi csoportok nagyobb átfedést mutattak, mint a különböző domének fogalmi csoportjai.

Hivatkozások

1. Carroll, J., Koeling, R., Puri, S.: Lexical acquisition for clinical text mining using distributional similarity. In: Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II. pp. 232–246. CICLing'12, Springer-Verlag, Berlin, Heidelberg (2012)
2. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P., Kopecek, I., Pala, K. (eds.) TSD. Lecture Notes in Computer Science, vol. 3206, pp. 41–48. Springer (2004)
3. de Cruys, T..: Semantic clustering in Dutch. In: Proceedings 16th Meeting of Computational Linguistics in the Netherlands. pp. 19–31 (2005)
4. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
5. Firth, J.R.: A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis pp. 1–32 (1957)

6. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries 3(2), 115–130 (August 2000)
7. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the 28th Annual Meeting on Association for Computational Linguistics. pp. 268–275. ACL '90, Association for Computational Linguistics, Stroudsburg, PA, USA (1990), http://dx.doi.org/10.3115/981823.981857
8. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2. pp. 768–774. COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998)
9. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
11. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://www.aclweb.org/anthology/N13-1090
12. Miller, G.A.: WordNet: A Lexical Database for English. COMMUNICATIONS OF THE ACM 38, 39–41 (1995)
13. Padó, S., Lapata, M.: Dependency-based construction of semantic space models. Comput. Linguist. 33(2), 161–199 (Jun 2007), http://dx.doi.org/10.1162/coli.2007.33.2.161
14. Pereira, F., Tishby, N., Lee, L.: Distributional Clustering of English Words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. pp. 183–190. ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993), http://dx.doi.org/10.3115/981574.981598
15. Ruge, G.: Experiment on linguistically-based term associations. Inf. Process. Manage. 28(3), 317–332 (Jan 1992), http://dx.doi.org/10.1016/0306-4573(92)90078-E
16. Senellart, P., Blondel, V.: Automatic discovery of similar words. In: Berry, M. (ed.) Survey of Text Mining. Springer-Verlag (2003)
17. Siklósi, B., Novák, A.: Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods, Lecture Notes in Artificial Intelligence, vol. 8791, pp. 233–243. Springer International Publishing, Heidelberg (2014)
18. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963), http://www.jstor.org/stable/2282967
19. Zhang, J.: Representations of health concepts: a cognitive perspective. Journal of Biomedical Informatics 35(1), 17 – 24 (2002), http://www.sciencedirect.com/science/article/pii/S1532046402000035

Gépi fordítás minőségbecslésének optimalizálása kétnyelvű szótár és WordNet segítségével

Yang Zijian Győző[1], Laki László[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
e-mail:{yang.zijian.gyozo, laki.laszlo}@itk.ppke.hu

Kivonat Napjainkban, a gépi fordítás minőségének becslése fontos feladat. Egy megbízható minőségbecslő rendszer időt és pénzt spórolhat meg cégek, kutatók és átlagfelhasználók számára. A hagyományos automatikus kiértékelő módszerek legnagyobb problémája, hogy referenciafordítást igényelnek és nem tudnak valós időben kiértékelni. A jelen kutatás egy olyan minőségbecslő rendszert mutat be, amely képes valós időben, referenciafordítás nélkül kiértékelni. A minőségbecslő rendszer felépítéséhez a QuEst keretrendszert implementáltuk és optimalizáltuk magyar nyelvre. Mindezek mellett, a QuEst rendszerhez új, saját jegyeket fejlesztettünk egy kétnyelvű szótár, illetve a WordNet segítségével. A saját jegyek alkalmazása minőségbeli javulást eredményezett a kiértékelésben. Az így létrehozott magyar nyelvre optimalizált jegyhalmaz 11%-kal jobb eredményt ad az alaprendszerhez képest. Az általunk implementált minőségbecslő rendszer megfelelő alapot képez egy angol-magyar gépi fordítást kiértékelő rendszerhez.

Kulcsszavak: minőségbecslés, gépi fordítás, kiértékelés

1. Bevezetés

A gépi fordítás használata mára széles körben elterjedt a hétköznapokban, azonban a létező rendszerek között, a fordítási minőségében jelentős különbségek mutatkoznak. Ezért egyre több helyen merül fel igényként a gépi fordítás minőségének becslése. Cégek esetében igen nagy segítséget nyújt egy minőségi mutató, ami nemcsak a gépi fordítás utómunkáját végző szakemberek munkáját támogatja és gyorsíthatja, hanem segíti a fordítócégeket a költségeik csökkentésében is. Másik alkalmazási területe, egy minőségi mérőszám létrehozása a gépi fordítórendszerek kombinációjához. Megfelelő minőségbecsléssel több gépi fordítást össze tudunk hasonlítani és a jobb fordítást kiválasztva javíthatjuk a rendszerünk végső minőségét. Végül, de nem utolsó sorban, ha ismerjük a fordítás minőségét, ki tudjuk szűrni a használhatatlan fordításokat, illetve figyelmeztetni tudjuk a végfelhasználót a megbízhatatlan szövegrészletekre.

A gépi fordítás minőségének automatikus mérése nem könnyű feladat. A hagyományos módszerek legnagyobb problémája, hogy referenciafordítást igényelnek, amelynek létrehozása igen drága és időigényes, ezért ezek a módszerek nem

alkalmasak valós idejű használatra. Másik nagy problémája, hogy mivel ember által fordított referenciafordítás alapján értékelnek, a minőségbecslés minősége jelentős mértékben függ a referenciafordítás minőségétől. Az elmúlt évek kutatásai azt bizonyítják, hogy a hagyományos módszerek kiértékelései alacsonyan korrelálnak az emberi kiértékelésekkel [1,2].

A kutatásunk során, a hagyományos kiértékelő módszerek problémáira keresünk megoldást. Létezik egy másik kiértékelő módszer, amit minőségbecslésnek hívnak. A minőségbecslő módszer nem igényel referenciafordítást, ezért valós időben is alkalmazható és magasan korrelál az emberi kiértékeléssel. A kiértékelt minőségi mutatók a fordítás pontosságára, a mondatok helyességére és egyéb nyelvi problémákra tud megoldást nyújtani, melyekre a hagyományos kiértékelő módszerek, mint a BLEU [3] vagy a NIST [4] nem képesek.

2. Kapcsolódó munkák

Az elmúlt évek során több WMT workshopot[3] rendeztek minőségbecslés témájában, különböző párhuzamos annotált korpuszokat biztosítva a kutatók számára. A korpuszokat szakértők értékelték ki HTER, METEOR vagy utómunka ráfordítás szempontja alapján. Magyar nyelvre azonban nem létezik korpusz, ezért készítettünk egy saját kiértékelt angol-magyar párhuzamos korpuszt.

A minőségbecslés témájában két fő irányban folynak kutatások. Az egyik irány az új releváns minőségi mutatók felfedezése [5], a másik irány a minőségi mutatók optimalizálása gépi tanulás módszerek kísérletezésével [6,7]. A kutatásunk során mindkét területre fókuszálunk.

Korábbi cikkünkben [8] bemutattunk egy működő minőségbecslő rendszert angol-magyar nyelvre. A jelen kutatás a felépített rendszer hibáira keres megoldásokat, illetve további jegyeket tár fel, amelyek javítják a kiértékelő rendszer minőségét. Az előző cikkben felépített rendszer tanítóhalmaza 500 mondatpárral dolgozik, amelyek közül némelyik mondatot csak egy ember értékelt ki, valamint a bemutatott eredmények nem keresztvalidációval készültek. Ezzel szemben a mostani rendszert 600 mondatpárral tanítottuk, amiket legalább három ember értékelt ki. Továbbá, a kiértékelést keresztvalidálással végeztük.

3. A gépi fordítás minőségének becslése

A referenciafordítás nélküli minőségbecslő modell (lásd 1. ábra), a forrásnyelvi és a gép által lefordított mondatokból különböző nyelvfüggetlen és nyelvspecifikus minőségi mutatószámokat nyer ki. A minőségi mutatókat különböző jegyek (feature) segítségével nyerjük ki a forrás- és a gép által lefordított mondatokból. A jegyeket négy csoportba sorolhatjuk:

1. Komplexitással kapcsolatos jegyek: forrásmondatokból kinyert minőségi mutatók.

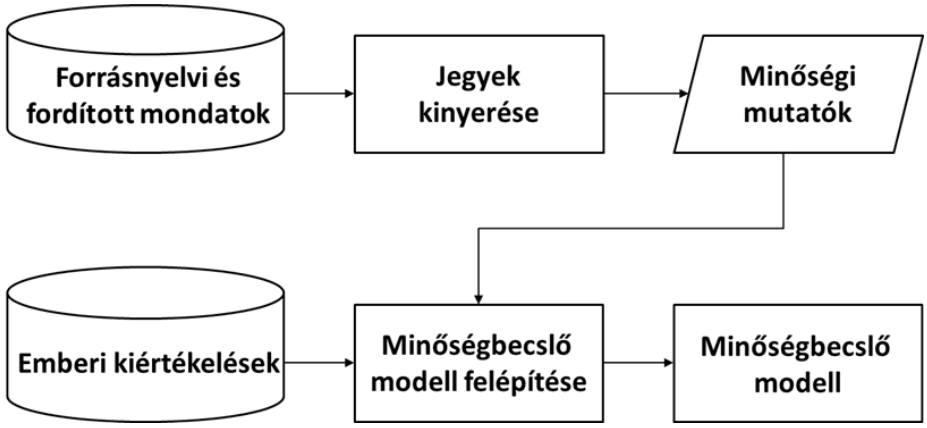
[3] http://www.statmt.org/wmt15/quality-estimation-task.html

2. Helyességgel kapcsolatos jegyek: fordított mondatokból kinyert minőségi mutatók.
3. Megfeleléssel kapcsolatos jegyek: forrásnyelvi és fordított mondatok közötti viszonyából számított minőségi mutatók.
4. Megbízhatósággal kapcsolatos jegyek: gépi fordítórendszerből kinyert minőségi mutatók.

Egy másik szempont alapján, két kategóriába sorolhatjuk a jegyeket:

1. „Black-box" jegyek: gépi fordítórendszertől független jegyek.
2. „Glass-box" jegyek: gépi fordítórendszerből kinyert jegyek.

A jegyek kinyerése után, a minőségi mutatókat egy gépi tanuló algoritmussal betanítjuk. A tanítás után a létrejött modell segítségével tudunk új, ismeretlen mondatokat kiértékelni valós időben. A célunk, hogy a minőségbecslő modell kiértékelése magasan korreláljon az emberi kiértékeléssel, ezért a tanulófázisban a minőségi mutatókat emberi kiértékeléseken tanítjuk be. A kutatásunkban csak „black-box" jegyeket használtunk, mivel több, egymástól független gépi fordítórendszert (amelyek belső paraméterei nem érhetőek el számunkra) is alkalmaztunk a tanítókorpusz létrehozásához.



1. ábra. Minőségbecslő modell működése

4. A HuQ Korpusz

A minőségbecslő modell felépítéséhez szükség van egy annotált párhuzamos korpuszra. A kutatásunk során a minőségbecsléshez létrehoztunk egy tanítókorpuszt, a HuQ (Hungarian Quality Estimation) korpuszt. A korpuszt úgy hoztuk létre, hogy angol mondatokat lefordítottuk különböző gépi fordítórendszerekkel. A fordításokat az alábbi rendszerekkel végeztük el: MetaMorpho [9] szabályalapú

gépi fordítórendszer, Google translate, Bing translator és MOSES [10] statisztikai gépi fordító keretrendszer. Annak érdekében, hogy a rendszer tanuljon a jó fordításokból is, a korpusz a Hunglish [11] korpuszból vett emberi fordításokat is tartalmazza. Ilyen módon, 1950 fordított mondatpár jött létre. Majd az így elkészült korpuszt (C1 korpusz) emberekkel kiértékeltettük. A korpusz jelenleg 600, ember által kiértékelt fordítást tartalmaz, de folyamatosan bővül. Az emberi értékelés pontszámainak létrehozásához készítettünk egy weboldalon elérhető kérdőívet[4]. A kiértékeléshez önkénteseket kértünk fel, akik közép-, illetve felsőfokú angoltudással rendelkeznek. A fordításokat két szempont alapján lehetett értékelni: *megfelelés* és *helyesség*. A megfeleléssel azt mértük, hogy a lefordított mondat tartalmilag mennyire felel meg a forrásnyelvi mondat mondanivalójának. A helyességgel azt mértük, hogy a lefordított mondat szerkezetileg és nyelvtanilag mennyire helyes. A minőséget 1-től 5-ig terjedő skálán osztályoztuk (lásd 1. táblázat). Ez a kiértékelési technika elterjedt a fordításkiértékelés területén [12]. A rosszul, vagy egyáltalán nem értelmezhető forrásoldali mondatok kiszűrésére, bevezettünk egy 0 („nem tudom értelmezni az eredeti (angol) mondatot") pontot. Minden fordítást legalább három ember értékelt ki.

1. táblázat. Értékelési szempontok

Megfelelés	**Helyesség**
1: egyáltalán nem jó	1: érthetetlen a mondat
2: jelentésben egy kicsit pontos	2: nem helyes a mondat
3: közepesen jó a pontosság	3: több hibát tartalmaz a mondat
4: jelentésben nagyrészt pontos	4: majdnem jó a mondat
5: jelentésben tökéletesen pontos	5: hibátlan a mondat
0: nem tudom értelmezni az eredeti (angol) mondatot	

5. Módszerek bemutatása

A minőségbecslő modell felépítéséhez szükség van egy jegykinyerő modulra, egy tanítókorpuszra és egy tanító algoritmusra. A jegyek kinyeréséhez a QuEst [13] keretrendszert alkalmaztuk, a tanító algoritmus kiválasztásához a Weka [14] rendszert használtuk.

Az általunk felépített minőségbecslő rendszer kiértékeléséhez az MAE (mean absolute error - átlagos abszolút eltérés), az RMSE (root mean square error - átlagos négyzetes eltérés gyöke) és a Pearson-féle korreláció mértékeket használtuk. A tanítás és a tesztelés során tízszeres keresztvalidálást alkalmaztunk.

[4] http://nlpg.itk.ppke.hu/node/65

Jelen kutatás során összesen 103 minőségbecslő jegyet (103F) próbáltunk ki. Ebből 76 jegy (76F) a Specia és társai [13] által kidolgozott jegyek és 27 jegy saját fejlesztésű jegyek. A 76F tartalmaz nyelvfüggetlen és nyelvspecifikus jegyeket. A nyelvspecifikus jegyekhez magyar nyelvi eszközöket integráltunk a QuEst keretrendszerbe. A szófaji egyértelműsítéshez a Humor [15] rendszerrel tanított PurePos 2.0-t [16] használtuk, a főnévi szerkezetek meghatározásához a Szeged Treebank [17] korpuszon tanított HunTaget [18]. További meglévő jegyek magyar nyelvre való implementálását, magyar nyelvi eszközök hiánya miatt, nem tudtuk elvégezni. A 27 saját fejlesztésű jegyből 3 jegy egy kétnyelvű angol-magyar szótárat használ és 24 jegy a WordNetet. A kinyert jegyek és az emberi értékelések segítségével felállítottuk a kiértékelési modellt, ami alapján becsülni tudtuk a gépi fordítás minőségét.

A tanításhoz több tanuló algoritmussal is kísérleteztünk, de a szupport vektor regresszió (SVR) és a gaussian process (GP) nyújtották a legjobb eredményeket.

Létezik egy nyelvfüggetlen alapjegyhalmaz, ami 17 jegyet (17F) tartalmaz (76F részhalmaza). Az alapjegyhalmazzal betanítottuk a QuEst keretrendszert angol-magyar nyelvre, de nem ért el elég jó minőséget, ezért optimalizáltuk a 103F-t angol-magyar nyelvre. Az optimalizáláshoz „forward selection" algoritmust alkalmaztuk, ami jobb eredményt adott, mint az előző cikkünkben bemutatott algoritmus. Az optimalizálás után 23 jegy (23F) alkotta a végső jegyhalmazt, ami a legjobb eredményt adja az angol-magyar nyelvre. A 23F közül 3 jegy saját fejlesztésű.

5.1. Kétnyelvű szótár és WordNet használata

A minőségbecslés pontosságának növelése céljából 27 új önálló jegyet fejlesztettünk. A 27 saját jegyből 3 jegy egy kétnyelvű szótárat [9] használ:

$$\frac{\text{illeszkedések száma}}{\text{forrásmondat hossza}} \tag{1}$$

$$\frac{\text{illeszkedések száma}}{\text{fordított mondat hossza}} \tag{2}$$

$$\text{(1) és (2) harmonikus középértéke} \tag{3}$$

A kutatásunk során fejlesztettünk 24 WordNetből kinyert jegyet. A feladathoz a Princeton WordNet 3.0-t [19] és a Hungarian WordNetet [20] használtuk. Első körben a forrás- és a fordított mondatból kigyűjtöttük az azonos szinonimahalmazba tartozó szavakat, majd két szinten a szinonimahalmazok hipernimáit is. Végül a kigyűjtött találatok számát súlyoztuk a szinteknek megfelelően.

$$\frac{\text{súlyozott (x illeszkedés) számossága}}{\text{forrásmondat hossza}} \tag{4}$$

$$\frac{\text{súlyozott (x illeszkedés) számossága}}{\text{x számossga a forrásmondatban}} \tag{5}$$

$$\frac{\text{súlyozott (x illeszkedés) számossága}}{\text{fordított mondat hossza}} \tag{6}$$

$$\frac{\text{súlyozott (x illeszkedés) számossága}}{\text{x számossága a fordított mondatban}} \tag{7}$$

$$\text{(4) és (6) harmonikus középértéke} \tag{8}$$

$$\text{(5) és (7) harmonikus középértéke} \tag{9}$$

ahol:

$$x = \text{főnév, ige, melléknév, határozószó}$$

$$\text{súlyozott (x illeszkedés)} = \sum \frac{\text{x illeszkedés}}{\text{szint}}$$

5.2. Mérések

A kutatásaink során, négy különböző mérést végeztünk:

- Első mérés (T1): A C1 korpuszt automatikus módszerekkel kiértékeltük: BLEU, NIST, TER [21].
- Második mérés (T2): A 103F-t használva felépítettünk egy minőségbecslő modellt. A C1 korpuszt az automatikus mértékekkel (BLEU, NIST, TER) tanítottuk be.
- Harmadik mérés (T3): A 103F-t használva felépítettünk egy minőségbecslő modellt. A HuQ korpuszt betanítottuk a megfelelés (M), a helyesség (H), illetve a megfelelés és a helyesség átlagának (MH) értékeivel.
- Negyedik mérés (T4): A HuQ korpuszt betanítva az MH értékekkel, különböző minőségbecslő modelleket építettünk az alábbi jegyhalmazokkal: 17F, 76F, 103F és a magyar nyelvre optimalizált 23F.

6. Eredmények

A T1 méréssel a C1 korpusz minőségéről kaphatunk képet. A BLEU és a TER értékei alapján, a C1 korpusz körülbelül 30%-ban tartalmaz helyes fordítást. A rendszer szintű automatikus módszerekkel mért értékeket lásd az 2. táblázatban.

A T2 és T3 mérése során több tanuló algoritmust is kipróbáltunk, de a GP és a SVR módszerek adták a legjobb eredményeket. Ahogy a 3. táblázatban láthatjuk, a TER kiértékelésben a GP érte el a legjobb eredményeket. A BLEU és a NIST kiértékelésben a SVR ért el jobb eredményt a korrelációban és a MAE-ban, míg RMSE-ben a GP. A 4. táblázat alapján minden esetben a SVR érte el a legjobb értékeket. A 4. táblázatban továbbá az látható, hogy a helyesség

2. táblázat. T1 kiértékelése

TER	0,6107
BLEU	0,3038
NIST	5,1359

3. táblázat. T2 kiértékelése

		TER	BLEU	NIST
GP	Korr	0,3672	0,4028	0,3254
	MAE	0,3202	0,2598	2,7680
	RMSE	0,4277	0,3335	3,4438
SVR	Korr	0,3550	0,4404	0,3669
	MAE	0,3275	0,2201	2,6695
	RMSE	0,4357	0,3474	3,4777

értékeivel magas korrelációt ért el a minőségbecslő modellünk, ami azt jelenti, hogy a használt jegyek magas mértékben jellemzik a célnyelv helyességét.

A T4 kísérletek során, először a 17F-fel tanítottuk be a modellt, majd a 76F-fel, ezután a 103F-fel és végül az angol-magyar nyelvre optimalizált 23F-fel. Az 5. táblázatban látható, hogy a 103F ∼1%-kal jobb eredményt adott a 76F-hez képest és ∼7%-kal jobb eredményt a 17F-hez képest. A magyar nyelvre való optimalizálás után, a 23F ∼11%-kal magasabb korrelációt mutat az alaprendszerhez (17F) képest. Sőt ∼4%-kal jobb eredményt mutat a 103F-hez képest. A 6. táblázatban látható a magyar nyelvre optimalizált 23F.

7. Összefoglalás

A jelen kutatásunkkal továbbfejlesztettük az előző cikkünkben felépített minőségbecslő rendszert. A tanítókorpusz minőségét javítottuk és a kiértékelt mondatok számát növeltük. A továbbfejlesztett korpusz segítségével újratanítottuk a rendszert és további új jegyeket fejlesztettünk. Az új jegyeket egy kétnyelvű szótár és a WordNet segítségével nyertük ki. A mérések során sikerült az általunk fejlesztett jegyek segítségével ∼11%-os minőségjavulást elérni a 17 alapjegykészlethez képest. A WordNet, valamint a szótár integrációjával sikerült javítani a gépi fordítás minőségének becslését.

Célunk az általunk fejlesztett új jegyeket további nyelvpárokra is alkalmazni, illetve új jegyeket kutatni.

4. táblázat. T3 kiértékelése

		Megfelelés	Helyesség	MH
GP	Korr	0,4934	0,5705	0,5536
	MAE	1,0347	0,9407	0,9279
	RMSE	1,1975	1,1208	1,0952
SVR	Korr	0,5058	0,6147	0,5851
	MAE	0,9642	0,8514	0,8621
	RMSE	1,2064	1,0827	1,0739

5. táblázat. Minőségbecslő modell optimalizálása magyar nyelvre (T4)

		Korr	MAE	RMSE
17F	GP	0,5101	0,9333	1,1217
	SVR	0,5112	0,912	1,1353
76F	GP	0,5763	0,9076	1,0925
	SVR	0,5784	0,9036	1,1214
103F	GP	0,5536	0,9279	1,0952
	SVR	0,5851	0,8621	1,0739
23F	GP	0,5859	0,8704	1,0578
	SVR	**0,6275**	**0,795**	**1,0292**

Hivatkozások

1. Tantug, A.C., Oflazer, K., El-Kahlout, I.D.: BLEU+: a Tool for Fine-Grained BLEU Computation. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) http://www.lrec-conf.org/proceedings/lrec2008/.
2. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. (2005) 65–72
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 311–318
4. Lin, C.Y., Och, F.J.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL '04, Stroudsburg, PA, USA, Association for Computational Linguistics (2004)

6. táblázat. Magyar nyelvre optimalizált 23 jegy

vesszők számosságának abszolút különbsége a forrás- és célmondatban
tokenek száma a célmondatban, amelyek nem csak a-z betűt tartalmaznak.
igék százaléka a célmondatban
szótár illeszkedés f-mérték
igék százaléka a forrásmondatban
célmondat perplexitás
tokenek száma a célmondatban
átlagos bigram gyakoriság, a második kvartilisben lévő gyakorisága a forrásnyelvi korpuszban
szótár illeszkedés / forrásmondat hossza
forrásmondat perplexitás
írásjelek aránya a célmondatban
átlagos unigram gyakoriság, az első kvartilisben lévő gyakorisága a forrásnyelvi korpuszban
kettőspontok számosságának abszolút különbsége a forrás- és célmondatban
WordNet illeszkedés a forrásmondatban: főnevek / főnevek száma
átlagos unigram gyakoriság, a második kvartilisben lévő gyakorisága a forrásnyelvi korpuszban
forrásmondatban lévő a-z tokenek százalékának és a célmondatban lévő a-z tokenek százalékának aránya
átlagos trigram gyakoriság, az első kvartilisben lévő gyakorisága a forrásnyelvi korpuszban
forrásmondat perplexitás, amelyik nem tartalmaz mondatvégi írásjelet
felkiáltójelek számosságának abszolút különbsége a forrás- és célmondatban, a célmondat hosszával normalizálva
felkiáltójelek számosságának abszolút különbsége a forrás- és célmondatban
átlagos bigram gyakoriság, a harmadik kvartilisben lévő gyakorisága a forrásnyelvi korpuszban
kettőspontok számosságának abszolút különbsége a forrás- és célmondatban, a célmondat hosszával normalizálva
tokenek száma a forrásmondatban, amelyek nem csak a-z betűt tartalmaznak

5. Beck, D., Shah, K., Cohn, T., Specia, L.: SHEF-Lite: When Less is More for Translation Quality Estimation. In: Proceedings of the Workshop on Machine Translation (WMT). (2013)
6. Biçici, E.: Feature Decay Algorithms for Fast Deployment of Accurate Statistical Machine Translation Systems. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, Sofia, Bulgaria, Association for Computational Linguistics (2013)
7. Camargo de Souza, J.G., Buck, C., Turchi, M., Negri, M.: FBK-UEdin participation to the WMT13 quality estimation shared task. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, Sofia, Bulgaria, Association for Computational Linguistics (2013) 352–358
8. Yang, Z.Gy., Laki, L., Prószéky, G.: Gépi fordítás minőségének becslése referencia nélküli módszerrel. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY

2015). (2015) 3–13
9. Novák, A., Tihanyi, L., Prószéky, G.: The MetaMorpho Translation System. In: Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 111–114
10. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. ACL '07, Stroudsburg, PA, USA, Association for Computational Linguistics (2007) 177–180
11. Halácsy, P., Kornai, A., Németh, L., Sass, B., Varga, D., Váradi, T., Vonyó, A.: A Hunglish korpusz és szótár. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Egyetem (2005)
12. Koehn, P.: Statistical Machine Translation. 1st edn. Cambridge University Press, New York, NY, USA (2010)
13. Specia, L., Shah, K., de Souza, J.G., Cohn, T.: QuEst - A translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Sofia, Bulgaria, Association for Computational Linguistics (2013) 79–84
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl. **11**(1) (2009) 10–18
15. Prószéky, G.: Industrial applications of unification morphology. In: Proceedings of the Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany, Association for Computational Linguistics (1994) 213–214
16. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: RANLP'13. (2013) 539–545
17. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue, Springer (2005) 123–131
18. Recski, G., Varga, D.: A Hungarian NP Chunker. The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics (2009) 87–93
19. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
20. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of the Fourth Global WordNet Conference (GWC 2008). (2008) 310–320
21. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas. (2006) 223–231

II. Morfológia, előfeldolgozás

Ékezetek automatikus helyreállítása magyar nyelvű szövegekben

Novák Attila[1,2], Siklósi Borbála[2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport ,
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
e-mail:{novak.attila, siklosi.borbala}@itk.ppke.hu

Kivonat Cikkünkben egy olyan rendszert mutatunk be, ami a statisztikai gépi fordítás módszereit használva megbízhatóan pótolja a hiányzó ékezeteket ékezetek nélkül írt magyar nyelvű szövegekben. Mivel magyar nyelv esetén elkerülhetetlen, hogy igen nagyméretű szöveges tanítókorpusz alkalmazása esetén is hiányozzanak bizonyos szóalakok a tanítóanyagból, morfológiai elemzőt integráltunk a rendszerbe, ami ékezetesített szóalakjelölteket generál ezekhez a szavakhoz. Az így létrejött rendszert kiértékelve a rendszer az esetek több mint 99%-ában helyes ékezetes alakot állított elő.

1. Bevezetés

Napjaink népszerű kommunikációs fórumai a közösségi oldalak. Az ezek használata során leírt szövegek létrehozása egyre inkább olyan mobil eszközökhöz kötődik, melyek szöveges beviteli felülete a magyar nyelv ékezetes karaktereinek elérésére nem nyújt kényelmes és gyors hozzáférést. Ezért egyre több olyan szöveg jön létre, ami nem tartalmaz ékezeteket. Az ebből adódó többértelműségek feloldása az emberek számára ritkán okoz problémát, ahhoz azonban, hogy ezek a szövegek a szokványos nyelvtechnológiai eszközökkel feldolgozhatóak, elemezhetőek legyenek, szükség van az ékezetek visszaállítására. Cikkünkben egy olyan magyar ékezetesítőrendszert mutatunk be, ami egy statisztikai gépi fordító (SMT) keretrendszer és egy morfológiai elemző kombinációjából áll.

Ugyan léteznek más ékezetesítőrendszerek magyar nyelvre, azonban azok vagy nem elérhetőek, vagy rosszabbul teljesítenek. Jellemző továbbá a probléma karakteralapú megközelítése, azonban az ilyen rendszereknél óhatatlanul megjelennek értelmetlen szóalakok. Ezzel szemben, a szótáralapú megoldások a szótárban nem szereplő szavak ékezetesítésére nem tudnak javaslatot tenni.

2. Kapcsolódó munkák

Több kutatás célozta már meg az ékezetek helyreállításának megoldását magyar nyelvű szövegek esetén. [11] és [3] gépi tanulási módszereket alkalmaztak, ahol a

beszúrandó ékezetek pozícióját az ékezet nélküli betű közvetlen környezete alapján határozzák meg. Ezzel a módszerrel 95%-os pontosságot értek el. A módszer előnye, hogy a tanítóanyagban nem szereplő, ismeretlen szavakat is kezelni tudja, hátránya viszont, hogy nem létező szóalakokat is generál. A feladat egy másik megközelítése szótár használatán alapul. Ezek a módszerek nagy szöveges korpuszból becslik meg a különböző ékezetes alakok disztribúcióját. [11] ezzel a módszerrel 98%-os pontosságról számol be. Ez a rendszer viszont nem tudja kezelni az ismeretlen szavakat. [4] egy többszintű nyelvfeldolgozó rendszert mutat be, amit egy text-to-speech alkalmazáshoz hoztak létre. Ennek keretein belül az ékezetek helyreállításához morfológiai és szintaktikai elemzést is végeznek, így az ékezetesítés pontossága erősen függ az elemzők teljesítményétől (95%-os pontosságot sikerült elérniük).

A Charlifter [8] egy nyelvfüggetlen ékezetesítőrendszer, ami lexikonalapú statisztikai módszereket alkalmaz, illetve egy bigram környezeti modellt és az ismeretlen szavak kezelésére egy karakteralapú statisztikai modellt is használ. A rendszert kipróbáltuk magyarra. Ennek teljesítményét alább a saját rendszerünkével összevetve részletezzük.

Más nyelvekre is hasonló módszereket találunk. [10] átfogó elemzést mutat be francia és spanyol szövegek ékezetesítésére adott megoldásokról. Az esettanulmány a szövegkörnyezet jelentőségét hangsúlyozza, de mind a különböző szóalakok, mind az ékezetek száma jóval kevesebb ezekben a nyelvekben, mint a magyarban. [12] szintén francia nyelvre ad megoldást, azonban kifejezetten orvosi szakszövegekkel, szavakkal foglalkozik, aminek jellegzetessége az ismeretlen szavak magas aránya az általános nyelvhasználathoz képest. A módszer címkézési feladatként fogalmazza meg a problémát, amit transzducerekkel oldanak meg. A tesztek során 92%-os pontosságot értek el egy orvosi tezaurusz címszavain mérve, szövegkörnyezet nélkül.

A saját módszerünkhöz leginkább [6] módszere hasonlít. Ebben a kutatásban szintén gépifordító-rendszert alkalmaztak vietnami szövegek ékezetesítésére, 93%-os pontosságot érve el. Ez a rendszer azonban egy külső szótárt is használ, továbbá a vietnami[3] és a magyar nyelv sajátosságai közötti különbségek miatt az eredmények nem összemérhetőek.

3. Ékezetek helyreállítása

Az ékezetek helyreállításának problémáját fordítási feladatként fogalmaztuk meg, ahol a forrásnyelv az ékezet nélküli szöveg, a célnyelv pedig az ékezetes változat. Mivel ebben az esetben nagyon könnyű nagyméretű párhuzamos korpuszt létrehozni (hiszen egy egynyelvű korpuszból csak el kell távolítani az ékezeteket),

[3] A tonális vietnami nyelvben különböző mellékjeleket használnak egyes magánhangzófonémák megkülönböztetésére (négy különböző mellékjel) és a szóalakokban szereplő szótagok tónusának jelölésére (öt különböző mellékjel). A vietnamiban több az ékezet, mint a magyarban. Gyakran egy magánhangzót jelölő betűn két különböző mellékjel is megjelenik. Ugyanakkor a vietnami izoláló nyelv, ezért a produktív magyar morfológiából adódó rengeteg különböző szóalak a vietnamira nem jellemző.

magától értetődőnek tűnt a statisztikai gépi fordító (SMT) rendszer alkalmazása, melyben a fordítási modell az egyes frázisok lehetséges ékezetes változatainak eloszlását tartalmazza, a nyelvmodell pedig a szövegkörnyezetet képviselve az aktuális környezetben helyes alak kiválasztását biztosítja. A rendszer magjaként a Moses [2] keretrendszert használtuk a fordítási modell építéséhez és a dekódoláshoz, a nyelvmodellt pedig a SRILM [9] eszközzel hoztuk létre. A Moses használata során annak alapértelmezett konfigurációs beállításait használtuk a szóösszerendelő lépés kihagyásával, amire ebben a feladatban nem volt szükség, hiszen minden forrásoldali szó egyértelműen megfeleltethető a céloldali párjának. Ugyanez indokolta azt is, hogy a dekódolás során csak monoton fordítást engedélyeztünk, azaz a szórendnek változatlannak kellett maradnia.

3.1. Az alaprendszer

Az alaprendszerben csak a tanítóanyagból épített fordítási- és nyelvmodelleket használtuk. A dekóder bemenete a hiányzó ékezeteket tartalmazó magyar szöveg. A fordítási modell csupán unigramokat tartalmazott ebben a felállásban (magasabb rendű n-gramokkal is kísérleteztünk, ez azonban az eredményre nem volt hatással), a nyelvmodellben pedig legfeljebb 5 szó méretű frázisok szerepeltek. Így a fordítási modell meghatározta az egyes szavakhoz tartozó ékezetes alakok disztribúcióját, míg a nyelvmodell a szövegkörnyezetet képviselve választja ki a megfelelő alakot. Az így létrehozott baseline rendszer hiányossága azonban, hogy a tanítókorpuszban nem szereplő ismeretlen (OOV) szavakat egyáltalán nem kezeli.

Egy másik alaprendszert is létrehoztunk az SMT rendszer hatásának vizsgálatára. Ebben a rendszerben minden ékezet nélküli szót mindig a leggyakoribb ékezetes alakjára cseréltük a szövegkörnyezet figyelembevétele nélkül. Ehhez a gyakorisági adatokat a tanítókorpuszból határoztuk meg.

3.2. Morfológiai elemző integrálása

A korpuszban nem szereplő ismeretlen szavak kezelésére a Humor morfológiai elemzőt [5,7] integráltuk a rendszerbe. Az eredeti elemzőnek egy olyan módosított változatát hoztuk létre és használtuk ebben a feladatban, ami az ékezet nélküli szóalakokat közvetlenül leképezi a lehetséges ékezetes változataikra. Továbbá, a morfémahatárokat is jelöli és meghatározza a morfoszintaktikai kategóriacímkét a kapott szóalakokhoz. A szegmentálásra vonatkozó jelölések (pl. szóösszetételi határok, képzők) és a kategóriacímkék az ékezetes alakok rangsorolásához használt pontszám számításakor szükségesek. Az ékezetes szóalakokat újraelemezzük, hogy a közvetlenül nem kinyerhető lemmákat is megkapjuk. A kísérleteinknél használt, 1 804 252 token méretű tesztanyagban a szavak kb. 1%-a nem volt benne a fordítási modellben a legnagyobb, 440 millió token méretű, tanítóanyag esetén sem. Az 1. táblázatban látható az ismeretlen szavak (OOV) aránya a fordítási modell létrehozásához használt különböző méretű tanítóanyagok esetén.

1. táblázat. Az ismeretlen (OOV) szavak aránya a különböző méretű tanítóanyagból épített fordítási modellek esetén.

tanítóanyag	**mondatok száma**	**millió szó**	**OOV a tesztanyagban**
100K	100 000	1,738	9,63%
1000K	1 000 000	18,078	3,44%
5000K	5 000 000	89,907	1,23%
10M	10 000 000	180,644	1,68%
ALL	24 048 302	437,559	0,81%

A tesztanyagban előforduló ismeretlen szavak esetén az elemző ékezetesített szóalakokat javasol. Ezeket a javasolt szóalakokat a Moses rendszer esetén azoknak a fordítandó szövegbe való beágyazásával továbbítani tudjuk a fordítórendszer felé. Ehhez azonban minden egyes javasolt szóalakhoz valószínűséget kell rendelni. Először egyenletes eloszlást feltételeztünk, így azonban a gyakori ékezetes alakok és a gyakorlatilag értelmetlen (bár nyelvtanilag helyes) alakok is azonos valószínűséggel szerepeltek. Hogy ezek a szóalakok ne jelenjenek meg az eredményben, a második változatban kifinomultabb algoritmust alkalmaztunk a valószínűségek becslésére.

Az ékezetesített javaslatokhoz egy-egy pontszámot rendelünk, amely alapján rangsorolhatók a kapott szóalakok. Mivel maga a szóalak nincs benne a korpuszban, ezért a pontszám a következő tényezők lineáris kombinációjaként jön létre: (1) lemmagyakoriság (LEM), (2) a szóalakban megjelenő ragsorozat gyakorisága (INF), (3) a szóalakban előforduló produktív összetételek (CMP), és (4) produktív képzők száma (DER). Az első két tényezőt a tanítókorpuszból számított statisztika alapján határoztuk meg. A modell használatakor ezek a tényezők pozitív súlyozást kaptak, előnyben részesítve ezzel a gyakori lemmákat, illetve gyakori toldalékkombinációkat. A második két tényező ezzel szemben negatív súlyozást kapott, csökkentve ezzel a többszörös összetételeket és képzőket tartalmazó jelöltek pontszámát. Az egyes ékezetes jelöltekhez rendelt pontszámot tehát az (1) egyenlet alkalmazásával határoztuk meg.

$$score = -\lambda_c \#CMP - \lambda_d \#DER + log_{10} LEM + \lambda_i log_{10} INF + MS \ , \quad (1)$$

ahol

$$MS = \begin{cases} |minscore| + 1 \text{ ha} & minscore \leq 0 \\ 0 & \text{egyébként} \end{cases} \quad (2)$$

Az MS komponens a pontszámok felskálázása miatt került bevezetésre, a kapott pontszámhoz $|minscore| + 1$-et adva hozzá, azaz az aktuális jelöltlistában szereplő legkisebb pontszámmal növelve az összes jelölt pontszámát, ezzel védve ki a negatív pontszámok megjelenését. A λ súlyokat a Moses `mert` optimalizáló programjával állítottuk be. Ehhez a korpusz egy előre elkülönített részén megvizsgáltuk a morfológiai elemző által elemzett OOV szavakban az összetételek, képzők és ragok eloszlását, majd a megfigyelt eloszlásnak megfelelően, de

véletlenszerűen kiválasztottunk 1000 szót. A `mert` optimalizálás célváltozója a rendszer pontossága volt erre az 1000 szóra, az így kapott λ súlyokat használtuk a modellben. Bár lineáris regressziót alkalmaztunk, melynek során bevett szokás egy torzító súly (bias weight) hozzáadása is, ezt nem tartottuk szükségesnek, mivel nem kellett a kapott becsléseket más forrásokból való becslésekkel szinkronba hozni. Végül normalizálással valószínűségi eloszlássá alakítottuk a kapott pontszámokat.

Bár a pontszámok megfelelő skálázásával a rangsorolt ékezetesített szóalak-jelöltek a fordítási modellben meglévő frázisokhoz hasonlóan mind felhasználhatóak lettek volna, a rendszerbe csak a legmagasabb pontszámot kapott jelöltet továbbítottuk. Ennek oka, hogy mivel a nyelvmodellben ezek a szóalakok nem szerepeltek (tehát a nyelvmodell szerint mindegyiknek azonos a valószínűsége), ezért mindenképp a legnagyobb valószínűségű szóalakot választja a rendszer az egyes ékezetlen szóalakokhoz generált jelöltek listájából.

4. Eredmények

A rendszer tanításához és teszteléséhez a magyar webkorpuszt használtuk [1]. Ebből 100 000 mondatot (1 804 252 token) tettünk félre teszteléshez, másik 100 000 mondatot optimalizáláshoz, a többit pedig több különböző felbontásban a tanításhoz használtuk. A legkisebb tanítóanyag 100 000 mondatból állt, a legnagyobb pedig több mint 24 millió mondatot tartalmazott. Az egyes tanítóanyagok méretét az 1. táblázatban foglaltuk össze. A kiértékelés során a tanítóanyag méretének növelése mellett vizsgáltuk a rendszer teljesítményét, illetve a morfológiai elemző integrálásának hatását.

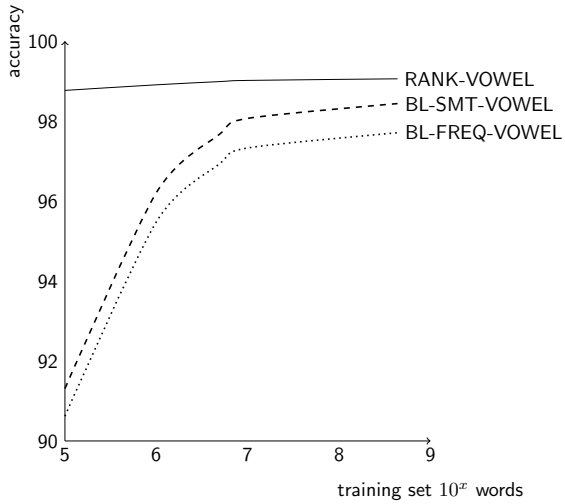
A teszthalmaz ékezetmentes, eredeti állapotában a tokenek 56,84%-a volt helyes szóalak, a magánhangzókat tartalmazó szavaknak pedig (ahol egyáltalán történhet ékezetesítés) 47,09%-a. Ebből az állapotból a morfológia nélküli SMT rendszer (BL-SMT) a legkisebb tanítóanyag esetén 91,31%-ra javította a magánhangzót tartalmazó szavak helyességének arányát, míg a teljes tanítóanyag felhasználása esetén ez az arány 98,44%-ra nőtt. A morfológiai elemző használata mellett (RANK) a pontosság 98,77%, illetve 99,06% volt a két tanítóanyag esetén. A minden szót mindig a leggyakoribb ékezetesített alakra cserélő alaprendszer pontossága (BL-FREQ) pedig 90,62%, illetve 97,71% volt. A 2. táblázatban láthatóak a részletes eredmények a legkisebb és a legnagyobb tanítóanyag használata mellett az összes szó (ALL) és a magánhangzókat tartalmazó szavak (VOWEL) esetén. Látható, hogy a rendszer pontossága a tanítóanyag növelésével csekély mértékben növekszik, viszont a fedés és a helyesség drasztikusan nő a morfológiai elemzőt nem használó rendszerek esetén.

A morfológiai elemző integrálásával azonban pótolni lehetett a kis tanítóanyagból hiányzó információt, amivel jelentősen megnőtt a fedés. Még a legnagyobb tanítóanyag esetén is 39,74%-os hibaarány-csökkentést eredményez az elemző használata, a hibás szavak arányát 1,56%ról 0,94%-ra csökkentve. A legkisebb tanítóanyag esetén a hibaarány-csökkenés 85,85%. Az elemzőt használó rendszer tehát a legkisebb tanítóanyag mellett is jobban teljesít, mint az elemző

nélküli SMT rendszer a legnagyobb tanítóanyagon tanítva. A 1. ábra az egyes rendszerek tanulási görbéjét mutatja, azaz a rendszer helyességét a tanítóanyag méretének függvényében.

2. táblázat. A pontosság alakulása az egyes rendszerparaméterek és a tanítóanyag mérete függvényében.

	100K			ALL		
rendszer	prec	rec	acc	prec	rec	acc
BL-FREQ-ALL	98,25	82,82	92,34	98,37	96,26	98,13
BL-FREQ-VOW	98,25	82,82	90,62	98,37	96,26	97,71
BL-SMT-ALL	99,03	83,88	92,91	99,09	97,36	98,72
RANK-ALL	98,81	98,08	98,99	99,01	98,56	99,23
RANK-VOW	98,82	98,08	**98,77**	99,02	98,56	**99,06**



1. ábra. A magánhangzót tartalmazó szavakon mért pontosság a tanítóanyag mérete függvényében az egyes rendszerek esetében.

Az eredményeinket összehasonlítottuk a Charlifter rendszerrel elért eredményekkel. Ennek teljesítménye 89,75% helyesség a leggyakoribb ékezetes alakok használata esetén, 90,00% a *lexicon-lookup+bigram* kontextuális modell esetén és 93,31% a *lookup+bigram context+character-n-gram* modell esetén. Az összehasonlításból látható, hogy az SMT modellben használt nyelvmodell jobban növeli a rendszer helyességét, mint a Charlifter által használt bigram kontextus modell,

a morfológiai elemzővel kiegészített SMT rendszer pedig szintén jobban teljesít, mint a karakteralapú n-gram modell.

5. Hibaelemzés

A tesztanyag egy 5000 mondatos részén részletes hibaelemzést is végeztünk. Ennek részletes eredményeit a 3. táblázatban foglaltuk össze.

A részletes elemzés során kiderült, hogy az eredeti és a rendszer által ékezetesített szöveg szavai közötti eltérés 14,7%-a nem valódi hiba. 3,55%-ban ekvivalens alakot kaptunk (pl. *lévő*∼*levő*), míg a többi a referenciában szerepelt hibásan, a rendszer által adott eredmény volt a helyes.

A referencia egy másik jelentős hányada (17,91%) szintén hibás volt, azonban ezekben az esetekben a hiba nem az ékezetek hiányából fakadt, ezért nem tudta a rendszerünk javítani. Ezek a hibák leggyakrabban hiányzó vagy felcserélt betűkből adódnak (10,81%), további 6,42% pedig valamilyen központozási hiba az eredeti referenciaszövegben.

A hibák kb. 2/3-a volt valódi hiba. Ezek 5,57%-ában a szótő ismeretlen volt a morfológiai elemző számára. Az esetek 3,55%-a olyan hiba volt, amikor a rendszer egy tulajdonnevet egy gyakoribb szóalakra alakított át: vagy egy másik tulajdonnévre, vagy csupán egy gyakori szóalakra. Hasonló hiba, amikor egy köznevet a rendszer egy gyakoribb tulajdonnévre alakít át (további 1,35%). Ezeknek a hibáknak egy részét kezelni lehetne, ha a rendszer figyelembe venné a kisbetű-nagybetű megkülönböztetést. Ez azonban más esetekben okozhatna hibát, a rendszer általános teljesítménye feltehetőleg romlana az adathiány miatt.

A hibák 2,20%-a a tanítóanyagban lévő hibákból fakadt. Mivel a magyarban gyakoriak a ritka szóalakok, ezért könnyen előfordulhat, hogy egy szó többször szerepel hibásan, mint helyesen (különösképpen igaz ez a korpuszban csupán egyszer szereplő szóalakokra). A vizsgált tesztanyagban előforduló hibák további 3,72%-a abból adódott, hogy a rendszer a szándékosan ékezet nélkül írt szóalakokat (fájlnevek, url-ek) is átalakította azok ékezetes alakjára, vagy valami más, értelmes szóalakra, vagy éppen ennek ellenkezője történt, a szövegben furcsa mód ékezetesen írt url-t nem ékezetesített (pl. *www.valamicég.hu*).

A hibák legnagyobb része (51,01%) olyan eset, amikor a rendszer nem tudta a környezet alapján sem eldönteni, hogy mi lenne a helyes szóalak. Ezeknek az eseteknek több, mint fele olyan hiba, amikor a rendszer felcserélte egy birtokos és a nem birtokos alakját egy adott főnévnek (pl.*gyereket*∼*gyerekét*, *gyereken*∼*gyerekén*, *gyereke*∼*gyereké*). További 26% a igék definit és indefinit alakjának hasonló tévesztéséből fakad (pl. *hajtottak*∼*hajtották*, *hajtanak*∼*hajtanák*, *hajtana*∼*hajtaná*).

3. táblázat. Hibaelemzés a rendszerkimenet és az eredeti szöveg eltéréseinek vizsgálatával egy 5000 mondatos tesztanyagon.

Hibatípus	**Arány**	**Példák**
A rendszer kimenete helyes	**14,70%**	
Ekvivalens alakok	3,55%	lévő→levő fele→felé áhá→aha periférikus→periferikus
Javított hibás név	1,01%	USÁ-ban→USA-ban Szóladon→Szóládon
Más javított hiba	10,14%	un.→ún. kollegánk→kollégánk lejto→lejtő lathato→látható
Valódi hibák	**67,40%**	
Hiányzik az elemzőből	5,57%	hemokromatózis-gén→hemokromatozis-gen
Helyes névből hibás kimenet	3,55%	MIG→míg Bösz→Bősz Ladd→Ládd Márton→Marton
Más helyes eredetiből hibás kimenet	2,20%	megőrzést→megorzést routeréhez→routerehez
Más helyes eredetiből a kontextusban hibás név	1,35%	logó→logo eperjeskein→eperjeskéin
Más helyes eredetiből a kontextusban hibás egyéb szó	51,01%	még→meg termek→termék gépét→gépet címét→címet vágyók→vagyok érméket→érmeket képé→képe
Az eredeti fájlnév vagy ékezetet tartalmazó URL	3,72%	latok→látok viz→víz szantok→szántók telepok→telepók www.valamicég.hu→www.valamiceg.hu
Nem javított hiba az eredetiben	**17,91%**	
Központozási hiba az eredetiben	6,42%	közalk.tan→kozalk.tan 1922.évi→1922.evi
Elválasztási hiba az eredetiben	0,68%	bemuta-tásra→bemuta-tasra
Egyéb hiba az eredetiben	10,81%	véri→veri ra→rá gonolkozásában→gonolkozasaban imátkoztok→imatkoztok hirújsághoz→hirujsasghoz változaban→változában környezetkíméli→kornyezetkimeli

6. Konklúzió

A cikkben egy magyar szövegek ékezetesítésére alkalmas rendszert mutattunk be. A statisztikai gépi fordítón alapuló alaprendszer fix méretű tanítókorpuszból épített fordítási-, és nyelvmodell használatával az esetek 98,44%-ában tudta helyesen ékezetesíteni az ékezet nélküli szóalakokat. Ez a rendszer azonban csak a tanítóanyagban szereplő szavak kezelésére képes. Ennek a problémának a megoldására morfológiai elemzőt integráltunk a rendszerbe, ami a tanítóanyagból hiányzó szavakhoz ékezetesített szóalakjelölteket generál. Ezzel a megoldással a helyesen ékezetesített szavak aránya 99,06%-ra nőtt. A rendszer további jellemzője, hogy a szövegkörnyezetet is figyelembe veszi nyelvmodell alkalmazásával, emellett nem generál értelmetlen szóalakokat, ami az ismeretlen szavak karakteralapú kezelésénél elkerülhetetlen lenne.

Hivatkozások

1. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating Open Language Resources for Hungarian. In: LREC. European Language Resources Association (2004)
2. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions. pp. 177–180. Association for Computational Linguistics, Prague (2007)
3. Mihalcea, R., Nastase, V.: Letter level learning for language independent diacritics restoration. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. pp. 1–7. COLING-02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1118853.1118874
4. 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 3(3-4), 217–236 (2000)
5. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
6. Pham, L.N., Tran, V.H., Nguyen, V.V.: Vietnamese text accent restoration with statistical machine translation. In: Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27). pp. 423–429. Department of English, National Chengchi University (2013), http://aclweb.org/anthology/Y13-1044
7. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
8. Scannell, K.P.: Statistical unicodification of african languages. Language Resources and Evaluation 45(3), 375–386 (2011)
9. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: Update and outlook. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Waikoloa, Hawaii (Dec 2011)

10. Yarowsky, D.: A comparison of corpus-based techniques for restoring accents in Spanish and French text. In: Proceedings of the 2nd Annual Workshop on Very Large Text Corpora. pp. 19—-32. Las Cruces (1994)
11. Zainkó, Cs., Németh, G., Olaszy, G., Gordos, G.: Eljárás adott nyelven ékezetes betűk használata nélkül készített szövegek ékezetes betűinek visszaállítására (2000)
12. Zweigenbaum, P., Grabar, N.: Accenting unknown words in a specialized language. In: Johnson, S. (ed.) ACL Workshop on Natural Language Processing in the Biomedical Domain. pp. 21–28. ACL (2002)

Utilizing Word Embeddings for Part-of-Speech Tagging

Gábor Berend

Szegedi Tudományegyetem,
TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.,
e-mail: berendg@inf.u-szeged.hu

Abstract. In this paper, we illustrate the power of distributed word representations for the part-of-speech tagging of Hungarian texts. We trained CRF models for POS-tagging that made use of features derived from the sparse coding of the word embeddings of Hungarian words as signals. We show that relying on such a representation, it is possible to avoid the creation of language specific features for achieving reliable performance. We evaluated our models on all the subsections of the Szeged Treebank both using MSD and universal morphology tag sets. Furthermore, we also report results for inter-subcorpora experiments.

1 Introduction

Designing hand-crafted features for various natural language processing tasks, such as part-of-speech (POS) tagging or named entity recognition (NER) has a long going history [1,2]. Systems that build upon such (highly) language/task-specific features can often perform accurately, however, at the cost of losing their ability to work well across different languages and tasks. A further drawback of such approaches is that the human-powered design of features can be a time consuming and expensive task without any guarantees that the features work well under multiple circumstances or at all.

There is now a recent line of research gaining increasing popularity, which aims at building more general models that require no feature engineering at all but relying on large collections of (unlabeled) texts alone [3,4,5,6]. For the above reason these models can be regarded language independent, making them more likely to be applicable across languages.

Sparse coding aims at expressing observations as a sparse linear combination of 'basis vectors'[1] [7]. The goal of our work is to combine two popular approaches, i.e. sparse coding and distributed word representations.

In our work we propose a POS tagging architecture which was evaluated on the Szeged Treebank using MSD and universal morphology tag sets. We report our POS tagging results on the levels of the six subcorpora the Szeged Treebank comprises of. Also, we evaluated our trained models in a cross-genre setting.

[1] The term basis vectors is used intuitively throughout the paper, as they need not be linearly independent.

2 Related work

The line of research introduced in this paper relies on distributed word representations [8] and dictionary learning for sparse coding [7], both area having a substantial literature. This section introduces the most important previous work along these topics.

2.1 Distributed word representations

Distributed word representations provided by approaches such as `word2vec` [6] and `GloVe` [9], enjoy great popularity these days as they have been shown to accurately model the semantics of words [10]. This property makes them available to perform successfully in semantic and syntactic word analogy tasks. There exist previous results claiming that distributed word representations are also useful in the word analogy task in Hungarian (and other lower-resourced Central European languages) [11]. There exist a variety of approaches on how continuous word embeddings can be determined, e.g. [8,3,4,6,9].

The Polyglot [8] neural net architecture is one such possible alternative to determine word embeddings. In their proposed model, word embeddings were trained on the passages of Wikipedia, while preprocessing of texts was kept at a minimal level by not performing lowercasing or lemmatization. Applying such a generic approach for preprocessing not favoring any specific language makes this neural network architecture applicable for a variety of languages without any serious modifications. Indeed, the authors also made their pre-trained word embeddings for over 130 languages publicly available[2] providing basis for cross-, and multi-lingual experimentation. Since we wanted to give an approach that is not sensitive to the hyperparameters of the word embedding model, we applied those Polyglot word embedding vectors trained for Hungarian that are available for download at the Polyglot project website.

2.2 Sparse coding

Sparse coding has it roots in the computer vision community, and its usage is perhaps no so common in natural language processing literature. The general purpose of sparse coding is to express signals in the form of a *sparse* linear combinations of basis vectors, while the task of finding an appropriate set of basis vectors is referred to as *dictionary learning* problem [7]. Generally, given a data matrix $X \in \mathbb{R}^{k \times n}$ with its i^{th} column $\mathbf{x_i}$ representing the i^{th} k-dimensional signal, the task is to find $D \in \mathbb{R}^{k \times m}$ and $\alpha \in \mathbb{R}^{m \times n}$, such that the product of matrices D and α approximates X. Mairal et al. [7] formalized this problem as an ℓ_1-regularized linear least-squares minimization of the form

$$\min_{D \in \mathcal{C}, \alpha} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(\|x_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right),$$

[2] https://sites.google.com/site/rmyeid/projects/polyglot

with $\mathcal{C}$ being the convex set of matrices that comprise of column vectors having an ℓ_2 norm at most one, matrix D acts as the shared dictionary across the signals, and the columns of the sparse matrix α contains the coefficients for the linear combinations of each of the n observed signals. [7] describes an efficient algorithm for solving the above optimization that we also applied in our experiments[3].

3 Sequence labeling framework

This section introduces the sequence labeling framework we employed for POS tagging. During our experiments the main source of features for the tokens in a sentence was the dictionary learning based sparse coding of their word embedding vector. Once the dictionary matrix D is given α_i, the sparse linear combination coefficients for a word embedding vector w_i, can be determined efficiently by solving the kind of minimization problem described in Section 2.2. The way we turned these sparse coefficients into features was that we regarded those indices of α_i as features that had a non-zero value, i.e. $f(w_i) = \{j : \alpha_i[j] \neq 0\}, \alpha_i[j]$ denoting j^{th} coefficient stored in the sparse vector α_i. It can be illustrative if we check out the kind of features that got determined for semantically related words. Table 1 includes such a set of words and their corresponding features. In Table 1 any feature ID appearing more than got boldfaced.

The only language dependent feature we made use of was the identity of words. For the calculation of this feature we performed no preprocessing, i.e. the words were not lemmatized and even their capitalization was left unchanged.

Table 1: Example words all being body parts and the sparse features induced for them. Features with multiple occurrences across words are in **bold** typeface. Within parentheses are the English equivalents of the Hungarian example words.

Word	Sparse features induced
kéz (hand)	{144, 218, **309**, 472, 713, 870, **916**}
láb (leg)	{138, 186, **250**, **309**, 324, 583, 626, 796, **948**}
fej (head)	{101, **250**, 271, **309**, 516, **783**, **916**, **948**}
törzs (trunk)	{81, **309**, **783**, 867, 948}
csukló (wrist)	{84, 194, **309**, 607, 815, 957}

When assigning features to a target word at some position within a sentence, we determined the same set of feature functions for the target word itself and its neighboring words of window size 1. We then used the previously described set of features in a linear chain CRF [12], using the CRFsuite implementation [13]. The coefficients for ℓ_1 and ℓ_2 regularization were set to 1.0 and 0.001, respectively.

[3] `http://spams-devel.gforge.inria.fr/`

4 Results and discussion

We evaluated our proposed POS tagging framework on the Szeged Treebank [14], which has six subcorpora, namely text related to *computer*s, *law*, *literature*, short news (referenced as *newsml*), *newspaper* articles and *student* writing. The performance of our POS tagger models is expressed as the fraction of correctly tagged tokens (per-token) evaluation and as a fraction of the correctly tagged sentences (per-sentence) evaluation when a sentence is regarded as correct if all the tokens it comprises are tagged correctly. Evaluation was performed according to the reduced tag set of the MSD v2.5 and the universal morphologies as well. In the two distinct tag sets, we faced a 93-class and a 17-class sequence classification problem, respectively. The dictionary learning approach we made use of relied on two parameters, the dimensionality of the basis vectors and the regularization parameter effecting the sparsity of the coefficients in α. We chose the former parameter to be 1024 and the latter to be 0.4, nevertheless we should also add the general tendencies remained the same when we chose other pairs of parameters.

The first factor that could influence the performance of our approach is the coverage of the word embedding vectors employed, i.e. what extent of the training/test tokens/word forms do we have a distributed representation determined for. Table 2 includes these information. We can see that due to the morphological richness of Hungarian, the word form coverage of the roughly 150,000 word embedding vectors we had access to is relatively low (around 60%) for all the domains in the treebank. Due to the Zipfian distribution of word frequencies, however, we could experience a much higher (almost 90%) coverage for all the domains in the treebank on the level of tokens. It is interesting to see that student writings have one of the lowest word form coverage, while it is among the genres with the highest token coverage. It might indicate that student writing is not as elaborate and standardized as news writing for instance.

Table 2: The token and word form coverages of the Polyglot word embeddings on the Szeged Treebank. In parentheses are the ranks for a given domain.

	Training		Test		Average
Domain	Tokens	Word forms	Tokens	Word forms	Tokens
computer	88.54% (4)	60.13% (3)	88.76% (4)	69.42% (3)	88.59% (4)
law	86.04% (6)	58.80% (4)	86.10% (6)	65.15% (5)	86.06% (6)
literature	90.12% (1)	58.56% (5)	89.97% (1)	68.58% (4)	90.09% (1)
newsml	87.67% (5)	63.15% (2)	87.72% (5)	69.85% (2)	87.68% (5)
newspaper	89.22% (3)	63.69% (1)	89.25% (3)	72.48% (1)	89.22% (3)
student	89.68% (2)	54.32% (6)	89.70% (2)	63.04% (6)	89.69% (2)
Total	**88.59%**	—	**88.61%**	—	**88.60%**

Regarding our POS tagging results, in all our subsequent tables, we report three numbers per each cross-domain evaluation. The three numbers refer to the three kinds of experiments below:

1. only word identity features are utilized,
2. both word identity and sparse coding-derived features are utilized,
3. only sparse coding-derived features are utilized.

Next, we present our evaluation across the six distinct categories of Szeged Treebank according to the reduced MSD v2.5 tag set consisting of 93 labels. Table 3 and Table 4 contain our results depending on whether accuracies were calculated on the per-token or per-sentence level, respectively.

Table 3: Per-sentence cross-evaluation accuracies across the subcorpora of Szeged Treebank using a reduced tag set of MSD version 2.5 consisting of 93 labels.

Train \ Test	computer	law	literature	newsml	newspaper	student
computer	88.47%	80.00%	74.11%	81.37%	79.70%	76.55%
	92.57%	88.19%	83.86%	88.75%	89.28%	82.84%
	90.07%	85.91%	80.73%	86.66%	86.49%	80.34%
law	76.35%	93.52%	64.89%	70.61%	72.87%	67.70%
	86.24%	95.47%	75.65%	83.32%	85.41%	76.83%
	83.95%	92.69%	73.06%	80.90%	82.84%	74.48%
literature	73.63%	68.01%	88.17%	64.16%	75.21%	84.71%
	85.81%	82.51%	91.65%	81.40%	86.97%	88.66%
	83.34%	80.79%	89.15%	79.03%	84.65%	85.81%
newsml	86.73%	86.02%	76.72%	95.79%	87.20%	77.73%
	77.91%	76.64%	67.57%	93.28%	77.94%	70.88%
	84.57%	84.37%	75.27%	93.79%	85.11%	75.43%
newspaper	82.21%	80.90%	79.68%	86.61%	85.78%	81.00%
	89.26%	88.75%	86.48%	91.48%	91.32%	85.69%
	87.04%	86.44%	84.02%	88.77%	88.94%	82.70%
student	75.27%	70.65%	82.74%	72.71%	77.80%	91.53%
	85.15%	82.50%	88.18%	83.45%	87.23%	93.21%
	82.24%	79.32%	85.42%	80.12%	84.11%	89.80%

Subsequently, we evaluated our models according to all the possible combinations of the subcorpora relying on the coarser-level universal morphologies tag set which includes 17 POS tags. Results for the per-token and sentence-level evaluations are present in Table 5 and Table 6, respectively.

Comparing the results when evaluating according to the MSD tagset and the universal morphologies, we can observe that better results were achieved when evaluation took place according to the universal morphologies. This is not so surprising, however, as the task was simpler in the latter case, i.e. we faced a

Table 4: Per-sentence cross-evaluation accuracies across the subcorpora of Szeged Treebank using a reduced tag set of MSD version 2.5 consisting of 93 labels.

Train \ Test	computer	law	literature	newml	newspaper	student
computer	21.21%	3.79%	8.31%	2.92%	6.16%	6.39%
	30.93%	12.71%	18.88%	11.35%	18.20%	12.79%
	21.26%	9.54%	13.87%	8.42%	12.32%	9.54%
law	4.64%	31.17%	3.28%	0.81%	3.22%	3.01%
	13.37%	41.08%	6.68%	4.74%	10.90%	7.25%
	9.57%	24.38%	5.25%	3.68%	7.44%	5.50%
literature	3.70%	1.50%	36.43%	0.40%	6.26%	19.76%
	11.00%	5.08%	43.86%	2.62%	14.60%	26.49%
	8.24%	3.79%	34.91%	2.12%	10.09%	18.64%
newsml	4.64%	2.23%	3.22%	42.56%	4.79%	3.35%
	13.37%	8.97%	7.27%	50.68%	12.42%	7.23%
	9.92%	6.85%	6.68%	35.30%	8.58%	6.01%
newspaper	8.68%	4.62%	14.38%	6.61%	12.27%	11.75%
	19.14%	12.24%	25.03%	14.52%	23.36%	17.59%
	12.97%	9.08%	19.76%	10.14%	16.97%	13.07%
student	3.55%	0.99%	22.08%	0.76%	6.21%	40.09%
	10.71%	5.50%	31.58%	5.14%	14.41%	45.79%
	7.70%	3.37%	24.05%	3.23%	9.43%	31.49%

Table 5: Per-token cross-evaluation accuracies across the subcorpora of Szeged Treebank using the universal morphology tag set.

Train \ Test	computer	law	literature	newsml	newspaper	student
computer	90.66%	84.05%	78.54%	83.62%	81.84%	83.28%
	94.56%	91.63%	88.38%	91.63%	91.59%	90.52%
	92.35%	89.32%	86.29%	90.21%	89.30%	88.35%
law	78.18%	96.07%	70.07%	72.91%	75.94%	73.81%
	88.18%	97.67%	82.38%	86.90%	87.00%	84.38%
	86.43%	95.65%	80.35%	85.76%	85.51%	82.21%
literature	76.70%	75.64%	91.54%	66.17%	78.19%	88.90%
	87.54%	87.87%	95.16%	82.38%	90.05%	93.36%
	85.70%	85.69%	92.92%	80.49%	88.11%	91.23%
newsml	79.83%	81.36%	69.71%	94.50%	79.62%	75.02%
	89.51%	90.42%	85.19%	97.07%	90.70%	85.62%
	87.88%	88.96%	83.30%	95.58%	88.53%	83.33%
newspaper	84.08%	85.89%	83.48%	88.29%	88.38%	86.51%
	91.43%	91.93%	91.23%	93.59%	94.01%	91.96%
	89.89%	90.28%	89.55%	91.32%	91.85%	89.61%
student	77.49%	75.77%	85.41%	69.89%	79.61%	93.88%
	88.73%	87.97%	92.08%	85.74%	90.56%	96.04%
	85.83%	84.45%	90.28%	82.69%	88.22%	94.04%

17-class sequence classification problem, opposed to the 93-class problem for the MSD case.

Applying either kind of evaluation, the domain of newspapers seems to be the hardest one in the intra-domain evaluation, as the lowest accuracies are reported here. Also, we can notice that the *literature* and *student* domains are the most different from the others, as training on these corpora and evaluating against some other yields the biggest performance drops. Although *literature* and *student* writing being substantially different from all the other genres, they seem to be similar to each other, as the performance gap when training on one of these domains and evaluating on the other has milder performance gaps compared to other scenarios.

It can be clearly seen that models using features for both the word identities and sparse coding have the best results often by a large margin. It is not surprising as this model had access to the most information. When comparing the results of the models which either solely relied on word identity or sparse coding features, it is interesting to note that the model not relying on the identity of words ar all, but the sparse coding features alone, tends to perform better. A final important observation to make is that when sparse coding features are employed, domain differences seem to be expressed less, i.e. the performance drops in cross-domain evaluation settings tend to lessen.

Table 6: Per-sentence cross-evaluation accuracies across the subcorpora of Szeged Treebank using the universal morphology tag set.

Train \ Test	computer	law	literature	newml	newspaper	student
computer	26.64%	8.25%	13.85%	5.24%	10.66%	16.69%
	41.54%	23.91%	28.63%	20.42%	26.49%	31.89%
	29.26%	17.12%	22.88%	13.77%	19.62%	24.64%
law	5.97%	47.93%	5.49%	1.31%	4.50%	6.13%
	18.55%	63.28%	14.35%	7.87%	14.69%	15.59%
	13.37%	42.48%	11.96%	5.95%	12.23%	12.11%
literature	5.33%	3.53%	48.34%	0.61%	9.10%	31.23%
	17.56%	14.11%	60.51%	5.40%	22.70%	45.29%
	12.93%	9.75%	48.87%	3.53%	17.58%	35.07%
newml	6.36%	5.29%	5.17%	48.41%	7.58%	6.79%
	19.39%	17.84%	19.63%	59.51%	20.76%	17.73%
	13.32%	13.74%	16.14%	44.13%	14.88%	14.04%
newspaper	10.71%	8.82%	21.44%	12.15%	19.95%	23.16%
	27.97%	23.55%	39.52%	25.67%	36.35%	36.92%
	19.68%	17.43%	32.89%	17.35%	27.01%	27.84%
student	6.22%	2.75%	29.03%	1.01%	9.67%	50.89%
	17.07%	14.06%	44.82%	7.56%	24.08%	62.46%
	12.33%	9.60%	36.99%	4.74%	18.63%	48.76%

5 Conclusion

In this paper, we described our CRF-based POS-tagging model relying on the sparse coding of distributed word representations. We evaluated our proposed method on the subsections of the Szeged Treebank and found that the sparse coding derived features help to lessen the domain differences in cross-genre evaluation settings. We also found that relying on sparse coding features alone, it is possible to obtain better tagging accuracies than using word identity features and that combining the two sources of information can yield the best accuracies.

References

1. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 168–171
2. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 173–180
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3** (2003) 1137–1155
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08, New York, NY, USA, ACM (2008) 160–167
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12** (2011) 2493–2537
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11** (2010) 19–60
8. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed Word Representations for Multilingual NLP. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, Association for Computational Linguistics (2013) 183–192
9. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proceedings of EMNLP. (2014)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR **abs/1310.4546** (2013)
11. Makrai, M.: Comparison of distributed language models on medium-resourced languages. XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015) (2015) 22–33

12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
13. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007)
14. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)

Módosított morfológiai egyértelműsítés és integrált konstituenselemzés a magyarlanc 3.0-ban

Farkas Richárd[1], Szántó Zsolt[1], Vincze Veronika[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged Árpád tér 2.
e-mail:{rfarkas,szantozs,zsibrita}@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat Cikkünkben bemutatjuk a magyarlanc programcsomag [1] legújabb fejlesztéseinek számítógépes nyelvészeti szempontból érdekes tanulságait. Röviden érintjük a webes szövegek elemzésére kiterjesztett tokenizálót, majd hosszabb összehasonlítást közlünk a legmodernebb szófaji egyértelműsítők eredményeiről. Végül a szintaktikai elemzés terén vizsgáljuk különböző morfológiai kódkészletek hatását a függőségi elemzésre és bemutatjuk a magyarlancba integrált statisztikai konstituenselemzőnket. A magyarlanc programcsomag ingyenesen elérhető honlapunkon[3].

1. Bevezetés

A magyarlanc programcsomag [1] magyar nyelvű szövegek nyelvi előfeldolgozását hajtja végre a mondatra bontástól a morfológiai elemzésen át a szintaktikai elemzésig. Ebben a cikkben bemutatjuk az elemző lánc legújabb változatát. Először röviden ismertetjük a webes szövegekre (is) optimalizált tokenizálót, majd összehasonlítjuk a különféle szófaji egyértelműsítők teljesítményét. Bemutatjuk azt is, hogy az eltérő kódrendszerek miképpen befolyásolják a morfológiai és függőségi elemzők eredményességét, továbbá végül kitérünk a magyarlancba újonnan integrált konstituenselemző modulra is.

2. Robosztus tokenizáló

A magyarlanc v2.1 mondatra és tokenekre bontó első modulját újraimplementáltuk, majd robosztusabbá tettük, hogy a közösségi média sajátosságait is figyelembe vegye. A két legfontosabb ilyen kiegészítés az URL-ek felismerését célzó reguláris kifejezések beépítése, illetve egy emotikonazonosító szabályrendszer, ezek esetében két korábbi megoldásra építettünk[4,5], de kiegészítettük azokat speciális szabályokkal.

[3] rgai.inf.u-szeged.hu/magyarlanc

[4] https://gist.github.com/uogbuji/705383

[5] https://github.com/twitter/commons/blob/master/src/java/com/twitter/common/text/extractor/EmoticonExtractor.java

Egy fő elvárás a szegmentálóval kapcsolatban, hogy továbbra is a Szeged Korpusz [2] tokenhatárainak megfelelően szegmentáljon annak érdekében, hogy a korpuszon gépi tanított módszerekkel kompatibilis maradjon. A tokenizálót ezért a fejlesztés során folyamatosan lefuttattuk a Szeged Korpuszon és a Szeged Web Korpuszon [3], és annak kimenetét a gold standard tokenhatárokkal összevetettük, majd az esetlegesen szükséges változtatásokat beépítettük a rendszerbe.

3. Morfológiai egyértelműsítő rendszerek

A fejlesztés során módosítottuk a magyarlanc által alkalmazott szófaji egyértelműsítő rendszert. A módosításnak többféle motivációja is volt, egyrészről olyan licencet szerettünk volna használni, amely segítségével a magyarlanc alkalmazható ipari projektek részeként, másrészről az eddigiekben használt maximum entrópia Markov-modellre (MEMM) építő Stanford POS Tagger [4] mellett a legmodernebb szófaji egyértelműsítő rendszerek hatékonyságát is szerettük volna összevetni más szófaji egyértelműsítőkkel.

A kísérleteinkhez a Stanford POS Taggert két másik szófaji egyértelműsítővel hasonlítottuk össze. A PurePOS [5] egy morfológiai elemzővel kiegészített trigramokat használó rejtett Markov-modell (HMM) alapú elemző, míg a MarMoT [6] egy magasrendű feltételes véletlen mezőkre (CRF) építő szófaji egyértelműsítő.

A három elemző a háttérben használt matematikai modell mellett több dologban is eltér, ezek egyike a nyelvi erőforrások használata. A magyarlanc több nyelvi erőforrást is igénybe vesz a szófaji egyértelműsítés folyamatához. A meglévő szófaji címkéket először leképezi egy sokkal kisebb szófajicímke-halmazra (amelyből a szóalak ismeretében egyértelműen visszanyerhető az eredeti címke), majd az elemzés során morfológiai egyértelműsítő használatával szűri le az egyes szóalakokhoz tartozó lehetséges címkéket. Ezzel szemben a PurePos képes hatékony elemzést adni tisztán statisztikai módon, viszont a programban lehetőség van morfológiai elemző bekötésére, amivel tovább javítható a rendszer pontossága. A MarMoTot csak tisztán statisztikai módon, a tanítókorpuszon kívüli bármiféle nyelvi erőforrás használata nélkül alkalmaztuk.

Erőforrások szempontjából bár a kiértékelés mindhárom elemző esetén gyorsnak mondható, a tanítási időben nagy eltérések vannak a rendszerek között. A leggyorsabb a PurePos, amely másodpercek alatt képes egy modellt felépíteni a teljes Szeged Korpuszból. Ez a folyamat a MarMoT esetén azonos hardver mellett pár órát, míg a Stanford POS Tagger esetén napokat vesz igénybe.

3.1. Eredmények a Szeged Korpuszon

A rendszerek doménen belüli hatékonyságának vizsgálatához a Szeged Korpuszt vettük alapul. A Szeged Korpusz mind a 6 alkorpuszát véletlenszerűen felosztottuk 80-20 arányban tanító és kiértékelő korpuszra.

Az 1. táblázat az egyes rendszerek hatékonyságát tartalmazza a Szeged Korpusz egyes doménjein tanítva és kiértékelve. Az eredmények meghatározásához a

1. táblázat. Szófaji egyértelműsítők hatékonysága a Szeged Korpusz alkorpuszain.

	sz. tech.	jog	irodalom	rövidhír	újság	iskolás
magyarlanc	94,08	97,51	95,89	95,92	94,07	96,00
PurePos	94,15	97,09	94,06	97,35	93,63	95,27
Purepos + MA	94,75	97,39	95,90	96,88	94,33	96,01
MarMoT	95,88	97,73	95,74	98,03	95,75	96,32

teljes morfológiai leírás szerinti pontosságot használtuk, azaz mind a fő szófajnak, mind a morfológiai jegyeknek egyezniük kellett. A *magyarlanc* a magyarlancban eddigiekben is használt Stanford POS Tagger eredményeit tartalmazza. A $PurePos + MA$, illetve $PurePos$ a PurePos morfológiai elemzővel kibővített, illetve a nélküli változatát jelölik.

A legjobb eredményeket az – irodalmi szövegek kivételével – minden esetben a MarMoT érte el. Az irodalmi szövegek esetén a PurePos és Stanford POS Tagger holtversenyben végzett az első helyen.

A PurePos esetén átlagosan 0,61 százalékpontot javítva hat esetből ötször szerepelt jobban a morfológiai elemzőt is használó változat. A magyarlánc és a PurePos versenyében az előbbi több esetben tudott jobban szerepelni a morfológiai elemzőt nem használó PurePos változatnál. A morfológiai elemző használata mellett viszont a PurePos három alkorpuszon jobb, kettőn pedig közel azonos eredményt el, mint a magyarlanc.

3.2. Eredmények közösségimédia-szövegeken

A vizsgálatok során cél volt az is, hogy az elemző ne csak előre megszerkesztett (regények, újságcikkek, ...) szövegeken tudjon jól működni, hanem a nyelvi szabályokat sokkal kevésbé betartó internetes közösségi médiából származó szövegeken is hatékonyan működjön. Az elemzők számára a tanítóhalmaztól eltérő domén mellett az is kihívást jelent, hogy ezek a szövegek sokkal kevésbé szerkesztettek és ellenőrzöttek, mint a Szeged Korpuszban található egyéb szövegek. A mondat szerkezetében lévő eltérések mellett a közösségi médiából származó szövegek nagy mennyiségben tartalmazhatnak helyesírási hibákat vagy olyan szóalakokat, amelyek egyáltalán nem jellemzők az irodalmi, újságírói nyelvre.

A vizsgálatainkhoz két, közösségi médiából származó tesztkorpuszt [3] használtunk, mindkét esetben a teljes Szeged Korpuszon tanítottunk. A gyakori kérdések korpusz (faq) a gyakorikerdesek.hu oldalon feltett kérdésekből és arra érkező válaszokból áll, míg a $facebook$ korpusz Facebookról származó bejegyzéseket és a hozzájuk tartozó kommenteket tartalmazza. A két korpusz szerkesztettsége erősen eltér, hiszen míg a gyakori kérdések általában előre átgondolt és megszerkesztett kérdéseket és válaszokat tartalmaz, addig a facebookról származó bejegyzések sokszor csak egy hirtelen jött gondolatot fogalmaznak meg, és az alattuk található kommentek sokkal inkább hasonlítanak valós idejű társalgásra, mint átgondolt és előre megszerkesztett szövegre.

2. táblázat. Szófaji egyértelműsítők hatékonysága közösségi médiából származó szövegeken.

	facebook	faq
magyarlanc	67,17	84,46
PurePos	67,86	86,08
PurePos + MA	70,40	86,61
MarMoT	67,76	87,49

3. táblázat. Szófaji egyértelműsítés és lemmatizáció együttes hatékonysága a közösségi médiából származó szövegeken.

	facebook	faq
magyarlanc	65,00	82,37
PurePos	66,22	85,49
PurePos + MA	66,51	83,37
MarMoT	63,59	84,61

Az egyes rendszerek eredményeit a 2. táblázat tartalmazza. Minden esetben az egész Szeged Korpuszt használtuk tanításhoz és az egyes közösségi média korpuszokon értékeltünk ki. Ezúttal az elemzők sorrendje mindkét korpuszon azonos. A facebook esetén a PurePos teljesített a legjobban, a morfológiai elemzős változat 2,64 százalékponttal ér el jobb eredményt, mint a MarMoT. A gyakori kérdéseken viszont 1 százalékpont alatti különbséggel, de a MarMoT jobban teljesített. A magyarlanc mindkét esetben alulmaradt, ennek az indoka, hogy a rendszer nagyban támaszkodik a morfológiai elemző kimenetére. A morfológiai elemző viszont helyesírási hibák, lemaradt ékezetek esetén sokszor nem tud lehetséges elemzéseket meghatározni, az ilyen esetekben az adott szót mindig *X* (ismeretlen szó) címkével látja el a rendszer.

3.3. Lemmatizáció

A szófaji egyértelműsítés mellett fontos kérdés volt az egyes szóalakokra a megfelelő szótövek meghatározása. A Stanford POS Tagger külön szótövesítésre nem képes. A magyarlanc eddigiekben arra az állításra építve tudta meghatározni a szótöveket, hogy a magyarban a szóalak és a morfológiai címke ismeretében a szótő egyértelműen meghatározható. Így a szótő megadásához egy adott szóalakra a morfológiai elemző által adott lehetséges elemzéseket használtuk.

Ezzel szemben mind a PurePos, mind a MarMoT (Lemming [7]) tartalmaz beépített statisztikai lemmatizálót. A PurePos a lehetséges lemmákat képes szóvégződések alapján statisztikai módon, vagy ha rendelkezésre áll morfológiai elemző, akkor az alapján meghatározni.

A 3. táblázat tartalmazza a szótövesítés eredményeit. Az egyes értékek a teljes morfológiai címke és a szótő együttes eltalálásának a pontosságai. Amennyiben

a címkéket is nézzük, a PurePos mindkét esetben jobban teljesített a MarMoT-nál. Viszont meglepő módon a gyakori kérdéseken jobb eredményeket ért el a morfológiai elemzőt nem használó PurePos, mint az azt használó modell.

4. Morfológiai kódrendszer

Morfológiai címkekészletben áttértünk az ún. univerzális morfológia kategória-rendszerére [8]. Az univerzális morfológia célja – az Univerzális Dependencia Projekt keretében –, hogy egy olyan univerzális, azaz nyelvfüggetlen morfológiai kódkészletet hozzon létre, mely számítógépes nyelvészeti oldalról elősegíti a morfológiai elemzők és szófaji egyértelműsítők fejlesztését, továbbá elméleti nyelvészeti oldalról megkönnyítik az egyes nyelvek kontrasztív morfológiai vizsgálatát. A projekt további célkitűzése, hogy ezen elméleti reprezentációt szorosan követő korpuszokat és treebankeket hozzon létre. Jelenleg 33 nyelvre áll rendelkezésre univerzális dependencia és/vagy morfológiai annotáció, melyek egyike a magyar.

A Szeged Korpusz 2.5-ben használatos morfológiai kódokat automatikusan alakítottuk át az univerzális morfológiai kódkészletre. Ezt a folyamatot részletesebben [8] tárgyalja. Jelen munkánkban azt vizsgáljuk, hogy a két kódrendszer közti eltérések mennyiben befolyásolják a morfológiai és szintaktikai elemzés hatékonyságát. Ennek érdekében a Szeged Korpusz 2.5 kódkészlete és az univerzális morfológia közötti különbségeket empirikus kísérletekkel támasztjuk alá.

Annak érdekében, hogy bizonyos nehezebb nyelvtani jelenségeket külön is megvizsgálhassunk, kézzel összeállítottunk egy mondathalmazt, melyet mindkét kódkészletnek megfelelően beannotáltunk, majd a teljes Szeged Korpusz anyagán tanítva a MarMoT szófaji egyértelműsítőt, automatikusan leelemeztettük a mondatokat. A számszerű eredmények szerint az univerzális morfológián tanítva jobb teljesítményt nyújtott az elemző (92,31%-os pontosság), szemben a 2.5-ös kódkészlettel (91,45%), a különbség azonban nem jelentős. Kíváncsiak voltunk azonban arra is, hogy a két kódrendszer esetében mik a nehézséget jelentő nyelvi jelenségek, így megvizsgáltuk a morfológiai egyértelműsítő rendszer tipikus tévesztéseit.

A gyakorító és műveltető igék elemzése mindkét kódrendszernek kisebb nehézségeket okozott, illetve bizonyos homonim alakok tévesztése is előfordult mindkét kódrendszer esetében (pl. *hozzátok* igei és névmási elemzése). Az univerzális morfológia ugyanakkor helyesen elemzi a kötőszavakat, ellenben a 2.5-ös kódrendszer tévesztéseivel. Itt azonban meg kell említenünk azt a tényt, hogy az univerzális morfológia mindösszesen a kötőszavak alá- vagy mellérendelő jellegét jelöli a morfológiai jegyek között, míg a 2.5-ös kódrendszer azt is jelöli, hogy tagmondatokat vagy frázisokat köt-e össze az adott kötőszó. Természetesen ez utóbbi megkülönböztetés inkább szintaktikai, semmint morfológiai természetű, így egy további érvet szolgáltat az univerzális morfológia használata mellett, hiszen ilyen jellegű megkülönböztetésekre nincs szükség a morfológia szintjén.

A kétfajta kódrendszer hasznosságát megvizsgáltuk aszerint is, hogy mennyire nyújtanak hasznos kimenetet a függőségi elemzéshez. Ehhez a Szeged Treebank Népszava alkorpuszának univerzális dependenciára annotált verzióját hasz-

4. táblázat. Szófaji egyértelműsítés és függőségi elemzés hatékonysága 2.5-ös és univerzális morfológiai kódkészlet mellett.

	LAS	ULA	POS
2.5 kódkészlet	77,41	81,81	91,54
univerzális morfológia	76,23	81,01	92,34

náltuk, melynek 80%-án tanítottuk a magyarlancba beépített Bohnet parsert, és a maradék 20%-án pedig kiértékeltük a rendszert. A tanítás során predikált szófaji elemzést használtunk mind a 2.5-ös kódrendszer, mind az univerzális morfológia esetében. A számszerű eredmények szerint függőségi elemzésre nézve jobb teljesítményt érünk el a 2.5-ös kódkészleten (l. 4. táblázat). A teljes morfológiai kódokat tekintve az univerzális morfológia jobb eredményeket ér el, ami arra enged következtetni, hogy az könnyebben gépi tanulható.

A szintaktikai elemzéseket részletesebben is megvizsgáltuk, így fény derült arra, hogy a fontosabb nyelvtani szerepek (pl. alany, predikátum) azonosításában közel hasonló teljesítményt nyújt a két rendszer. A 2.5-ös morfológia főleg a névutós szerkezetek és az igekötők azonosításában múlta felül az univerzális morfológiát. Az univerzális morfológia előnyei közvetlenül a minőség- és mennyiségjelzők azonosításában mutatkoznak meg, illetve hatékonyabban képes kezelni az alárendelő mellékmondatok több fajtáját is.

A magyarlanc jelenlegi verziójába a nemzetközi trendeknek megfelelően az univerzális morfológiai kódrendszert integráltuk. Jövőbeli terveink között szerepel, hogy a fenti tapasztalatok alapján a szintaktikai elemzés hatékonyságát segítendő a 2.5-ös morfológia egyes jegyeit nyelvfüggő kiegészítésként felvesszük az univerzális morfológiai kódrendszerbe.

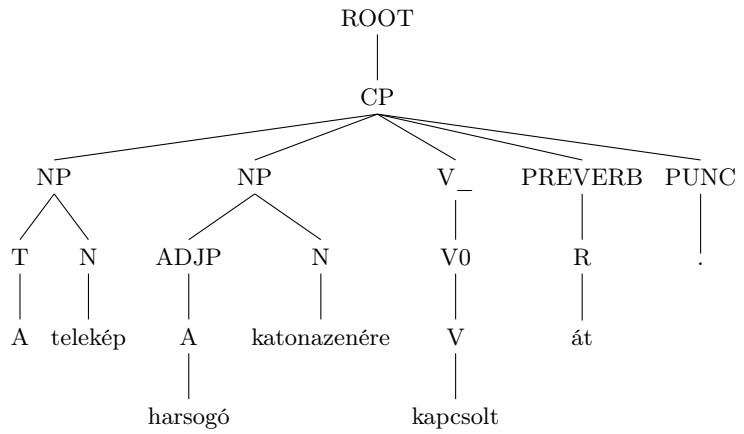
5. Konstituenselemzés

A szintaxis célja a mondatban rejlő nyelvtani kapcsolatok leírása. Az ilyen kapcsolatok megadására több eltérő reprezentáció is létezik. A számítógépes nyelvészetben a két legelterjedtebb reprezentáció a konstituensnyelvtanok és a függőségi nyelvtanok.

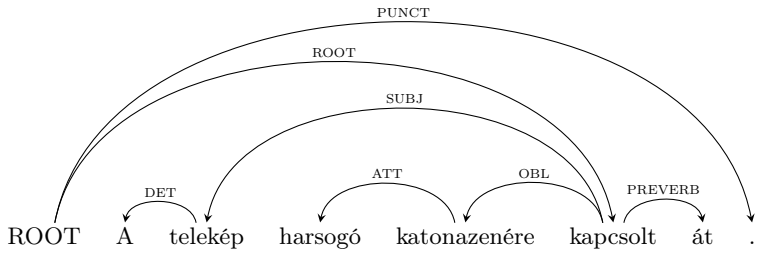
Az 1. ábra egy konstituensfát tartalmaz, amin jól látható, hogy a mondat úgynevezett konstituensekre van bontva, ezek a más szóval frázisoknak hívott egységek csoportba foglalják a szavakat (pl: NP – főnévi csoport). A fában a szavak a leveleken helyezkednek, a szófajok az úgynevezett preterminális rétegen a szavak felett, és e felett találhatók az egyes frázisok.

Ezzel szemben a függőségi fák (2. ábra) esetén a fa minden pontja egy szó és az élek a szavak közötti kapcsolatokat írják le.

A magyarlanc a korábbiakban már képes volt függőségi elemzések meghatározására, amihez a Bohnet parser [9] nevű nyelvfüggetlen függőségi elemzőt használta, a Szeged Dependencia Treebanken betanítva. Az új verzióban egy konstituenselemző modullal bővítettük a magyarlancot.



1. ábra. Konstituensfa.



2. ábra. Függőségi fa.

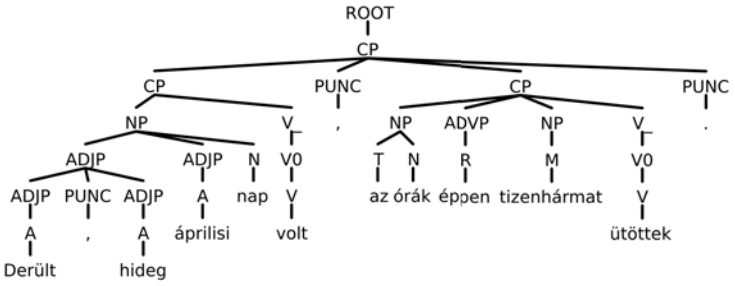
A magyar nyelv szintaktikai elemzése során probléma, hogy a meglévő rendszerek az angol nyelv igényeit figyelembe véve készültek. A magyar és az angol nyelv több szempontból nézve is alapjaiban különbözik. Az angol esetén a szintaktikai információk általában a szórendben tárolódnak, ezzel szemben a magyar nyelven a szintaxis a szavak szintjén jelenik meg toldalékok formájában. A konstituenselemzésre leggyakrabban alkalmazott valószínűségi környezetfüggetlen nyelvtanokra építő elemzők hatékonyságát nagyban rontja a toldalékolás következtében bekövetkező magas szóalakszám.

A morfológiailag gazdag nyelvek, köztük a magyar szintaktikai elemzésére hozták létre a Statistical Parsing of Morphologically Rich Languages workshop sorozatot. A magyarlancba készített rendszer a workshop keretében megrendezett SPMRL 2014 Shared Task [10] első helyezést elért rendszere [11] által bemutatott technikákra épül. Az elemző alapja a valószínűségi környezetfüggetlen nyelvtanokat alkalmazó Berkeley Parser [12].

A szóalakok nagy számának kezelésére a tanítóhalmazon nem, vagy csak ritkán látott szóalakokat lecseréljük a szófaji egyértelműsítés során megkapott fő szófaji kódra. A Berkeley Parser tanítása során kis mértékben szerepe van a véletlennek is, ennek a véletlennek a kiküszöbölésére 8 különböző modellt tanítottunk (eltérő random seed mellett), és predikáláskor a különböző modellek által egy mondatra adott valószínűségek szorzatát vesszük, így kiátlagolva a véletlen szerepét.

A valószínűségi környezetfüggetlen nyelvtanok hatékonyságának javítására úgynevezett újrarangsoroló rendszereket szoktak alkalmazni. Az alapgondolat az, hogy míg a környezetfüggetlen nyelvtannak az összes lehetséges elemzés közül kell választania, addig az általa választott legjobb k elemzésből egy lassú diszkriminatív gazdag jellemzőkészlettel rendelkező elemzővel kiválasztjuk a legjobbat. A magyarlancba is készítettünk egy újrarangsoroló rendszert, amit a területen általánosnak számító jellemzőkészletek [13,14] mellett morfológiai alapú jellemzőkkel bővítettünk [15]. Az így kapott rendszer a legaktuálisabbnak számít magyar nyelvű szövegek konstituenselemzésében.

A függőségi elemzéshez hasonlóan a magyarlanc képes vizuálisan is megjeleníteni a konstituenselemző által elkészített fákat. A megjelenítéshez a ParseTreeApplication[6] nevű fa vizualizációs szoftvert használtuk fel, a 3. ábra a magyarlanc egy példa kimenetét tartalmazza.



3. ábra. Konstituensfa a magyarlánc kimenetében.

6. Összegzés

Cikkünkben bemutattuk a magyarlanc programcsomag legújabb, 3.0 verzióját. Ennek része a webes szövegek elemzésre kiterjesztett tokenizáló, továbbá beépítettük a PurePOS morfológiai egyértelműsítőt és integráltunk egy konstituenselemző rendszert, amit morfológiailag gazdag nyelvek elemzésére dolgoztunk ki.

[6] https://github.com/ktrnka/ParseTreeApplication

A magyarlanc programcsomag ingyenesen elérhető: http://rgai.inf.u-szeged.hu/magyarlanc.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Hivatkozások

1. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013) 763–771
2. Alexin, Z., Gyimóthy, T., Hatvani, C., Tihanyi, L., Csirik, J., Bibok, K., Prószéky, G.: Manually annotated Hungarian corpus. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2, Association for Computational Linguistics (2003) 53–56
3. Vincze, V., Varga, V., Papp, P.A., Simkó, K.I., Zsibrita, J., Farkas, R.: Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, Szegedi Tudományegyetem (2015) 122–132
4. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 173–180
5. Orosz, Gy., Novak, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013) 539–545
6. Müller, T., Schmid, H., Schütze, H.: Efficient Higher-Order CRFs for Morphological Tagging. In: Proceedings of EMNLP. (2013)
7. Müller, T., Cotterell, R., Fraser, A., Schütze, H.: Joint Lemmatization and Morphological Tagging with Lemming. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics (2015) 2268–2274
8. Vincze, V., Farkas, R., Simkó, K.I., Szántó, Z., Varga, V.: Univerzális morfológia és dependencia magyar nyelvre. In: XII. Magyar Számítógépes Nyelvészeti Konferencia. (2016) 322–329
9. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. (2010) 89–97
10. Seddah, D., Kübler, S., Tsarfaty, R.: Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. (2014) 103–109
11. Björkelund, A., Özlem Çetinoğlu, Faleńska, A., Farkas, R., Müller, T., Seeker, W., Szántó, Zs.: Introducing the IMS-Wroclaw-Szeged-CIS entry at the SPMRL 2014

Shared Task: Reranking and Morpho-syntax meet Unlabeled Data. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. (2014) 97–102
12. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. (2006) 433–440
13. Collins, M.: Discriminative Reranking for Natural Language Parsing. In: Proceedings of the Seventeenth International Conference on Machine Learning. ICML '00 (2000) 175–182
14. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05 (2005) 173–180
15. Szántó, Zs., Farkas, R.: Special techniques for constituent parsing of morphologically rich languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden (2014) 135–144

Új integrált magyar morfológiai elemző

Novák Attila[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a
e-mail:{novak.attila}@itk.ppke.hu

Kivonat A morfológiai elemző a magyar nyelvtechnológiai alkalmazásokban kulcsszerepet játszik, hiszen minden nyelvi feldolgozási lánc első lépései között szerepel, így minden ezen a szinten keletkező hiba tovább terjed. Kulcsfontosságú tehát egy olyan morfológiai elemző és keretrendszer létrehozása, amely szabadon hozzáférhető, könnyen bővíthető és módosítható egy-egy adott felhasználási terület sajátos igényei szerint, és nem utolsó sorban jó minőségű elemzésre képes. Cikkünk célja egy ilyen morfológiai elemzőrendszer tervezési szempontjainak és a már létező morfológiai elemzőkre alapuló implementációjának bemutatása.

1. Bevezetés

Az MTA INFRA2 pályázatának keretein belül megvalósuló nyílt, integrált magyar nyelvtechnológiai kutatási infrastruktúra fejlesztésének célja egy olyan nyílt forrású, szabadon hozzáférhető nyelvtechnológiai infrastruktúra fejlesztése, melynek elemei a magyar nyelv gépi elemzésének alapvető eszközeit tartalmazzák egy integrált, szabványos keretben. A morfológiai elemző az infrastruktúra fejlesztésének központi eleme. A cél egy szabadon elérhető, az eddigi eszközök tudását szintetizáló, gyors és testre szabható elemző eszköz és a hozzá kapcsolódó komplex fejlesztőkörnyezet létrehozása.

2. Morfológiai elemzők magyar nyelvre

A jelenleg magyar nyelvre elérhető morfológiai elemzők (Humor [5,7], Xerox és Hunmorph/morphdb.hu [9]), illetve a hunmorph-foma[3] elemző minősége eltér, más-más nyelvi jelenséget tudnak jól kezelni, illetve tőtáruk is különböző, ezért különböző szókincset fednek le. Ugyanakkor az egyes elemzők leírásánál alkalmazott formalizmus is jelentősen eltér. Emellett az említett erőforrások jelentősen különböznek a leírás olvashatósága, illetve karbantarthatósága, a lefedett szókincs bővíthetősége és a forrás hozzáférhetősége szempontjából, illetve abból a szempontból is, hogy az erőforrás fejlesztői mennyire érhetőek el. Az utóbbi két szempont miatt a Xerox magyar elemzőjének forrásként való felhasználása

[3] http://freecode.com/projects/hunmorph-foma

nem jöhet szóba. A hunmorph-foma elemző a leírás olvashatósága, illetve karbantarthatósága, módosíthatósága, valamint a lefedett szókincs bővíthetősége szempontjából messze elmarad a Humor és a morphdb.hu alapú erőforrások mögött. Ez a leírás ugyanis nem nyelvtanon alapul, bővíteni kizárólag analógiás alapon lehet, a felvenni kívánt új szóval azonos morfológiai viselkedésű szó leírásának lemásolásával. Nem is beszélve a leírásban, illetve a paradigmákban levő esetleges hibák javításáról. Ezen kívül a gitorius.org lekapcsolásával a forrása pillanatnyilag nem hozzáférhető, és a fejlesztő elérhetősége is kérdéses. Mindezek miatt ennek az erőforrásnak a forrásként való használatáról is lemondtunk.

A morphdb.hu-n alapuló hunmorph (ocamorph) és jmorph elemzők nagy előnye, hogy nyílt forráskódúak, és maga a morfológiai adatbázis egyrészt nyelvtanalapú, így jól bővíthető, javítható, a forrása alapján értelmezhető, másrészt teljesen szabadon felhasználható és módosítható. A morphdb.hu és a morfológiai leírás alapjául szolgáló hunlex eszköz hátránya ugyanakkor, hogy ezek fejlesztése sok évvel ezelőtt leállt, a dokumentáció hiányos, és bár az egykori vezető fejlesztő, Trón Viktor kifejezte együttműködési készségét a projekt résztvevőivel, egyrészt csak nagyon korlátozott időben tud a rendelkezésünkre állni, másrészt a rendszer nem dokumentált tulajdonságai és a befejezetlen fejlesztések részleteivel kapcsolatban nem nagyon tudott segíteni, mert az évek folyamán a kérdéses tudás feledésbe merült. Ugyancsak nem könnyíti meg a helyzetet, hogy a hunlex implementációja OCaml nyelven íródott, és nem áll rendelkezésünkre ezen a nyelven kompetens programozó. Ugyancsak OCaml nyelven íródott az ocamorph elemző, ezzel szemben a jmorph Java alapú. A két elemzőeszköz azonban különbözik a bennük implementált elemzőalgoritmus, különösképpen az összetételi konstrukciók kezelése szempontjából. Ugyanakkor a különbségek nem dokumentáltak, pusztán az eszközök viselkedésének vagy forráskódjának tanulmányozása révén tárhatók fel.

A Humor elemző alapjául szolgáló morfológiai adatbázis forrása az ebben a cikkben említett projektum elindulásáig nem volt szabadon hozzáférhető, és maga a Humor elemző is zárt forráskódú. Ugyanakkor ez az erőforrás is nyelvtanalapú, így jól bővíthető, javítható, illetve viszonylag jól dokumentált. Az említett erőforrások közül egyedüliként a fejlesztő elérhető, és a rendszer fejlesztése annak létrehozása óta folyamatos. A Humor morfológiai adatbázis jellemzői az 1. táblázatban láthatók.

3. Az elemzők fedésének kiértékelése

Első lépésként a további fejlesztés szempontjából szóba jövő elemzőknek (Humor, ocamorph, jmorph) a részletes minőségi kiértékelése, és kritikai összehasonlítása volt a feladat egy nagy szöveges korpuszból készült gyakorisági lista segítségével. Egy 4 milliárd szavas, nagyrészt webről letöltött szövegből készült, elemi tokenizálóval tokenizált, nem szűrt 35 millió szavas gyakorisági listát elemeztettünk a Humor, az ocamorph és a jmorph elemzőkkel. Az ocamorph elemző produktív összetettszó-elemző üzemmódját bekapcsolva sajnos az input szavak egy részére végtelen ciklusba került, ezért egy idő után kikapcsoltuk ezt az üzemmódot.

1. táblázat. A magyar Humor morfológia jellemzői

tőlexikon		**lemmák/lexémák**	**allomorfok**
általános szókincs		**95811**	**141718**
	eredeti lexikon kibővítve	75132	105473
	zárt tőosztályok (névmások, számnevek, stb.)	744	3675
	szótárakból és korpuszokból	19935	32570
terminológiai lexikonok		**110129**	**178324**
	Földrajzi és személynevek	40262	
	Nukleáris terminológia	911	
	Gazdaság/adminisztráció	4736	
	Angol	1920	
	Orvosi	40813	
	Katonai	21487	
összes		**205940**	**320042**
	ebből összetett	89415	126728
	sokmorfémás/toldalékolt		7720

toldaléklexikon		**lexémák**	**allomorfok**
	összes	**283**	**12041**
	sokmorfémás		10959

szabály file-ok	**műveletek**	**szabályok**	**sorok**
tőszabályfile			
	45 deklaráció	520 szabály	2074 sor
	596 allomorfgeneráló művelet	220 tő allomorfiaszabály	
toldalékszabályfile		50 szabály	233 sor
	86 allomorfgeneráló művelet	34 allomorfiaszabály	

	állapotok	**átmenetek**	**flagek**
szónyelvtan-automata			
	47 állapot	602 átmenet	20 flag

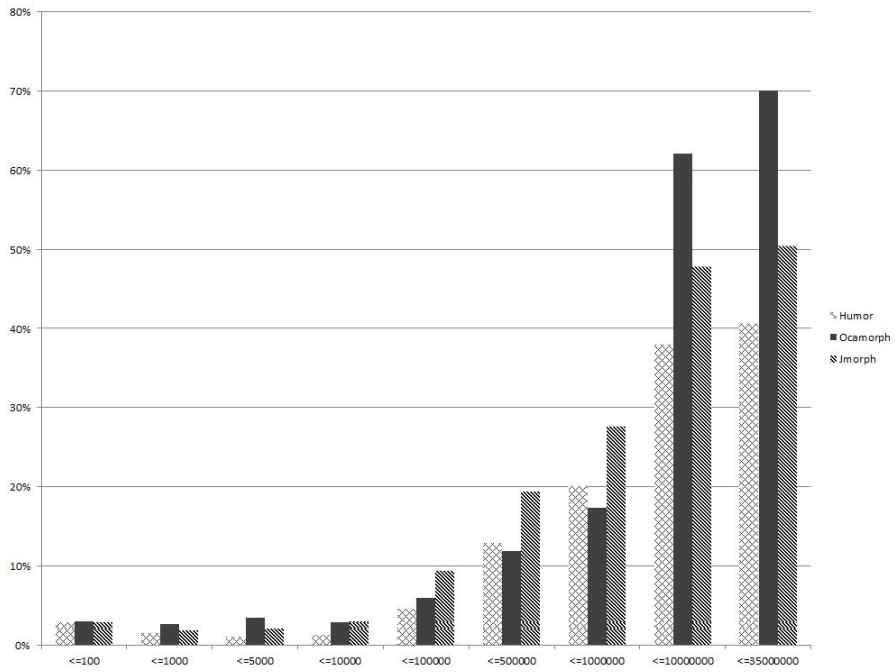
	kategóriák	**tulajdonságok**
a jegyek és szónyelvtan-kategóriák kódolásának definíciója		
	102 szónyelvtan kategória	102 vektorkódolt jegy
		187 mátrixkódolt jegy

Emellett mivel a morphdb.hu nyelvtanban az összetett szavak szerkezetét leíró konstrukció egyszerűen tetszőleges nominális (főnévi, melléknévi, számnévi) tövek tetszőleges sorrendben tetszőleges számban való megjelenését megengedi

(tehát pl. a *tevepirosnegyven* szó helyes számára), ezért sokkal több értelmetlen elemzést ad a produktív elemzés bekapcsolása esetén, mint a másik két elemző. A futtatás során a 2. táblázatban látható eredményeket kaptuk az egyes elemzőkre. A 1. ábrán látható, hogy az 500000. és a 10000000. szó közötti régióban az ocamorph adta a legnagyobb lefedést, a többi régióban a Humor.

2. táblázat. A nem elemzett szavak száma a 35 millió szóalak közül az egyes elemzők esetén.

Elemzőrendszer	Nem elemzett szavak száma
Humor	13 754 680
ocamorph	23 248 165
jmorph	17 152 815



1. ábra. A három elemző által nem elemzett szavak aránya gyakorisági rangsor egyes régióiban

A jmorph és az ocamorph ugyanazon lexikai adatbázis feltehetőleg különböző változatainak felhasználásával készült. Abban, hogy jelentősen különböző eredményt adnak, az elemzőkben implementált algoritmusok különbsége is szerepet játszik. A gyakori régióban előforduló hiányok elsősorban tokenizálási, központozási és helyesírási hibákból adódnak.

4. Az új morfológiai elemzőrendszer

A kifejlesztendő morfológiai elemzőrendszer felépítése három rétegből áll. Az első a felhasználó, nyelvész szakértő által olvasható morfológiai forrásadatbázis, azaz a tőtár és a morfo(fono)lógiai nyelvtan. A második réteg egy forrásadatbázis-konverter, mely a harmadik réteg számára szükséges erőforrásokat állítja elő az első rétegből. A harmadik réteg pedig maga az elemzőfuttatási keretrendszer.

Mind az ocamporh és jmorph elemzők , mind a Humor elemző által használt adatbázisok egy ugyanígy hármas tagolású rendszerben jönnek létre. A projekt jelenlegi fázisában a Humor nyelvtan és a [6] cikkben leírt véges állapotú nyelvleírást előállító konverter segítségével generáljuk az elemző adatbázisát.

4.1. A tőtár

A forrásadatbázis a Humor és morphdb.hu adatbázisok tőadatbázisának és morfológiai szabályrendszerének szintéziseként áll elő. Támogatja a pragmatikai, nyelvhasználati, szemantikai és morfológiai jellemzők rögzíthetőségét és kezelését. Pragmatikai jellemzők alatt a különböző stílusminősítő jegyeket, a helyesírási normától való eltérést jelző tulajdonságokat, illetve gyakorisági információkat értünk. Ezek a korábbi elemzők egyikében sem voltak egyszerre jelen. A forrásadatbázis a szemantikai jegyek, az ontológiai besorolás mellett a tematikus/vonzatkereteket és az időaspektusra vonatkozó jegyeket is tartalmazza. A morfológiai jegyek a Humor és a morphdb.hu lexikonok kategória-rendszerének egyesítésével lettek meghatározva, a kettő uniójaként, az esetleges ütközések és ellentmondások feloldása mellett. Bár a Humor adatbázis eleve tartalmazott az elemzésben meg nem jelenített szemantikai jellegű címkéket, ezeket a projektum során disztribúciós szemantikai modellek felhasználásával kibővítjük és ellenőrizzük [8].

A Humor adatbázisában szereplő mintegy 200000 lemmán túl a morphdb.hu adatbázisa kb. 13000 új lemmát tartalmaz, bár ezek egy része a helyesírási normának nem megfelelő alak. Jelenleg folyamatban van ennek a 13000 alaknak az ellenőrzése és a helyesírási normáknak nem megfelelő alakok leképezése a megfelelő helyes alakokra. Ehhez a Humor tőadatbázisából és az Osiris Helyesírás [3] szótári részében szereplő szavakból és többszavas kifejezésekből épített listán az A* algoritmust [2] futtatva és hibamodellként tévesztési mátrixot definiálva rangsorolt javítási javaslatokat generáltunk, amelyeket a kézi ellenőrzés támogatására használunk.

4.2. A kategória-rendszer

A főkategóriák mellett alkategóriákat is bevezetünk, melyek egyrészt a Humor/morphdb.hu rendszerben szereplő, de a végső kategória-rendszer főkategóriái közé nem kerülő kategóriák, másrészt az egyéb szemantikailag vagy morfoszintaktikailag releváns kategóriák. Továbbá a szokásos morfológiai jegyek: inflexiók, képzők, szóösszetételi határok és típusok is a két elemző adatbázisának egyesítésével kerülnek bevezetésre. A létrejött elemző fonológiai jellemzők kezelésére is alkalmas. Ez magában foglalja a szóalakok CV-vázának meghatározását (a rövid-hosszú szegmentumok opcionális megkülönböztetésével). Kezeljük továbbá a a felszíni alakból automatikusan nem levezethető kiejtéseket, illetve kiejtésváltozatokat és a nyelvjárási és szociolektális megkülönböztetéseket.

Ehhez áttekintettük a morphdb.hu nyelvtan hunlex nyelven megírt forrását és megkerestük azokat a pontokat, ahol a Humor leírásba átemelhető eltérések vannak. A morphdb.hu adatbázisban a határozószavak számos alosztályra vannak osztva. Ezt az osztályozást disztribúciós módszerrel validáltuk és kiterjesztettük. Ehhez a nyelvtechnológiai kutatások egyik kurrens módszerét, a neurális hálózattal létrehozott folytonos vektoros reprezentációkból álló modellt alkalmaztuk (*word embedding*). Ez a módszer nyers szöveges korpuszból szemantikai és grammatikai információk kinyerésére alkalmazható úgy, hogy az eredményül kapott modellben a lexikai elemek egy valós vektortér egyes pontjai, melyek konzisztensen helyezkednek el az adott térben, azaz a hasonló disztribúciójú, egymáshoz szemantikailag, szintaktikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. A modell létrehozásához a word2vec[4] eszközt használtuk a [8] cikkben ismertetett módon. Az eredmények számos hibára rámutattak a morphdb.hu határozószavaihoz rendelt kategorizációban, melyek kézzel történő javítása a modell alapján jelentős mértékben egyszerűsödött, hiszen a hasonló kategóriába tartozó szóalakok egymáshoz közel helyezkednek el a térben, így a csoportok egyszerre kategorizálhatók, illetve a hibásan kategorizált szavak kakukktojásként könnyen feltűnnek a hozzájuk hasonló viselkedést mutató, de másképp címkézett szavak között.

Érdekes megfigyelés a modellel kapcsolatban, hogy különböző távolságmetrikákkal számolva, illetve a korpusz különböző (pl. lemmatizált és elemzett vs. elemzetlen) változatain betanítva a modellt, láthatóan különböző szempontok dominálnak az osztályozásban. Ez egyben a Humor adatbázisban különböző egyébként nagyjából ekvivalens elemzésekkel kapcsolatos problémákra is rávilágított. Mikor a morphdb.hu határozószóként annotált lexémáit az (alapvetően a Humor elemző és a Szeged Korpusz felhasználásával) egyértelműsített és lemmatizált korpuszváltozatot használva klasztereztük, létrejött egy nagy méretű, grammatikai szempontból igen heterogén elemeket tartalmazó klaszter, amely – mint kiderült – olyan elemeket tartalmazott, amelyek a korpuszban főleg valamilyen mélyebb elemzéssel a szóalaktól különböző lemmára visszavezetve szerepeltek, de néhány elemzésük mégis atomi határozószói elemzést kapott (a tanítóanyag nem egységes annotációja miatt). Ezek a ritka elemzések azonban nem

[4] https://code.google.com/p/word2vec/

adtak megbízható disztribúciós modellt, így a furcsa klaszterben ezek a heterogén elemek jelentek meg, amelyekben leginkább az volt a közös, hogy az atomi határozószói elemzésük inkább ritka és zajszerű kivétel volt a másik részletesebb elemzésük mellett.

A disztribúciós elemzés hatásosan kimutatott egyéb annotációs anomáliákat is a korpuszban. Például jól elkülönülő klaszterekben gyűltek össze a Humor elemző által fel nem ismert, és így a PurePos tagger guessere által azonos módon hibásan elemzett és lemmatizált szavak (pl. azok, amelyekről mind tévesen vágta le az *-it* végződést, azt az *-i* képző tárgyragos alakjának vélve, vagy a téves anaforikus birtokos elemzés nyomán az *-é* levágásával, illetve a *-ba/be* végződés téves levágásával kapott hibás lemmák). A 3. táblázatban ilyen, illetve a korpuszban eleve hibásan írt vagy elválasztott és a tokenizáló által helyre nem állított szóalakokhoz tartozó klaszterek első néhány elemére láthatunk példát.

3. táblázat. Példák hibásan lemmatizált, tokenizált vagy hibásan írt kifejezésekre és a rendszer által azonosított hozzájuk hasonló szavak listájának első néhány eleme

pufidzsek	angolúl	kony	gyűrűj	sebti
rövidnac(43)	magyarúl(486)	nű(165)	királynőj(30)	Juteszem(33)
napszemcs(37)	németűl(132)	lyos(227)	Manass(104)	útköz(323)
szemcs(37)	francziául(25)	legha(17)	Hekat(91)	juteszem(94)
szmöty(45)	angolol(27)	komo(167)	oké-ok(412)	subscri(55)
zacs(170)	írül(95)	latos(283)	Lüzisztrat(21)	neszójjá(11)
suzuk(131)	mindenrõl(422)	legki(36)	juhéé(97)	kizom(25)
sap(374)	minderrõl(129)	csó(183)	jóskán(60)	akurvaélet(33)
törcs(11)	ilyenröl(58)	pontosab(59)	örüjj(35)	Egyfolyta(21)
kispolszk(41)	Amirõl(143)	nyolult(18)	Béb(774)	hébehó(25)
févör(8)	olyasmirôl(38)	kes(2114)	hoppár(189)	CsimbaWam(20)

A kötőszavak osztályozása a Humor adatbázisban és a morphdb.hu-ban részben különbözik. A Humor leírásban csak azok a szavak kaptak kötőszó címkét, amelyeknek a tagmondaton belüli disztribúciója egyértelműen kötőszószerű eloszlást mutat, és a pusztán a pragmatikai funkciójuk szempontjából tagmondatok közötti kötőelemként működő, de egyébként a tagmondatbeli elhelyezkedésük szempontjából határozószószerű viselkedést mutató szavak következetesen határozószóként vannak osztályozva. Ezzel szemben a morphdb.hu-ban ezek egy része is kötőszó címkével szerepel. Az egységesített címkézési rendszerben ezeket a határozószók egy alosztályaként címkézzük. Ugyancsak nem kötőszóként szerepelnek az egyébként a kötőszók eloszlásának megfelelő viselkedést muta-

tó vonatkozó névmások, hanem ezeket a főkategóriájukat megtartva vonatkozó névmásként alkategorizáltuk. A morphdb.hu nyelvtan áttekintése után a morphdb.hu lexikonban szereplő elemek jegyeit is automatikusan át tudjuk vinni a Humor lexikonba. Ugyanakkor kézi ellenőrzést is igényel a folyamat, mert a megadott jegyek sok esetben tévesek, vagy hiányoznak. Például sok helyesírási anomáliát tartalmazó szótő nincs ilyenként megjelölve.

4.3. A keretrendszer

A létrejött forrásadatbázis nyelvész szakértő számára olvasható, értelmezhető és plain text editorral szerkeszthető. Olyan kiegészítő alkalmazás fejlesztésére is sor kerül, amely a tőtár bővítését és megváltoztatását számítógépes nyelvészeti szakértelem nélkül is lehetővé teszi.

A második réteget képező forrásadatbázis-konverter a forrásadatbázis tőtárából *lexc* [1] lexikont állít elő, melyet a HFST keretrendszerben [4] implementált harmadik réteg használ fel.

A létrejövő keretrendszer lehetővé teszi a forrásadatbázisban specifikált információ alapján testre szabható domén- és regiszterspecifikus elemzők előállítását. Kimeneti kódkészletként bármely, magyar nyelvre elterjedt kódrendszer (KR, Humor, MSD) választható, illetve a már meglévő annotált korpuszokkal való kompatibilitás is biztosított. A lemmatizálás során a tő részét képező toldalékolás, illetve az elemzés során figyelembe vett morfofonológiai jellemzők konfigurálhatók.

5. Konklúzió

A cikkben egy olyan készülő nyílt forráskódú magyar morfológiai elemző létrehozására irányuló fejlesztést mutattunk be, amely több korábban készült magyar számítógépes morfológiai leírást harmonizálva és egyesítve, illetve azt további lexikai tudással kiegészítve terveink szerint minden korábbinál teljesebb és pontosabb eszköz lesz. Folyamatban van a rendszer alapjául szolgáló Humor és a morphdb.hu magyar morfológiai leírások harmonizációja és egyesítése. A lexikon ellenőrzéséhez és további lexikai jegyek félautomatikus felvételéhez folytonos vektortéralapú disztribúciós modelleket és automatikus hierarchikus klaszterezőalgoritmust használunk, amelyek igen hatékony eszköznek bizonyultak az elvégzendő lexikográfiai munka támogatásához.

Hivatkozások

1. Beesley, K., Karttunen, L.: Finite State Morphology. No. 1 in CSLI studies in computational linguistics: Center for the Study of Language and Information, CSLI Publications (2003)
2. Huldén, M.: Fast approximate string matching with finite automata. Procesamiento del Lenguaje Natural 43, 57–64 (2009)
3. Laczkó, K., Mártonfi, A.: Helyesírás. A magyar nyelv kézikönyvtára, Osiris (2004)

4. Lindén, K., Silfverberg, M., Pirinen, T.A.: Hfst tools for morphology - an efficient open-source package for construction of morphological analyzers. In: Mahlow, C., Piotrowski, M. (eds.) SFCM. Communications in Computer and Information Science, vol. 41, pp. 28–47. Springer (2009)
5. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
6. Novák, A.: A new form of humor – mapping constraint-based computational morphologies to a finite-state representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207
7. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
8. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In: Tanács, A., Varga, V., Vincze, V. (eds.) XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2016)
9. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of LREC 2006. pp. 1670–1673 (2006)

III. Beszédtechnológia

Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása

Tarján Balázs[1,2], Varga Ádám[2], Tobler Zoltán[2], Szaszák György[1,2], Fegyó Tibor[1,3], Bordás Csaba[4], Mihajlik Péter[1,2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
tarjanb@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.
mihajlik@thinktech.hu
[3] SpeechTex Kft.
tfegyo@speechtex.com
[4] Médiaszolgáltatás-támogató és Vagyonkezelő Alap (MTVA)

Kivonat: Cikkünkben egy valós idejű, kis erőforrás-igényű gépi beszéd-szöveg átalakító rendszert mutatunk be, melyet elsősorban televíziós közéleti társalgási beszéd feliratozására fejlesztettünk ki. Megoldásunkat összevetjük a tématerületen legelterjedtebben használt nyílt forráskódú keretrendszer, a Kaldi dekóderével is. Ezen felül különböző adatbázis-méretek mellett és újrabeszélés alkalmazásával is végzünk felismerési kísérleteket. Kísérleti rendszerünkkel, mely egy több mint 70 millió szót tartalmazó szövegkorpuszon és egy közel 500 órás beszédadatbázison lett tanítva sikerült az eddig publikált legalacsonyabb szóhibaarányt elérnünk magyar nyelvű, televíziós híradók és közéleti társalgási beszéd témakörén.

1. Bevezetés

Világszerte egyre szigorúbb törvények írják elő a televíziós társaságoknak, hogy a kép és hang mellett feliratot is sugározzanak, melynek célja a műsorok akadálymentesítése a siket és nagyothalló nézők számára. Bizonyos műsorok feliratozása kis ráfordítással is megoldható, mert a feliratok rendelkezésre állnak (pl. filmek) vagy elkészíthetőek kézi úton (pl. felvett műsorok). **Élő műsorok** esetén azonban nincs, vagy csak nagyon korlátozottan van lehetőség a hagyományos, manuális módszerek alkalmazására. Cikkünkben bemutatunk egy nagyszótáras beszédfelismerő rendszert, melyet elsősorban közéleti- és hírműsorok gépi úton történő, élő feliratozásához fejlesztettünk a Médiaszolgáltatás-támogató és Vagyonkezelő Alappal (MTVA) folytatott kutatás-fejlesztési együttműködésünk keretében.

Egy ilyen automata feliratozó rendszer fejlesztése többféle kihívást tartogat. A nemzetközi irodalomban a legtöbb eredmény híradók felismerésével született, mely jól artikulált, felolvasott szövegen alapuló beszédnek tekinthető. Ezzel szemben a közéleti, politikai műsorok gyakran két- vagy akár többszereplős párbeszédeket, illetve spontán megfogalmazásokat tartalmaznak, melynek felismerése a **párhuzamos beszédszakaszok** és a **lazább artikuláció** miatt jóval nehezebb feladat. További kihívás, hogy az

élő műsorok feliratozásához valós időben működő rendszert kellett terveznünk, mely ráadásul akár öt közszolgálati csatorna párhuzamos feliratozására is képes. Mindez még egy manapság korszerű szerveren sem egyszerű feladat a nagyszótáras felismerő rendszerek magas erőforrásigénye miatt.

Célunk tehát, hogy bemutassuk a rendszer fejlesztése során kipróbált módszereket és azok hatását a felismerési hibára valamint az erőforrásigényre. Összehasonlítjuk az igen elterjedt, nyílt forráskódú **Kaldi** programcsomag [1], valamint a SpeechTex Kft. által rendelkezésünkre bocsátott **VOXerver** [2] súlyozott véges állapotú átalakítókon (Weighted Finite State Transducer – WFST) alapuló beszédfelismerő dekódereket. Megvizsgáljuk továbbá, hogy mekkora előnnyel járhat, ha a műsorokat nem közvetlenül feliratozzuk, hanem közbeiktatunk egy ún. **újrabeszélőt**, aki elismétli az elhangzottakat. Mindezek mellett különböző méretű akusztikus és szöveges tanítókorpusz mellett is meghatározzuk a rendszer hibáját és erőforrásigényét, valamint **mély neurális hálózatokon** (Deep Neural Network - DNN) alapuló akusztikus modelleket is alkalmazunk a további hibacsökkentés érdekében.

A következő fejezetben cikkünk témaköréhez legjobban illeszkedő nemzetközi és hazai eredményeket mutatjuk be. Ezután a kísérleti feliratozó rendszerünk felépítését, valamint tanító- és tesztadatbázisait, majd a negyedik fejezetben az erőforrás-igényeket és pontosságokat meghatározó méréseink eredményét ismertetjük. Végül az utolsó fejezetben összefoglalását adjuk vizsgálataink legfontosabb eredményeinek.

2. Kapcsolódó eredmények

A legtöbb televíziós műsor felismerésével kapcsolatos publikáció **híradók** leiratozásával foglalkozik. Kutatócsoportunk korábbi eredményei [2–4] 10-50 óra híradós kézi leirat alapján tanított Gaussian mixture modell (GMM) alapú akusztikus modellel és webről gyűjtött szövegeken alapuló nyelvi modellekkel készültek, melyekkel átlagosan 21-27%-os szóhiba-arányt értünk el híradókon. Hasonló mértékű, kb. 24%-os szóhibaarányt említenek [5]-ben, ahol a felügyelet nélküli tanításra helyezték a hangsúlyt. Sajnos azonban ezt az eredményt nem könnyű összehasonlítani másokéval, mivel a tesztanyag egyszerre tartalmazott híradókat és közéleti társalgási beszédet is. Az eddigi legjobb magyar nyelvű híradó-felismerési eredmény legjobb tudomásunk szerint [6]-ban található, ahol 17%-os szóhiba-arányról számoltak be, melyet web-alapú tanítószöveg és DNN akusztikus modell segítségével kaptak.

Közéleti társalgási beszéd közvetlen, tehát újrabeszélés nélküli átírására kevesebb példa van, különösen magyar nyelven. Az egyetlen, melyről tudunk egy korábbi munkánk [2], ahol a híradók felismerésére optimalizált rendszerünket teszteltük televíziós beszélgetéseken is. Az elért 50%-os hibaarány azonban magasnak mondható. Általában a nemzetközi sztenderd ezen a feladattípuson 20-30% között mozog [7–9]. Természetesen az ilyen kis hibájú rendszerek nagy mennyiségű feladatspecifikus hang- és szöveganyagon lettek tanítva, ám szerencsére ilyenek a jelenlegi feladatnál már számunkra is rendelkezésre állnak.

Cikkünkben bemutatott **újrabeszélés**i eredményeket nehéz összehasonlítani a külföldi megoldásokkal. A gyakorlatban is működő újrabeszélt feliratozás általában meg-

lepően nagy, 5-12 másodperces, de eseteként akár 18 másodpercesnél is nagyobb késleltetéssel dolgozik [10]. Ennek egyrészt az az oka, hogy minőségbiztosítási célból az újramondott és felismert feliratok még egy manuális hibajavítási fázison is átesnek, mely nyilvánvalóan késleltetéssel jár. Másrészről a tapasztalatok szerint a feliratok információtartalmát érdemes 125-160 szó/perc körüli értékre csökkenti, hogy ne vonja el a néző figyelmét túlzottan a képről [11]. Ez utóbbi azonban szintén azt jelenti, hogy az újrabeszélőnek be kell várnia bizonyos mennyiségű információt, hogy aztán abból **kivonatot** készíthessen. Ezzel szemben jelenlegi kísérleteinkben **szó szerinti** újrabeszélést kértünk az újrabeszélőktől, és a teljesen **valós idejű** működésre koncentráltunk.

3. A kísérleti rendszer

Ebben a fejezetben a gépi beszéd-szöveg átalakító rendszerünk tanítása és tesztelése során felhasznált adatokat és módszereket ismertetjük.

3.1. Akusztikai modellezés

3.1.1. Akusztikai tanító-adatbázisok

A kísérleteink során alkalmazott akusztikus modelleket két különböző méretű beszéd-adatbázison tanítottuk. Az első adatbázis 64 óra, kézzel annotált, webes híradót tartalmazott (**webes híradók**). Ezt az adatbázist használtuk a kezdeti modellek tanításához és a dekóderek erőforrásigényének felméréséhez.

A második adatbázisba (**kiterjesztett adatbázis**) a kezdeti modell elemei mellé más beszéd-adatbázisokat is felvettünk, melyek reményeink szerint tovább növelik a feliratozó rendszer pontosságát és robosztusságát. A négy kiegészítő adatbázis tartalma és jelölése a következő:

- *Közéleti társalgási beszéd leirata* (**Közéleti hírműsorok**): 31 óra manuálisan címkézett közéleti televíziós beszélgetést tartalmaz, melyet az MTVA bocsátott rendelkezésünkre
- *Félig felügyelten annotált MTVA adatbázis* (**FF**): Az MTVA által rendelkezésünkre bocsátott televíziós felvételek egy részéhez felirat is tartozott. Ezek a feliratok nem mindenhol követték hűen a műsorban elhangzottakat, így közvetlenül nem voltak alkalmasak akusztikus modell tanítására. Ezt a problémát félig felügyelt tanítóanyag-válogatás [12] segítségével kezeltük, így a 136 órányi felirattal rendelkező hanganyagból összesen 100 órát válogattunk ki tanítási célra.
- *Egri Katolikus Rádió adatbázis* (**EKR**): 65 órányi beszélgetést és hírfelolvasást tartalmaz ez az adatbázis, melyet az Egri Katolikus Rádióban rögzítettek.
- *Speecon beszédadatbázis* (**Speecon**): A 2000-ben indult Speecon projekt [13] célja az volt, hogy változatos környezetben rögzített beszédadatbázisokkal segítse a beszédfelismerő rendszerek tanítását. Mi ennek az adatbázisnak a magyar nyelvű változatát használtuk, azon belül is az irodai és otthoni környezetben gyűjtött felvételeket. A négy rögzített mikrofonjel közül kettőt használtunk fel az akusztikus modellünkben, így adódott a 2x114 órás adatbázisméret.

A kísérleteink utolsó fázisában használt rendszer akusztikai tanító-adatbázisa így összesen közel 500 óra hanganyagot tartalmazott (lásd **1. táblázat**).

1. táblázat: Az akusztikai tanító-adatbázisok mérete.

	Kezdeti adatbázis	**+Kiterjesztett adatbázis**			
	Webes híradók	Közéleti hírműsorok	FF	EKR	Speecon	∑
Időtartam [óra]	64	31	100	65	228	488

3.1.2. Akusztikus modellek tanítása

Az akusztikus modellek tanítása Kaldi keretrendszerben [1] történt, de front-endként a VOXerver [2] lényegkiemelő modulját használtuk. Emellett a VOXerver dekóderét is alkalmassá tettük az elkészült akusztikus modellek fogadására.

A 64 órás korpuszon a tanítás a state-of-the-artnak megfelelő módon, MFCC39 jellemzőkön, trifón GMM/HMM modellek elkészítésével indult. A tanított modellek 9453 osztott állapottal, állapotonként átlagosan 10 Gauss-komponenssel rendelkeztek. Az ebből kiindulva készített DNN bemeneti rétege 351 dimenziós (aktuális keret ±4 keret összefűzve), 3 rejtett rétege 400 egységből, egységenként 5 neuronból állt (összesen 2000 neuron rejtett rétegenként), p-norm aktivációs függvényekkel. A p-norm nemlinearitást a maxout nemlinearitás általánosításaként kapjuk [14]:

$$y = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}$$

ahol p szabad paraméter, N pedig a neuronok száma egységenként. Esetünkben p=4, N=5.

Az 500 órás korpuszon tanított modellek jellemzővektorait MFCC13 kiindulási jellemzőkön, a front-endben hét (aktuális ±3 keret) összefűzése után alkalmazott LDA eljárás után 40 dimenziós keretenként kaptuk. A state-of-the-art GMM/HMM 9677 osztott állapotot, állapotonként átlagosan 10 Gauss komponenst tartalmazott. Az ebből kiindulva tanított DNN bementi rétege 9 keret összefűzésével 360 dimenziós, 6 rejtett rétegű, rejtett rétegenként 400x5=2000 neuront tartalmazott, p-norm (p=4) aktivációs függvényekkel.

Az újrabeszélt felvételek kiértékelése során egy a tesztanyaghoz maximum a posteriori (MAP) módszerrel adaptált GMM modellt alkalmaztunk. A lexikai elemek fonetikus átírását a magyar nyelv hasonulási tulajdonságait figyelembe vevő, automatikus eljárással készítettük.

3.2. Nyelvi modellezés

3.2.1. Szöveges tanító-adatbázisok

Kísérleti rendszerünk nyelvi modelljének betanításához négy, különböző forrásból származó szövegkorpuszt használtunk fel (lásd **2. táblázat**):

- *MTVA feliratok*: Az MTVA által rendelkezésünkre bocsátott televíziós felvételekhez tartozó feliratok közül kiválogattuk a közéleti- és hírműsorokhoz tartozókat. Az így nyert 15 millió szót tartalmazó szöveges adatbázis képezi a kezdeti rendszer tanítószövegét, melyet a dekódolási eszközök összevetése során használtunk
- *Webes híradók*: Az azonos nevű akusztikai tanító-adatbázisunk szöveges leirata alkotja ezt a szövegkorpuszt
- *Közéleti hírműsorok*: A webes híradókhoz hasonlóan ez is az azonos nevű beszédadatbázis kézi leiratát tartalmazza
- *Webkorpusz*: A kisebb méretű, de feladatspecifikus tanítószövegek mellett kiegészítő adatbázisként egy webes hírportálokról gyűjtött, 55 millió tokent tartalmazó korpuszt is felhasználtunk a kísérleti rendszerben

2. táblázat: A szöveges tanító-adatbázisok statisztikai adatai.

	Kezdeti adatbázis	**+Kiterjesztett adatbázis**			
	MTVA feliratok	Webes híradók	Közéleti hírműsorok	Webkorpusz	∑
Token [millió szó]	15,1	0,454	0,409	54,8	70,8
Type [ezer szó]	586	67	49	613	931

3.2.2. Nyelvi modellek tanítása

A szövegkorpuszok előkészítése során eltávolítottuk a nem fonetizálható elemeket, meghatároztuk a mondathatárokat, majd statisztikai módszer segítségével átalakítottuk a mondatkezdő szavakat, oly módon, hogy csak a feltételezhető tulajdonnevek őrizzék meg a nagy kezdőbetűs írásmódot. Ezután bizonyos, nem hagyományos lexikai elemeket (pl. számok) átírtunk kiejtett alakjukra, így segítve a kiejtési modell generálását.

A normalizált tanítószövegek alapján minden korpuszon független, 3-gram nyelvi modellt tanítottunk az SRI nyelvi modellező eszköz segítségével [15]. A kísérletek során felhasznált nyelvi modellek ezután úgy készültek, hogy az egyes modelleket lineáris interpoláció segítségével a beszédfelismerési feladathoz adaptáltunk egy paraméterhangolási célokra elkülönített tesztanyagon. Entrópia-alapú modellmetszést nem alkalmaztunk, azonban a webkorpuszból készített nyelvi modell szótárából eltávolítottuk az egyszer előforduló szavakat.

3.3. Tesztelés

A feliratozó rendszer tesztelésére összesen 3,75 óra televíziós közéleti társalgási beszédet különítettünk el, melyből 2,75 órát használtuk a közvetlen feliratozó rendszer tesztelésére és 1 órát az újrabeszélt felvételek kiértékelésére. Ezen felül 10 televíziós híradón is teszteltük a feliratozót, mely összesen további 3 óra tesztanyagot jelentett.

A kísérletek során két súlyozott, véges állapotú átalakítót alkalmazó dekóder teljesítményét vetettük össze. Az első a népszerű Kaldi [1] toolkit **FasterDecoder** nevű eszköze, melyet a SpeechTex Kft. WFST dekóderével a **VOXerver**-rel [2] hasonlítottunk össze. Az eredmények minél pontosabb összevethetősége érdekében minden tesztet ugyanazon a számítógépen (3.5 GHz Core i7), és ugyanazon az operációs rendszeren (Ubuntu 12.04) futtattuk. A különböző implementációk dekódolási sebességét, memóriaigényét és pontosságát is mértük. Az MTVA által rendelkezésükre bocsátott szerveren összesen 24 GB memóriát allokáltak az 5 csatornát feliratozó rendszer üzemeltetésére, így legfeljebb **4.8 GB** állt rendelkezésünkre egy csatorna feliratozásához.

4. Eredmények

A fejezet első felében a Kaldi és a VOXerver dekóder erőforrásigényét hasonlítjuk össze közéleti társalgási beszéd felismerési feladatán. Utána bemutatjuk az összes tanítóanyag felhasználásával készült feliratozó rendszerünk pontosságát és erőforrásigényét immáron nem csak közéleti beszélgetések, hanem híradók esetén is. Végül az újrabeszélők alkalmazásával kapott eredményeket ismertetjük.

4.1. Erőforrásigények összehasonlítása

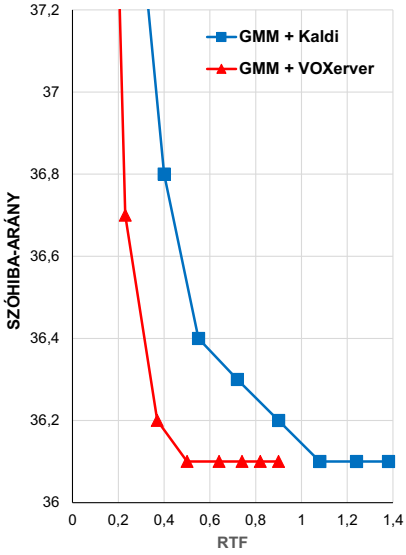
Ennél a vizsgálatnál mind az akusztikus, mind a nyelvi modell tanításához az előző fejezetben bemutatott kezdeti adatbázisokat használtuk, és a közéleti társalgási beszéd közel 3 órás tesztanyagán értékeltük ki őket. Az egyes dekóder és akusztikus modell párosokkal a szaturációs pontban mérhető szóhiba-arányokat és dekódolási sebességeket a **3. táblázat**ban foglaltuk össze. Míg GMM esetén az alkalmazott dekódertől nem függ a hibaarány, VOXerver-rel dekódolva 1%-kal jobb hibaarányt kapunk DNN modell esetén, melynek okát egyelőre vizsgáljuk. A futási sebességek között már azonban lényegesebb különbséget látunk. A Kaldi dekóderével több mint **kétszer annyi időbe telik** a szaturációs pont elérése, mint VOXerver-rel. Talán még ennél is szembetűnőbb a különbség a két dekóder memóriahatékonysága között, ugyanis itt majdnem **hatszoros a különbség,** ismét a VOXerver javára. A Kaldi több mint 7 GB-os memóriaigénye azt jelenti, hogy már egy ilyen szűkített modellel sem tudnánk kiszolgálni az MTVA szerverén az összes csatornát.

A VOXerver kiemelkedő erőforrás-hatékonysága az adattárolási és dekódolási stratégiájában rejlik. A VOXerver-ben az akusztikai állapotok nem részei a WFST-nek, hanem az akusztikus modellek és a CLG (fonetikai környezet, szótár, nyelvtan) szintű WFST kerülnek együtt eltárolásra egy speciális, bináris struktúrában. Ez a struktúra a gyors hozzáférésre, az optimális cache-használatra és a modellek nagyon kompakt reprezentációjára lett kifejlesztve.

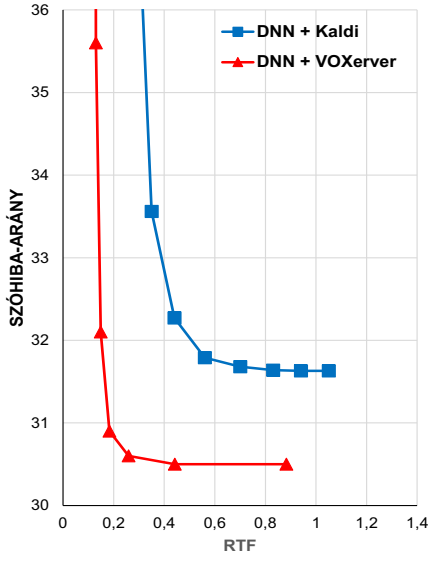
3. táblázat: A közéleti társalgási beszéd közvetlen felismerésével nyerhető legjobb eredmények (**RTF: Real Time Factor**, a dekódoláshoz szükséges idő és a tesztfelvétel hosszának hányadosa, ha RTF ≤ 1 a rendszer képes valós idejű feldolgozásra).

Akusztikus modell	Dekóder	Szóhiba-arány	RTF	Memória használat
GMM	Kaldi	36,1 %	1,23	7,6 GB
	VOXerver	36,1 %	0,5	1,2 GB
DNN	Kaldi	31,6 %	0,9	7,6 GB
	VOXerver	30,5 %	0,44	1,2 GB

A két dekóder futási sebességének alaposabb összevethetőségének érdekében különböző beam szélességgel is futtattunk dekódolást, melynek eredményét az **1.** és **2. ábrán** mutatjuk be. Mint látható mindkét esetben nagyjából fele akkora RTF-nél történik meg a szaturáció a VOXerver-t használva. Ha a lehető legnagyobb pontosság érdekében szaturáció közelében szeretnénk üzemeltetni a feliratozó rendszert, akkor a Kaldi dekóderével a valósidejű működés határán (RTF ~ 1) mozgunk, ezzel szemben a VOXerver képes valósidejű feldolgozásra (RTF ~ 0,4-0,5) az akusztikus modell típusától függetlenül.



1. ábra: Szóhiba-arány a futásidő függvényében közéleti társalgási beszéden mérve, GMM-alapú akusztikus modellel.



2. ábra: Szóhiba-arány a futásidő függvényében közéleti társalgási beszéden mérve, DNN-alapú akusztikus modellel.

4.2. Teljes rendszer

A teljes méretű nyelvi modelleket csak VOXerver környezetben tudtunk futtatni (lásd **4. táblázat**), ugyanis nem állt rendelkezésünkre olyan szerver, mellyel ki tudtuk volna elégíteni a Kaldi keretrendszer WFST hálózat építése közben keletkező memóriaigényét. Becsléseink szerint Kaldi hálózatot használva 25 GB memóriára lett volna szükségünk egyetlen felismerési szál futtatására, és a modell megépítéséhez szükséges memória ennek akár háromszorosa is lehetett volna.

A 4. táblázat adatai alapján elmondhatjuk, hogy a nyelvi modell bővítésével 6%-os, további akusztikus tanítóanyagok bevonásával 8%-os, összegségében pedig mintegy **13%-os relatív hibaarány csökkenést** értünk el a közéleti társalgási beszéd feladatán. Annak érdekében, hogy láthassuk a különbséget a két feladat nehézsége között, egy 3 órás híradókat tartalmazó adatbázissal is teszteltük a modelleket. Látható, hogy még a kezdeti modellekkel is jelentősen alacsonyabb hibaarány érhető el híradókon (~12%), a kiterjesztett modellekkel pedig ez tovább csökkenthető. Az így elért kicsivel **10% alatti szóhiba-arány** annyira alacsonynak mondható, hogy akár újrabeszélő nélküli, közvetlen feliratozását is lehetővé teszi a híradóknak.

Felmerül a kérdés, hogy a nagyobb nyelvi modell és a több rejtett réteget használó DNN akusztikus modell milyen hatással van a feliratozó rendszer erőforrásigényére. A kiterjesztett nyelvi modell hatására a VOXerver memóriaigénye **4 GB-ra növekedett**, mely azonban még így is alatta marad a kitűzött 4,8 GB-os határnak. A dekódolási sebesség tekintetében méréseink alapján nincs változás a kezdeti rendszerhez képest.

4. táblázat: Közéleti társalgási beszéd és híradók feliratozásának hibaaránya a kiterjesztett adatbázisokkal tanított modellek alapján.

Tesztadatbázis	Tanítószöveg	Szóhiba-arány	
		Kezdeti DNN (64 óra)	Kiterjesztett DNN (500 óra)
Közéleti társalgási beszéd	Kezdeti	30,5 %	28,0 %
	+Kiterjesztett	28,7 %	26,4 %
Híradók	Kezdeti	12,4 %	11,1 %
	+Kiterjesztett	10,6 %	9,9 %

4.3. Újrabeszélési kísérletek

Bár a 26%-os hibaarány még nemzetközi összehasonlításban is alacsonynak mondható, ez még nem jelenti azt, hogy a mostani megoldás közvetlenül alkalmas lenne közéleti társalgási beszéd feliratozására. A hibaarány további csökkentése céljából úgy döntöttünk, hogy kipróbáljuk a más országokban már nagy népszerűségnek örvendő **újrabeszélés**t (re-speaking) [11]. Ehhez 1 órányi közéleti társalgási beszéd valós körülményeket szimuláló újrabeszélését rögzítette számunkra az MTVA. Mivel gyakorlott, szakképzett újrabeszélő jelenleg nem érhető el Magyarországon, az MTVA-val közösen úgy döntöttünk, hogy első kísérleteinket tömörítés nélkül, szószerinti újramondással végezzük. Így lehetőségünk nyílt szóhiba-arány számítására, azonban egy gyakorlatban is

működő rendszer esetén a jövőben meg kell fontolni az összefoglaló jellegű újrabeszélés alkalmazását, ugyanis az így létrejött nagy mennyiségű szöveg mind a befogadó, mind az előállítói oldalon problémát jelenthet.

Az újrabeszélt műsorokat két módon teszteltük. Először elkészítettük a feliratot közvetlenül a hangsávból (**közv.**), majd utána az újrabeszélt változatból is (**újrab.**). Akusztikus modellként a közvetlen feliratozás esetén az 4.1 pontban használt GMM és DNN modellt alkalmaztuk, újrabeszélt változaton pedig a DNN-t, illetve a GMM modell egy MAP módszerrel beszélőadaptált változatát. Az eredményeket az **5. táblázat**ban foglaltuk össze. Látható, hogy összesen négy felismerési feladatot és három újrabeszélőt vizsgáltunk, és mindkét tényezőtől erősen függött az újrabeszéléssel kapható javulás. DNN modellek esetén 2-30% között mozog a relatív hibaarány csökkenés, mely tisztán az újrabeszélésnek tulajdonítható. A legjobb eredményt az adaptált GMM modellel kaptuk, mely a közvetlenül feliratozott DNN-es eredményhez képest átlagosan 9%-kal jobb. Elmondható tehát, a mostani egyszerű kísérletek is igazolták az újrabeszélés hatékonyságát. Meggyőződésünk, hogy az újrabeszélők képzésével a jelenleginél sokkal jobb eredmények is elérhetők lesznek a jövőben.

5. táblázat: Közéleti társalgási beszéd közvetlen és újrabeszélés utáni gépi feliratának szóhiba-aránya. A műsor és újrabeszélő azonosítókat az első oszlopban jelöltük arab illetve római számmal.

Műsor / Újrab.	**GMM (közv.)**	**DNN (közv.)**	**DNN (újrab.)**	**GMM + MAP (újrab.)**
1 / I	31,5 %	23,4 %	20,5 %	19,2 %
2 / I	30,7 %	24,8 %	23,4 %	18,8 %
3 / II	33,6 %	25,4 %	18,1 %	17,6 %
4 / III	34,1 %	26,8 %	26,2 %	25,1 %
Átlag	**32,5** %	**25,1** %	**22,1** %	**20,2** %

5. Összefoglalás

Cikkünkben bemutattunk egy elsősorban televíziós közéleti társalgási beszéd feliratozására optimalizált gépi beszéd-szöveg átalakító rendszert, melyet az MTVA-val együttműködésben fejlesztünk. Többféle adatbázis alapján, többféle technikával tanítottunk beszédfelismerő modelleket, melyeket kétféle WFST dekóder segítségével értékeltünk ki. Méréseink azt mutatták, hogy az akusztikai modellezéstől függetlenül a VOXerver **kétszer gyorsabb és hatszor kevesebb memóriát fogyaszt**, mint a Kaldi keretrendszer FasterDecoder nevű eszköze. A több mint 70 millió szón és 500 óra beszéden, mély neurális hálózatok felhasználásával tanított rendszerünk kb. 26%-os szóhiba-aránnyal ismerte fel a közéleti társalgási beszédet és kevesebb mint 10%-os hibával a híradókat. Legjobb tudomásunk szerint ezek a **legalacsonyabb publikált értékek** mindkét magyar nyelvű beszédfelismerési feladaton.

A közéleti társalgási beszéd feliratának javítása céljából újrabeszéléssel is végeztünk kísérleteket. Bár ezzel a technikával sikerült 20% körülire csökkenteni a feliratozás hibaarányát, ez még mindig túl magas a gyakorlati alkalmazhatósághoz. Véleményünk szerint azonban képzettebb, gyakorlottabb újrabeszélőkkel és az összefoglaló jellegű újramondás megengedésével a jövőben jó minőségű feliratok hozhatóak majd létre.

Köszönetnyilvánítás

Ezúton is szeretnénk megköszönni a Médiaszolgáltatás-támogató és Vagyonkezelő Alapnak minden segítséget, mellyel munkánkat támogatta. Kutatásunk részben a Patimedia (PIAC_13-1-2013-0234) projekt támogatásával készült.

Bibliográfia

1. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Others: The Kaldi speech recognition toolkit. In: Proc. ASRU (2011).
2. Tarján, B., Mihajlik, P., Balog, A., Fegyó, T.: Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection. In: 2nd International Conference on Cognitive Infocommunications (CogInfoCom). pp. 1–5. , Budapest, Hungary (2011).
3. Tarján, B., Mihajlik, P.: On morph-based LVCSR improvements. In: Spoken Language Technologies for Under-Resourced Languages (SLTU-2010). pp. 10–16. , Penang, Malaysia (2010).
4. Tarján, B., Fegyó, T., Mihajlik, P.: A Bilingual Study on the Prediction of Morph-based Improvement. In: SLTU 2014: 4th International Workshop on Spoken Languages Technologies for Under-Resourced Languages. pp. 131–138. , Saint Petersburg (2014).
5. Roy, A., Lamel, L., Fraga, T., Gauvain, J., Oparin, I.: Some Issues affecting the Transcription of Hungarian Broadcast Audio. In: 14th Annual Conference of the International Speech Communication Association (Interspeech 2013). pp. 3102–3106 (2013).
6. Tamás, G., György, K., László, T.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY. pp. 3–13 (2014).
7. Sundermeyer, M., Nussbaum-Thom, M., Wiesler, S., Plahl, C., Mousa, A.E.-D., Hahn, S., Nolden, D., Schluter, R., Ney, H.: The RWTH 2010 Quaero ASR evaluation system for English, French, and German. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 2212–2215 (2011).
8. Winebarger, J., Nguyen, B., Gehring, J., Stüker, S., Waibel, A.: The 2013 KIT Quaero Speech-to-Text System for French. In: Proceedings of the 10th International Workshop for Spoken Language Translation (IWSLT 2013). , Heidelberg (2013).
9. Kobayashi, A., Oku, T., Imai, T., Nakagawa, S.: Risk-Based Semi-Supervised Discriminative Language Modeling for Broadcast Transcription. {IEICE} Trans. 95-D, 2674–2681 (2012).
10. Ofcom: Measuring live subtitling quality: Results from the first sampling excercise. (2014).
11. Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., Remael, A.: Live Subtitling with Speech Recognition Causes and Consequences of Text Reduction. (2010).
12. Mihajlik, P., Balog, A.: Lightly supervised acoustic model training for imprecisely and asynchronously transcribed speech. In: Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on. pp. 1–5 (2013).

13. Siemund, R., Höge, H., Kunzmann, S., Marasek, K.: SPEECON-speech data for consumer devices. Second Int. Conf. Lang. Resour. Eval. 883–886 (2000).
14. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 215–219. IEEE (2014).
15. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proceedings International Conference on Spoken Language Processing. pp. 901–904. , Denver, US (2002).

Egy magyar nyelvű beszédfelismerő rendszer szószintű hibáinak elemzése

Gosztolya Gábor[1,2], Vincze Veronika[1], Grósz Tamás[2], Tóth László[1]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: {tothl, ggabor, vinczev}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail: groszt@inf.u-szeged.hu

Kivonat Az automatikus beszédfelismerő rendszerek szószintű hibáját hagyományosan egy illesztési távolságon alapuló metrikával mérjük, amely a szóalakok pontos egyezésének vizsgálatán alapszik. Mint a legtöbb beszédfelismerési technika, ez is jól illeszkedik az angol nyelvre, más (pl. ragozó) nyelvekre azonban ez nem feltétlenül igaz. Ebben a cikkben azt vizsgáljuk, hogy egy hagyományosnak számító beszédfelismerő rendszer (mély neuronhálós akusztikus modell és szó-trigram nyelvi modell) milyen jellegű hibákat vét. Ehhez száz hangfelvétel hibáit gyűjtöttük ki, annotáltuk manuálisan, majd elemeztük. Végkövetkeztetésünk, hogy a szótárban nem szereplő elemek mellett nagy gondot okoz a magyar nyelvben az egybe- és különírások kezelése, melyet a hagyományos pontosságmetrika különösen nagy mértékben büntet. [1]

Kulcsszavak: beszédfelismerés, illesztési távolság, N-gram modell

1. Bevezetés

Az automatikus beszédfelismerő rendszerek fejlesztésében hamar dominánssá vált, és mindmáig az is maradt az angol nyelvű beszéd felismerése. Emiatt a beszédfelismeréssel foglalkozó kutatók általában olyan technikákat dolgoztak ki, melyeknél a fő kritérium az volt, hogy angol nyelvre jól működjenek; természetes, hogy a más nyelvekre beszédfelismerő rendszert fejlesztők és a téma kutatói ugyanezeket vagy nagyon hasonló technikákat alkalmaznak saját nyelvükre. Ez azonban sokszor nem a legszerencsésebb, hiszen egy-egy nyelv speciális tulajdonságai indokolhatják, hogy az adott problémát árnyaltabban közelítsük meg. Erre egy jó példa az ún. N-gram nyelvi modell, melyben egy-egy szóalak előfordulási valószínűségét statisztikai módszerekkel, a megelőző $N-1$ szóalak alapján becsüljük; ez a megközelítés értelemszerűen elég jól illeszkedik az angol nyelvhez, egy ragozó nyelvnél (mint pl. a magyar) viszont indokolt (lenne) az eljárás finomítása (pl. nyelvtani kategóriákra [1] vagy morfokra [2] számolt N-gram).

[1] A jelen kutatás során használt TITAN X grafikus kártyát az NVIDIA Corporation ajándékozta csoportunknak.

Egy másik elterjedt beszédfelismerési technika, mely szintén a szóalakok különbözőségének elvén alapszik, maga a beszédfelismerő rendszer pontosságát mérő metrika: általában az illesztési (vagy Levenshtein-) távolságon [3] alapuló pontosságértéket szokás alkalmazni. Ebben a beszédfelismerő rendszer kimenetét a helyes szöveges átirathoz hasonlítva megszámoljuk a beszúrt, törölt illetve kicserélt szóalakok számát, és ezen értékek (valamint a helyes átiratban szereplő szóalak-előfordulások száma) alapján számítjuk ki a rendszer százalékban kifejezett pontosságát. Nyilvánvaló, hogy egy agglutinatív nyelv esetében ez a megközelítés is vezethet problémákhoz, ugyanakkor tudtunkkal itt fel sem merült más metrika használata. (Bár az eljárás finomításaként értékelhetjük, hogy a távolkeleti nyelvek (pl. japán vagy kínai) esetén ugyanilyen módon, de nem szóalak-, hanem karakteralapon számítják ki a beszédfelismerő rendszer pontosságát.)

Ha felmérjük, hogy beszédfelismerő rendszerünk milyen jellegű hibából vét (relatíve) sokat, akkor az adott jelenséget célirányosan kezelve jelentősen javíthatjuk a felismerés pontosságát. Emiatt érdekes lehet, hogy egy beszédfelismerő rendszer szószintű kimenetében milyen jellegű tévesztések fordulnak elő gyakrabban; ugyanakkor a jellemző hibatípusokat nyugodtan nevezhetjük közismertnek. Köztudott, hogy a hibák egy jelentős része visszavezethető a felismerési szótárból hiányzó (Out-of-vocabulary, OOV) szóalakokra, azon belül különösen két nagyobb csoportra: a tulajdonnevekére és a számnevekére (ez utóbbiba beleértünk egyéb, számokhoz kapcsolódó szóalakokat is, pl. *harmincezren*). Tulajdonnevekből és egyéb névelemekből nagyon sok alak létezik, ráadásul ezek gyakorisága a beszéd témájától, sőt a beszéd elhangzásának idejétől függően változik [4]. A számnevek esetében szintén a szóalakok nagy (gyakorlatilag végtelen) száma okoz gondot. A két típusban közös, hogy egyrészt nem lehet az összes lehetséges szóalakot felvenni a szótárba, másrészt az OOV szó helyére beerőltetett, hibás szóalak a nyelvi modell által a következő szavakra adott becslést is lerontja.

Az OOV szóalakok további köztudott tulajdonsága, hogy hajlamosak „elrontani" az előfordulásuk közvetlen környezetét is. Ha egy beszédfelismerő rendszer egy általa ismeretlen (és így felismerhetetlen) szóalakkal találkozik, jellemző, hogy az érintett részre valamely akusztikailag nagyon hasonló, ám a szótárban szereplő szóalakot illeszt. Amennyiben az akusztikailag hasonló szó rövidebb vagy hosszabb, akkor a szótévesztések a szomszédos szavakra is kiterjednek (pl. *biztosít többséget* vs. *biztosítók siket*). A másik mellékhatás, mikor a rendszer az ismeretlen szót több, a szótárban szereplő szóból próbálja meg kirakni (pl. *huszonkilencedikére* vs. *ózon kilencedikére*), ekkor ugyanis két vagy több szótévesztést (egy szócserét és egy vagy több szóbeszúrást) tapasztalunk, amelyek nagyobb súllyal esnek latba a rendszer szószintű pontosságának számításakor.

Jelen cikkünkben azt vizsgáljuk, hogy egy, a fönti szempontok alapján hagyományosnak számító megközelítés hogyan viselkedik magyar nyelvre. Ehhez egy, ma a legmodernebbnek számító akusztikus, és egy hagyományos szóalak trigram (3-gram) nyelvi modellel rendelkező beszédfelismerő rendszerrel végzünk felismerést magyar nyelvű híradófelvételeken [5], majd a szószintű hibákat manuálisan kategorizáljuk és elemezzük. Végül kísérletet teszünk arra is, hogy felmérjük a felismerő kimenetének olvashatóságát.

2. A tesztkörnyezet

A következőkben bemutatjuk a tesztkörülményeket: a beszédfelismerő rendszer akusztikus modelljét, a hangfelvételek szöveges átiratainak készítését, valamint az alkalmazott nyelvi modellt.

2.1. Akusztikus modell

Akusztikus modellként egy mély egyenirányított neurális hálót alkalmaztunk [5]. Mély neuronhálónk tanításához egy teljesen GMM-mentes módszert használtunk [6], melynek lényege, hogy az átirat időbeli illesztését és a környezetfüggő állapotok klaszterezését is csak mély neuronhálókra támaszkodó módszerek segítségével végezzük.

Mivel kezdetben nem állt rendelkezésünkre a szöveges átirat időbeli illesztése, első lépésként egy maximális kölcsönös információn (Maximum Mutual Information, MMI [7]) alapuló módszerrel tanítottunk egy környezetfüggetlen neuronhálót. A betanított háló kimenetei alapján elvégeztük az annotáció kényszerített illesztését, majd az illesztések további finomításához tanítottunk egy újabb hálót immár a hagyományos keresztentrópia-alapú kritériumra optimalizálva. Az új háló alapján újraillesztettük a címkéket, és a továbbiakban az így kapott keretszintű címkézést használtuk.

A kényszerített illesztés elvégzése után a tavaly bemutatott Kullback-Leibler–divergencián alapuló klaszterezést alkalmazva állítottuk elő az összevont környezetfüggő állapotokat [8]. A végső akusztikus modellként használt mély neuronhálót az így kapott környezetfüggő osztályok felismerésére tanítottuk a hagyományos keresztentrópiát optimalizáló hiba-visszaterjesztéses algoritmussal. A konkrét modellünkben 1843 környezetfüggő állapotot használtunk; a neuronháló öt rejtett rétegből állt, rétegenként ezer-ezer ún. rectifier neuronnal.

2.2. Nyelvi átírás és -modellezés

A beszédadatbázis ortografikus átiratának elkészítésekor a következő alapszabályokat alkalmaztuk. A szöveges átiratok csak nagybetűket tartalmaznak, hogy a tulajdonnevek nagy kezdőbetűjéből eredő szótévesztéseket kizárjuk. A számneveket minden esetben ortografikusan írtuk át. A kötőjelet tartalmazó összetett szavak esetén a kötőjel helyett szóközt írtunk. A rendhagyó szóalakokat (pl. idegen szavak) kiejtés szerint írtuk le, hogy a fonetikus átíró feladatát megkönnyítsük. A két utóbbi lépés a beszédkorpuszban szereplő átiratok és a nyelvi modell szókészlete között nyilván eltéréseket okoz, ennek hatását a továbbiakban elemezni fogjuk.

A szótár és a nyelvi modell kialakításához két forrást használtunk fel. A nyelvi modell alapját az origo hírportál anyagából készített szövegkorpusz képezte, ami kb. 50 millió szövegszóból állt. Mivel a korpusz sok hibás szóalakot tartalmazott, szükségünk volt egy megoldásra a hibák kiszűrésére. Ezért a rendszer által elfogadott szóalakok listáját leszűkítettük a Magyar Kiejtési Szótár szólistájára [9]. Ez kb. másfél millió, korpuszokból kigyűjtött szóalakot tartalmaz,

ezért azt feltételeztük, hogy a legtöbb fontos szóalak fellelhető lesz benne. A origo korpuszon megvizsgáltuk ezen másfél milló szóalak gyakoriságát, és csak a legalább kétszer előforduló alakokat tartottuk meg. Ez a lépés felismerőnk szótárának méretét 486 982 szóra redukálta. Az origo korpusz alapján trigram nyelvi modellt készítettünk a HTK nyelvi szubrutinjait az alapértelmezett értékekkel használva [10], míg a szótár szavainak kiejtését a Magyar Kiejtési Szótárból [9] vettük.

Úgy véljük, hogy a fönti (etalon) szöveges átirat és a nyelvi modell tekinthetőek standard és ésszerű módon elkészítettnek, így a cikkünkben ezután megjelenő tapasztalatok és következtetések tekinthetőek általános érvényűnek (legalábbis magyar nyelvre).

2.3. Egyéb kísérleti körülmények

Kísérleteinket a „Szeged" magyar nyelvű híradós beszédadatbázison [5] végeztük. Az adatbázis összesen 28 órányi hangzóanyagot tartalmaz, melyet a szokásos felosztásban használtunk: 22 órányi anyag volt a betanítási rész, 2 órányi a validációs halmaz, a maradék 4 órányi hanganyag pedig a tesztelésre szolgáló blokk.

Az elemzést az adatbázis teszthalmazán, annak is egy 100 hangfelvételből álló részhalmazán végeztük. A teljes teszthalmazon a felismerési pontosság 84, 14% volt, míg a vizsgált részen 85, 69%; utóbbiban összesen 4214 szóelőfordulást számoltunk.

3. A hibák elemzése

Az előforduló hibatípusok elemzéséhez manuálisan néztük át a teszthalmaz egy részének annotációját és az adott részre a beszédfelismerő rendszer kimenetét. Ehhez automatikusan kigyűjtöttük a tévesztett részeket, majd azokat és a felismerő illesztett eredményét egy-egy szomszédos szóval kiegészítve megjelenítettük. Ezután az egyes tévesztéseket manuálisan kategóriákba soroltuk.

Az egyes hibákat először nyelvészeti szempontok alapján kategorizáltuk. Ilyen volt például az egybeírás/különírás: ez esetben a beszédfelismerő rendszer által készített átirat és az etalon szöveg mindössze egy (vagy több) szóköznyi eltérést mutatott (pl. *a két százmilliárdos tétel* vs. *a kétszáz milliárdos tétel*, *az exportdinamikája is* vs. *az export dinamikája is*). Visszatérő hiba volt az is, ha egymás után következett két azonos hang, melyet egy (hosszú) hangnak tekintett a rendszer. Ebben a kategóriában különösen gyakran egy *a*-ra végződő szót követő *a* névelő okozta a hibát (*mondja bankszövetség* vs. *mondja a bankszövetség*). Sok esetben magát a szót/szótövet jól felismerte a rendszer, azonban a hozzá kapcsolódó toldalékok esetében hibázott: lemaradt a toldalék (*Mezőtúr polgármester* vs. *Mezőtúr polgármestere*), esetleg hibás toldalék került a szó végére (*szétdarabolják* vs. *szétdarabolták*). Bizonyos esetekben az átiratból kimaradt egy szó (*terén erősítik* vs. *terén ha erősítik*). Két olyan hibatípussal is találkozhattunk, amikor a beszédfelismerő kimenete helyes volt, mégis eltért az

etalontól. Az egyik hibatípusnál maga az etalon átirat tartalmazott hibát (*a szennyezetett víztől* vs. *a szennyezett víztől*), míg a másik hibatípus esetében az etalon átirat készítési elveinek megfelelően a rendhagyó kiejtésű tulajdonnevek fonemikus átiratban szerepeltek a szövegben, ugyanakkor a beszédfelismerő az eredeti helyesírás szerint tüntette fel ezeket (*Magyar Helszinki Bizottság* vs. *Magyar Helsinki Bizottság*). A szótárban nem szereplő szavak esetében megfigyelhetjük, hogy azokat a rendszer gyakran fonetikailag hasonló tulajdonságokkal bíró hangokból álló szóval helyettesíti, például a *be* és *de* szavak összecserélése során ugyanúgy zöngés zárhangot találunk a szó elején.

A hibatípusok megoszlásán kívül azt is vizsgáltuk, hogy az egyes hibatípusok jellemzően milyen jellegű szavak környezetében fordulnak elő. Hogy ezt megtehessük, négy tényezőt vizsgáltunk. Amennyiben az adott tévesztéshez tartozó etalon-átiratban bármelyik szóra igaz volt a vizsgált feltétel (pl. a három érintett szóból az egyik hiányzott a szótárból), az érintett hibaelőfordulásra bejelöltük az adott tulajdonságot.

Először azt vizsgáltuk, hogy szerepel-e az etalonban névelem (pl. *Balogh*, *Fidesz*, *tálibok*). Másodszor azt ellenőriztük, hogy szerepel-e benne számnév vagy számmal kapcsolatos szóalak (pl. *ezeréves*, *kétmilliós*, *ezerkilencszázötvenhatos*). Ezután azt is megnéztük, hogy van-e az adott annotációban olyan szó, amely nem szerepel a beszédfelismerő rendszer szótárában (OOV). Végül azt vizsgáltuk, hogy az etalonnak tekintett annotáció helyes-e, vagy esetleg hibát tartalmaz. Ez jellemzően egybe-különírási hiba volt; természetesen ez nem feltétlenül jelenti azt, hogy az etalon valóban hibás, hanem tükrözheti azt is, hogy az annotáció más elvek szerint készült, mint ahogyan a szótár és a nyelvi modell felépült.

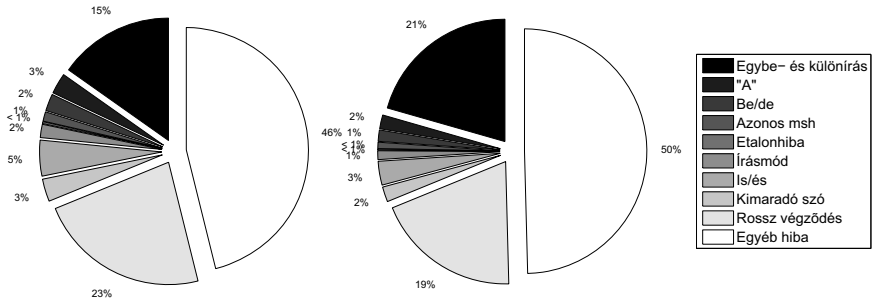
Mivel a beszédfelismerő rendszer szószintű hibáját a korábban ismertetett, illesztési távolságon alapuló módszerrel szokás mérni, logikus a hibák darabszáma mellett a hozzájuk tartozó *szótévesztések* számát is vizsgálni, ezért ezeket is feljegyeztük.

4. Eredmények

Az 1. ábrán látható az egyes hibatípusok megoszlása a hibák darabszámának és a szótévesztések számának arányában.

Az egyes hibatípusok, illetve azokon belül az egyes annotált szótípusok megoszlása az 1. és 2. táblázatokban található. Látható, hogy a tévesztések kicsit több mint 50%-át lehetett besorolni valamilyen informatív hibakategóriába, így kb. 46%-uk az „Egyéb hibák” közé került. A szótévesztéseknek ez valamivel nagyobb részét, szinte pontosan a felét tette ki, ami arra vezethető vissza, hogy bizonyos hibakategóriák (pl. be/de vagy is/és tévesztés, egymás után két „a”, kimaradó szó, írásmódeltérés) esetében jellemzően egy hibára egyetlen szótévesztés jut, míg ez az érték átlagosan 1,5. Az egybe- és különírás hibatípus azonban a felismerési hiba nagyobb részéért felelős, hiszen ilyenkor legalább két szótévesztés jut minden felismerési hibára.

A névelemeket érintő hibák között természetszerűleg voltak írásmódeltérési hibák, illetve gyakori volt a rossz végződés is. Ez nem meglepő, hiszen az egyes



1. ábra. A hibák megoszlása az egyes hibakategóriák között a hibák darabszámának (balra) és a szótévesztések számának (jobbra) arányában

1. táblázat. Az egyes hibatípusok előfordulásának száma, illetve ezen belül az egyes annotált szótípusok előfordulásának száma

Hibatípus	Névelem	Számnév	OOV	Annot.	Össz.
Egybe- és különírás	3	25	25	14	61
Egymás utáni két „a”	0	0	0	0	11
Be/de tévesztés	0	0	9	0	9
Egymás után két azonos msh	0	0	0	0	5
Etalonhiba	0	0	1	1	1
Írásmódeltérés	7	0	6	0	7
Is/és tévesztés	0	0	0	0	19
Kimaradó szó	0	0	0	0	12
Rossz végződés	7	3	20	0	91
Egyéb hiba	96	6	114	0	185
Hibák összesen	113	34	175	15	401

névelemek eleve elég ritkán fordulnak elő a tanítószövegben, így a ragozott alakjaik sem túl gyakoriak. Mégis, a névelemek nagy részét érintő hibák az Egyéb kategóriába estek.

A számneveket érintő hibák nagy többsége egybe- és különírási tévesztés volt. Kézenfekvő lenne ezt betudni annak, hogy nagyon sok számnévi szóalak képezhető, melyeket képtelenség felsorolni egy szótárban, ugyanakkor a 25 esetből csak 5 olyan volt, ahol egyúttal OOV szó is szerepelt az átiratban. A gondot a számneveknél valószínűleg az okozta, hogy a nyelvi modell *mindkét* írásmódot képes előállítani (pl. a *kétszázharmincezer* szó esetén mind a *kétszázharminc*, mind az *ezer* szó szerepelhet (és szerepelt is) a szótárban); illetve tizenegy esetben a számneveket érintő egybe- és különírási hiba annotációs hibával is egybeesett.

Az OOV szónál történt tévesztéseknek együtt kb. negyedét tették ki az egybe- és különírási, valamint a szuffixhibák, a nagy többségüket az egyéb hibák közé soroltuk. Ennek valószínűleg az a magyarázata, hogy ehhez a két kategóriához az szükséges, hogy legalább a szó egy eltérő ragozású alakja szerepeljen a szótárban;

2. táblázat. Az egyes hibatípusokhoz tartozó szótévesztések száma, illetve az egyes hibatípusokon belül az egyes annotált szótípusokhoz tartozó szótévesztések száma

Hibatípus	Névelem	Számnév	OOV	Annot.	Össz.
Egybe- és különírás	6	52	52	28	124
Egymás utáni két „a”	0	0	0	0	11
Be/de tévesztés	0	0	9	0	9
Egymás után két azonos msh	0	0	0	0	5
Etalonhiba	0	0	1	1	1
Írásmódeltérés	7	0	6	0	7
Is/és tévesztés	0	0	0	0	19
Kimaradó szó	0	0	0	0	12
Rossz végződés	11	5	32	0	116
Egyéb hiba	157	11	188	0	299
Hibák összesen	181	68	288	29	603

3. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák száma

Szótípus	Névelem	Számnév	OOV	Annot.	Össz.
Névelem	113	0	99	0	113
Számnév	0	34	10	11	34
OOV	99	10	175	1	175
Annot.	0	11	1	15	15
Összesen	113	34	175	15	216

amennyiben ez sem áll fenn, a beszédfelismerő rendszer valamilyen egyéb, hasonló hangzású szót fog beerőltetni az adott helyre (és ezzel esetleg a környezetet is elrontja). Az olyan tévesztési helyek, ahol az annotáció nem volt helyes (vagy konzisztens), általában egybe- és különírási hibához vezettek; egy esetben pedig az annotáció egyszerűen el lett gépelve (*szennyezetett*).

Az egyes tévesztési típusok felől közelítve látható, hogy az egybe- és különírási tévesztések nagyon nagy része történik olyan helyeken, ahol valamelyik jelölt szótípus előfordul az annotációban; ezek teszik ki az ilyen típusú hibák kb. 80%-át. A be/de tévesztések mindegyike egyúttal OOV hiba is, aminek az a triviális oka van, hogy a „be” szó valahogyan kimaradt a szótárból. Nem meglepő, hogy az írásmódeltérések kizárólag névelemeket érintenek, az már annál inkább, hogy egy esetben nincs szó OOV-ről. Ennek az az oka, hogy mind az *Attilának*, mind az *Atillának* szóalak szerepelt a szótárban.

Az egyéb, máshova besorolhatatlan hibák több mint felében névelem is előfordult, kétharmadukban pedig a szótárban nem szereplő szó is. Az összes előforduló hibát tekintve is magas (bár ennél alacsonyabb) arányokat láthatunk; összességében csak a tévesztések kb. 54%-ánál nem volt jelen egyik jelölt szókategória sem, igaz, ezek adták a felismerési hiba kb. 60%-át.

A 3. és 4. táblázat mutatja, hogy az egyes jelölt szótípusok mennyire estek egybe. (Értelemszerűen a táblázat főátlója megegyezik az összesítő sorral és -

4. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák szótévesztéseinek összege

Szótípus	Névelem	Számnév	OOV	Annot.	Össz.
Névelem	181	0	161	0	181
Számnév	0	68	22	22	68
OOV	161	22	288	1	288
Annot.	0	22	1	29	29
Összesen	181	68	288	29	360

oszloppal.) Látható, hogy a névelemmel egybeeső tévesztések nagyon nagy része (87%-a) egyúttal OOV is; fordítva ez értelemszerűen jóval kisebb (53%), hiszen sok más szóalak-típusra is jellemző lehet, hogy hiányzik a szótárból (pl. ragozott alakok). A számnevek kb. egy-egyharmada OOV és annotálási hiba. Föltűnő még, hogy az annotálási hibák milyen nagy része számnév; ez valószínűleg a számnevek helyesírásának bonyolultságára vezethető vissza (hiszen a szavakat a kötőjelek mentén feldaraboltuk, így a kötőjelezési hibák is egybe- és különírási hibaként jelennek meg).

Az egyes hibakategóriákra néhány példát láthatunk az 5. táblázatban.

Összességében, tapasztalataink szerint a hibák egy jelentős része arra vezethető vissza, hogy az átiratot és a szótárat (részben) eltérő módon állítottuk össze. A tulajdonnevek fonetikus átírása segített a bemondások fonetikai címkéinek meghatározásakor (és így az akusztikai modell tanításakor), a felismerő szótárába azonban ezek a szavak más alakban kerültek be, így, még ha meg is találta a kérdéses szavakat a beszédfelismerő rendszer, a kimenet az eltérő írásmód miatt hibásnak számított. Valószínűleg a kiejtési szótár és az N-gram modell felépítésére használt korpusz időnként eltérő írásmódja is felelős azért, hogy egy sor rövidítés és tulajdonnév végül kimaradt a nyelvi modellből; ilyenek voltak (az egy szál Fidesz kivételével) a pártok nevei, melyek pedig a híradófelvételeinkben erőteljesen felülreprezentáltak. Emellett valamilyen rejtélyes okból néhány igen gyakori szó (pl. *be*, *legalább*) is hiányzott a szótárból (vagyis a Magyar Kiejtési Szótárból).

A fentieken felül a felismerési hibák meglepően nagy része vezethető vissza egybe- és különírási hibákra, főleg számnevek esetében. Ekkor a felismerő kimenete „gyakorlatilag” helyes, jól olvasható és értelmezhető, „csak” helyesírási hibát tartalmaz. Ez a jelenség nem (vagy csak elhanyagolható mértékben) jelentkezik az angol nyelv esetén; magyar nyelvre végzett felismerésnél azonban ez fokozottan jelen van. Természetesen a rosszul tagolt szavak is hibának számítanak, azonban ezeket logikus lenne kisebb súllyal figyelembe venni, mint ha egy teljesen más jelentésű szót ismertünk volna fel az adott helyen. Véleményünk szerint ez a beszédfelismerés területén gyakorlatilag egyeduralkodó pontosságmetrika (magyar) nyelvspecifikus hiányossága.

5. táblázat. Példák az egyes hibatípusokra

Hibakategória	Etalon szöveg	Felismert szöveg
Egybe- és különírás	kettőszáz milliárdot	kettő százmilliárdot
	százhatvannégyezer	százhatvannégy ezer
	bankszektortól	bank szektortól
	feladatszabó állománygyűlésére	feladat szabó állomány gyűlésére
Kimaradó „a”	leszakította a vihar	leszakította vihar
Írásmódeltérés	smitt pál	schmidt pál
	balog andrást	balogh andrást
Egymás után két azonos msh.	ülést tart	ülés tart
	alkotmánybírók kiválasztásáról	alkotmánybíró kiválasztásáról
Rossz végződés	tihamért	tihamér
	miniszterelnököt	miniszterelnökhöz
	kivégzésére	kivégzését
Egyéb	védőhálóként	védőháló kint
	blogjában	blokkjában
	nem tervez tömeges	nem tervezte meg és
	huszonkilenc pontot	huszonkilenc bontott
	miniszterelnököket	miniszterelnök őket
	húsfeldolgozóba jártak	húsfeldolgozó bejártak
	képest	képes
	tévéjeiket	tévéje éket
Egyéb (szn.)	nulla egész kilenc tized	nyúl egész keresztttüzet
	huszonkilencedikére	ózon kilencedikére
Egyéb (ne.)	alkaida tag	ajkaid adtak
	az atévé híradóban	az a tévéhíradóban
	szuzuki szvift modell	hozó kiszűrt modell
	rogán antal	jogán antal
	kósa lajos	koós alajos
	emeszpés	ám ezt és
	be az emeszpé	bazár messzi
	oszama bin láden	asszam a világon
	biszku béla	whisky béla
	biszku béla	büszke béla
	vargasovszki	varga sóz ki
	európai unió	euró pari unió
	robert ficó	róbert fikció
	fidesz kádéenpé	fidesz káld ilyen ki

5. A kimenet olvashatóságának vizsgálata

Habár a pontos szóalak-írásmód is elvárható egy jól működő beszédfelismerő rendszertől, nyilvánvaló, hogy az (emberi) olvasó is képes valamilyen szintű hi-

6. táblázat. Az egyes jelölt szótípusokat és azok kombinációit tartalmazó hibák szótévesztéseinek összege

Javított hibák	Hibák száma	Relatív csökk.	Szótév. száma	Relatív csökk.
Semmi (eredeti kimenet)	401	—	603	—
Könnyű hibák	229	43%	400	34%
Könnyű és nehéz hibák	219	45%	368	39%

bajavításra. Ennek vizsgálatához a beszédfelismerő rendszer által szolgáltatott szószintű átiratokat egy tesztalanynak mutattuk meg, és megkértük, hogy próbálja megtalálni és kijavítani a hibákat. A javított hibákat könnyen és nehezen javítható csoportokra osztottuk, ezután két kijavított változatot vizsgáltunk: az egyikben csak a könnyen javítható hibákat korrigáltuk, a másikban pedig mindkét kategóriát. Ezután mindkét változatra kiszámítottuk a pontosságmetrikát.

A 6. táblázat mutatja, hogyan alakult a hibák és a szótévesztések száma a könnyen, illetve a könnyen és nehezen javítható hibák korrigálása után. Látható, hogy a hibák több mint 40%-át korrigálni lehetett olvasás közben, ezek azonban csak a szótévesztések egyharmadáért voltak felelősek. A nehezen korrigálható hibák kijavításával csak 10-zel csökkent a hibák száma, a szótévesztéseké azonban ennél sokkal jobban; erre részben az a magyarázat, hogy bizonyos hibák javítása nem sikerült tökéletesen, azonban szokszor a szétdaraboltan „felismert” szót egyetlen, ám helytelen szóra cserélt a tesztalany (így az illesztési távolságnál egy szócsere és több szóbeszúrás helyett már csak egyetlen szócsere jelent meg).

Amellett, hogy a kísérlet relevanciáját csökkenti, hogy csak egyetlen tesztalannyal végeztük, a fönti számértékeket egyébként is fenntartással kell kezelnünk. Ennek oka részben az, hogy bizonyos hibákat az emberi olvasó sokszor észre sem vesz (pl. bizonyos ragozási, egybe- és különírási hibák), ami mellett természetesen a szöveget tökéletesen megértette.

6. Konklúzió

Ebben a cikkben azt vizsgáltuk, hogy egy hagyományos felépítésű magyar nyelvű beszédfelismerő rendszer milyen jellegű hibákat vét. Ehhez a teszthalmaz egy részén előforduló szószintű hibákat kigyűjtöttük, majd manuálisan kategorizáltuk és elemeztük. Tapasztalataink szerint a hibák egy jelentős része vezethető vissza OOV szavakra, amely megfelel a várakozásoknak (bár a szótárból hiányzó szavak köre szokatlanul tág). Ugyanakkor a tévesztések egy jelentős része egybe- és különírási hiba, amely sokszor arra vezethető vissza, hogy a nyelvi modell mind az egybe-, mind a különírt formát megengedi. A szóalakok pontos egyezésén alapuló pontosságmetrika ráadásul ezt a típusú hibát nem enyhébb, hanem súlyosabb tévesztésnek tekinti, mint hogyha egy teljesen más jelentésű szót ismertünk volna

fel az adott helyen, mely a magyar (és feltehetőleg az egyéb agglutinatív) nyelvű beszédfelismerőket fokozottan sújtja.

Hivatkozások

1. Bánhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével. In: MSZNY, Szeged (2005) 337–347
2. Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T.: Improved recognition of spontaneous Hungarian speech: Morphological and acoustic modeling techniques for a less resourced task. IEEE Transactions on Audio, Speech, and Language Processing **18**(6) (2010) 1588–1600
3. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8) (1966) 707–710
4. Gosztolya, G., Tóth, L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal. In: MSZNY, Szeged (2010) 224–235
5. Grósz, T., Kovács, Gy.., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY, Szeged (2014) 3–13
6. Grósz, T., Gosztolya, G., Tóth, L.: GMM-free ASR using flat start sequence-discriminative DNN training and Kullback-Leibler divergence based state tying. In: ICASSP. (2016) (beküldve)
7. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: ICASSP. (2009) 3761–3764
8. Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel. In: MSZNY, Szeged (2015) 174–181
9. Abari, K., Olaszy, G., Zainkó, Cs., Kiss, G.: Magyar kiejtési szótár az Interneten. In: MSZNY, Szeged (2006) 223–230
10. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)

Szövegalapú nyelvi elemző kiértékelése gépi beszédfelismerő hibákkal terhelt kimenetén

Tündik Máté Ákos[1], Szaszák György[1]

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik, szaszak}@tmit.bme.hu

Kivonat A cikkünkben felvázolt vizsgálat fókuszában az áll, hogy kiderüljön, milyen mértékű szintaktikai elemzést képes végrehajtani a „magyarlánc" nyelvi elemző a beszédfelismerő által kibocsájtott, hibákkal terhelt szövegeken, és ez az elemzés mennyiben „hasonlít" a hibátlan referenciaszövegen futtatotthoz, illetve azonosítható-e az elemzésnek olyan szintje, részeredménye, amely nagyban korrelál a hibátlan szövegével. A feladathoz egy híradós adatbázis 535 mondatból álló részhalmazát használtuk fel. Ezen a „magyarlánc" nyelvi elemzővel szintaktikai elemzést hajtottunk végre, mely meghatározta a mondatokra a szófaji és függőségi címkéket. Ezt követően a szintaktikai / szemantikai elemzések elemi részekre (szavakra) történő azonosítása és felbontása következett, majd az ezek halmaza felett megvalósított bag of words reprezentáció vizsgálata, melyet a korreláció, hasonlóság mérésére használtuk fel. További összehasonlítás történt a kinyert szófaji és dependencia tagek távolságszámításával is, a szóhibaarány számításával analóg módon. Az eredmények alapján elmondható, a beszéd-szöveg átalakítással nyert szövegeken végzett elemzés nagyban korrelál a hibáktól mentes referenciaátiraton végzettel. [1]

Kulcsszavak: gépi beszédfelismerés, nyelvi elemzés, információkinyerés

1. Bevezetés

A csupán írott szöveget felhasználó tartalomelemző, jelentéskivonatoló, kulcsszókereső alkalmazásokból számosat ismerünk, melynek társadalmi-gazdasági haszna megkérdőjelezhetetlen (pl. az adatbányászat területén). Ugyanezen funkciók beszéden történő megvalósítása nagyobbrészt még várat magára (leggyakrabban a kulcsszó alapú keresés az egyetlen elérhető funkció), holott jelentős társadalmi-gazdasági haszna feltételezhető, hiszen számos archívum vagy egyéb adathalmaz csak beszélt nyelvi adatokat tartalmaz, legépelése, beszéd-szöveg átalakítása nem történik meg.

A beszédfelismerő rendszerek „csak" szöveggé konvertálják a beszédet, azonban egyre inkább előtérbe kerül, hogy legyen emögött valamilyen beszédértést, a

[1] A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely a PD-112598 projekt keretében az itt ismertetésre kerülő kutatást támogatta.

tartalom elemzését, értelmezését megvalósító funkció is. Napjainkban a rendelkezésre álló adatok egyre nagyobb hányada hangfelvétel, így a probléma megoldása egyre inkább hangsúlyossá válik. A hanganyagokon használható elemzők köre jóval szűkösebb, illetőleg két lehetőség kínálkozik: közvetlenül a beszédből nyerjük ki az információt, illetve a beszédet szöveggé alakítva szöveges elemzőket, keresőket alkalmazunk.

Az utóbbi lehetőség azt a potenciált is magában rejti, hogy az írott nyelvre már rendelkezésre álló elemzőeszközöket is felhasználhatjuk. Ennek során alapvető problémaként jelentkezik, hogy a beszédfelismerővel átalakított szöveg többkevesebb felismerési hibát – szócserét, -törlést vagy -beszúrást – tartalmaz. A kutatás első lépéseként azt vizsgáltuk, mennyire működik hatékonyan a nyelvi elemzés a beszédfelismerő által szolgáltatott kimeneten. Munkánkhoz a „magyarlánc" magyar nyelvű, függőségi nyelvtan alapú szintaktikai elemzőt [1] használtuk, egy médiából származó híranyagokon [2] alkalmazott beszéd-szöveg átalakítási lépés után. A beszédfelismerő közel 35% szóhibaaránnyal működött a választott anyagokon.[2]

Vizsgálatunk arra irányult, hogy kiderüljön, milyen mértékű szintaktikai elemzést képes végrehajtani az elemző a hibákkal terhelt szövegen, és ez az elemzés mennyiben „hasonlít" a hibátlan szövegen futtatotthoz, illetve azonosítható-e az elemzésnek olyan szintje, részeredménye, amely nagyban korrelál a hibátlan szövegével[3].

A cikkünk az alábbi struktúra szerint épül fel: elsőként bemutatjuk a korpuszt, illetve a „magyarlánc" nyelvi elemzőből kinyert adatokon végzett, összehasonlíthatóságot célzó utófeldolgozást. Ezután sor kerül a beszédfelismerő kimenetének és a referenciaszöveg feldolgozásának ismertetésére, amely több lépést is magában foglal. Végül a vizsgálati eredményeket ismertetjük az elemzések részletes, többszintű összehasonlítása mellett. A puszta korreláció mérése mellett további összehasonlításokat, hasonlósági és távolsági metrikákat is megadunk, a kinyert szófaji (POS) és függőségi (DEP) adatokra.

2. Anyag és módszer

2.1. A felhasznált híradatbázis

A kísérleteinkhez magyar nyelvű televíziós hírműsorok felvételeit használtuk fel. A felvételek két közszolgálati és két kereskedelmi csatornáról származtak, hozzávetőleg egyenletes eloszlásban, mondatszinten leiratozva. Összesen 535 mondatnyi anyagot választottunk ki vizsgálatra véletlenszerűen, de a hírblokkok egységét megtartva.

[2] A hivatkozott beszédfelismerő ennél lényegesen kisebb szóhibaarányt szolgáltat, esetünkben szándékosan állítottunk be ezt a magasabb értéket.

[3] A választott anyagokra nem áll rendelkezésünkre „gold standard" elemzés, ugyanakkor nem is célunk a „magyarlánc" elemző abszolút pontosságának mérése, munkánkhoz elegendőnek tartjuk a helyes referenciaszöveggel való összevetést.

Egy-egy hangfájl jellemzően egy hírblokkot tartalmazott, amelyet valós idejű médiafeliratozásra fejlesztett beszédfelismerő rendszerrel [2] szöveggé alakítottunk. A felismerést ezúttal szándékosan viszonylag magas, átlagosan 35%-ot közelítő szóhibaarányt szolgáltató akusztikai és nyelvi modell kombinációval végeztük, a felismert anyagokon pedig a szóhibaarány viszonylag nagy szórást mutatott (lásd 6. ábra), ami szempontunkból a teljes körű analízishez és az egyes eredmények szóhibaarány függésének megadásához kedvező beállítás.

2.2. Utófeldolgozás és adatreprezentáció

A referenciaszövegen és a beszédfelismerő kimenetét tartalmazó szöveges átiraton a „magyarlánc" nyelvi elemzővel végrehajtottuk a szintaktikai elemzést, mely meghatározta a mondatokra a szófaji és függőségi címkéket [1].

A feladat a beszédfelismerő kimenetének és a referenciaszöveg normalizálásával kezdődött, kézi központozással. A beszéd-szöveg átalakítás másik nehézsége a szóhibákon túl, hogy az írásjelek, központozás sem minden esetben megoldott. Jelen munkában ettől eltekintünk, és a központozást kézzel pótoljuk, amire különösen azért van szükség, mert a nyelvi elemző erre nagymértékben támaszkodik.

A szintaktikai / szemantikai elemző kimenetén előállt szófaji és dependencia tageket információ-visszakereső rendszerekben használt vektortér modellbe (vector space model) transzformáltuk. Ez a modell magában foglalja a szózsák (bag of words)-megközelítést is. Az információkeresésben ismert szózsák modellben a szavak dokumentumon belüli előfordulási gyakorisága az, ami számít, nem a sorrendjük. Ebben a modellben az *a fa zöld* és az *a zöld fa* rövid dokumentumok azonosan fognak viselkedni. Világos, hogy eltérő jelentésűek, azonban az is igaz, hogy mindketten relevánsak a fákat és a zöld színt kulcsként tartalmazó lekérdezésekre.

A vektortér modell eredeti ötlete szerint minden egyes dokumentumot (a mi esetünkben a dokumentumpárok egy-egy mondatpárnak felelnek meg) unigram szógyakoriságok vektoraként ábrázolnak. Ezt a modellt felhasználva, a szófaji tagek gyakoriságát és a dependencia tagek gyakoriságát vizsgáltuk az egyes mondatpárokra, valamint az indikátorvektorokat is megadtuk, ami prezencia / abszencia jellegű viselkedést ír le. Megfontolásaink szerint ugyanis egy információkinyerést célzó felhasználásban is legfőképpen a szófaji és a függőségi elemzésre való támaszkodás dominál [3].

Ezen kívül a mondatokban előforduló szófaji és függőségi címkék helyett azok szófaji és függőségi címkelistában elfoglalt sorszáma szerint is megcímkéztük a tokeneket (szavakat), melynek adatreprezentációja az 1. ábrán látható, egy adott példamondatpárra vonatkoztatva.

Prezencia vagy abszencia vizsgálata esetén az előbbi vektorok 0 értékei 0, 0-nál nagyobb értékei 1 értékűek lennének. A 2. ábrán a gyakorisági szózsák (pontosabban szófajzsák) reprezentáció látható ugyanerre a mondatra a szófaji címkék uniója alapján.

```
Referenciaszöveg: Szerbiában a pravoszláv karácsony állami ünnep.
Felismert szöveg: Szerbiában a pravoszláv karácsony áll    aminek.

Szófaji gyakorisági maszk
[N, V, A, P, T, R, S, C, M, I, X, Y, Z, 0]
Referencia szófaji gyakoriság
[3, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Felismerésre vonatkozó szófaji gyakoriság
[2, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

1. ábra. Szófaji alapú gyakorisági reprezentáció egy példamondatpárra

```
Szózsák maszk
[N, V, A, P, T ]
Referencia szózsák
[3, 0, 2, 0, 1 ]
Felismerésre vonatkozó szózsák
[2, 1, 1, 1, 1 ]
```

2. ábra. Szófaji alapú szózsák reprezentáció az előző példamondatpárra

2.3. Az adatsorok összehasonlításához használt mértékek

Az adatsorok összehasonlításánál több szempontot is figyelembe kellett venni, tekintettel arra, hogy kategorikus adatokról van szó. Az egyik fő megközelítés a **prezencia / abszencia** szempontú vizsgálat, amely azon alapul, hogy mely POS és DEP tagek fordulnak elő az egyes adatsorokban. Továbbhaladva, számításba vettük az egyes kategóriák előfordulásának **gyakoriság**át is, erre kétféle hasonlóságot vetettünk be; egyrészt az összes címke halmazát használtuk fel, másrészt az aktuális referenciamondat és beszédfelismerő kimenet szófaji, illetve függőségi címkéinek unióját vettük, és azon halmaz felett hajtottuk végre az összehasonlítást. Hasonlóságot kerestünk az adatsorok között úgy is, hogy a kategorikus címkéket az előre definiált címkelistában elfoglalt sorszámukkal helyettesítettük a vektorban, és így vetettük össze a vektorokat. Ez utóbbi eljárásra **sorrendi** összehasonlításként utalunk a továbbiakban.

Elsőként az adatsorok Pearson-korrelációját határoztuk meg. A korreláció jelzi azt, hogy két tetszőleges érték nem független egymástól. Az ilyen széles körű használat során számos együttható, érték jellemzi a korrelációt, alkalmazkodva az adatok fajtájához:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}}, \tag{1}$$

ahol a felülvonásos betűk a várható értéket, X és Y pedig az adatsorokat jelölik.

A hasonlóság másik lehetséges mértéke a felismert- és a referenciavektor skaláris szorzata. Geometriailag a két vektor skaláris szorzata az általuk bezárt

szög koszinusza, azaz ha két ilyen vektor koszinuszát maximalizáljuk (amennyiben azonos kvadránsban találhatók), akkor az általuk bezárt szög nullához közeli lesz. Ezen alapul az úgynevezett koszinusz-hasonlóság számítása:

$$sim(d_j, d_k) = \frac{d_j d_k}{|d_j||d_k|} = \frac{\sum_i w_{i,j} w_{i,k}}{\sqrt{\sum_i (w_{i,j})^2}\sqrt{\sum_i (w_{i,k})^2}} \quad (2)$$

A koszinusz-hasonlóságot és a Pearson-korrelációt a referenciában és a beszédfelismerő kimenetéből kinyert POS bigramok prezencia / abszencia és gyakoriság alapú vektorreprezentációira is kiszámítottuk.

Az adatsorokat megvizsgálva gyakran előfordult, hogy a beszédfelismerő kimenetén megjelenő szóbeszúrás vagy szókihagyás miatt az adatsorok hasonlósága rosszabb értéket mutatott annál, mint amivel intuitívan „ránézésre" rendelkezett, hiszen a hasonlósági mértékeink esetén főként az egyes címkék páronkénti összehasonlítására koncentráltunk. A POS-tagek alapján történő összehasonlításnál ennek kapcsán felhasználtuk a bioinformatikában használt Needleman-Wunsch globális szekvencia-illesztő algoritmust [4], melyet 4 pontozási értékkel súlyoztunk. Ha megegyeztek a i. indexen talált karakterek, ez 1 pontot ért, ha pedig nem, akkor az 0-t. Ha egy új hézagot kellett nyitni az igazításhoz, azt -0.5 ponttal büntette az algoritmus, ha pedig meghosszabbítani kellett, azt -0.1 ponttal. Az algoritmus hasonlít a sztringek összehasonlításához használt Levenshtein-távolsághoz [5], de annyiban meghaladja azt, hogy konkrét illesztési eredményeket szolgáltat a szekvenciákra, melyből a legnagyobb pontszámot választjuk ki, mivel ott a legnagyobb az egyezés.

Alább egy példát közlünk, ahol jól látszódik az igazítás haszna. Vegyük az alábbi referenciamintára és a beszédfelismerő kimenetére futtatott szófaji elemzést:

```
NRVTNS
CNRVTNS
```

Így Needleman-Wunsch igazítás nélkül a páronkénti összehasonlításból adódó korreláció értéke: -0,934 lesz. Ugyanakkor, ha felhasználjuk az igazító algoritmust, az alábbi rendezést kapjuk:

```
-NRVTNS
CNRVTNS
```

Így a korreláció értéke máris a valósághoz közelebb esően alakul: 0,895.

A függőségi címkékre ezt az illesztési módszert nem alkalmaztuk, helyette a nemzetközileg is használt kiértékelési paramétereket választottuk, azzal a kényszerrel élve, hogy csak az egyező hosszúságú mondatokra határoztuk meg. A LAS (Labeled Attachment Score) esetében azok a függőségi ívek érnek pontot, ahol a beszédfelismerő kimenetén lévő adott ív mind a szülőobjektumot tekintve (ez egy sorszám), mind az ívre írt függőségi élcímke megegyezik a referencia átirat függőségi ívéhez viszonyítva, míg az ULA (Unlabeled Attachment Score) esetében elégséges a szülő csomópont egyezése (itt nem számít hibának a rossz élcímke).

```
Referenciaszöveg: [...] amit látok, az tényleg megtörténik [...]
Felismert szöveg: [...] amit látok, azt tényleg megtörténik [...]

LAS[%]=94,12; UAS[%]=100; LA[%]=94,12.
```

3. ábra. Függőségi alapú összehasonlítás

A LA (Label Accuracy) esetén pedig a függőségi élcímkék egyezése számít [6]. A 3. ábrán egy példát is láthatunk.

Megállapítható tehát, hogy *az-azt* páros eltérése a függőségi kapcsolatokat az élcímkék szintjén befolyásolta, viszont maguk a függőségi ívek nem változtak. Láthatjuk, hogy ebben az esetben a globális illesztő függvény alkalmazása további alapos megfontolásokat igényelne (pl. milyen karakterrel jelöljük az igazítási hézagokat, és milyen címkét kapjanak?), így ezt nem alkalmaztuk.

A következő összehasonlítás a szóhibaarány (WER: Word Error Rate) mintájára történt. Ennek a képlete:

$$WER = \frac{S + D + I}{N}, \tag{3}$$

ahol S jelenti a szócserék számát, D a törölt szavak számát, I a szóbeillesztések számát, N pedig az eredeti szóhalmaz méretét. Ennek mintájára megalkottuk a csak szótövekre értelmezett SER, mint Stem Error Rate; a szófaji címkékre értelmezett PER, mint POS Error Rate; valamint a függőségi címkékre értelmezett DER, mint Dependency Error Rate mérőszámokat. Az utóbbi összefüggéseket a címkékből képzett bigramokra is meghatároztuk.

A szófaji és függőségi tagek vizsgálata mellett mondatszintű jellemzőket is meghatároztunk, úgy, mint pl. az átlagos Levenshtein-távolság és Jaro-Winkler távolság [7].

3. A vizsgálati eredmények

3.1. Szófaji címkék hasonlósága

Az 1. táblázatban láthatóak a referenciaszöveg és a beszédfelismerő hibákkal terhelt kimenetének Pearson-korreláció és koszinusz-hasonlóság alapú kapcsolatára vonatkozó összehasonlítás eredményei, melyeket az unigram szófaji címkék, valamint a belőlük képzett bigramok között határoztunk meg.

A továbbiakban értékeljük a különböző megközelítéseket, elsősorban az automatikus rendszerekben történő felhasználhatóság szempontjából. A prezencia / abszencia (1 / 0) jellegű értékek a kapcsolat erősségének szempontjából a leggyengébbek, hiszen nem veszik figyelembe az egyes szófaji címke gyakoriságokat. A gyakoriságot figyelembe vevő értékek közül az összes szófaji címkére vonatkozó a gyengébb a szózsák megközelítéshez képest, hiszen az összehasonlításnál az összes címke halmazán több közös abszencia adódik, ami így pozitívan súlyozza az eredményt, míg a szózsák modellt a referenciaszöveg és a beszédfelismerő

1. táblázat. Pearson-korreláció és koszinusz-hasonlóság szófaji címkékre és belőlük képzett bigramokra

Vizsgálattípus	Pearson-korreláció		Koszinusz-hasonlóság	
	unigram	bigram	unigram	bigram
Prezencia / abszencia jellegű (összes címke)	0,86	0,74	0,91	0,75
Gyakoriság jellegű (összes címke)	0,89	0,76	0,92	0,77
Páronkénti, sorrendi alapú	0,54	–	0,84	–
Páronkénti, sorrendi, Needleman-Wunsch	0,65	–	0,88	–
Gyakoriság jellegű (szózsákra, címkeunió)	0,73	0,30	0,92	0,77

kimenetének szófaji címkéinek uniójából képzett tér felett értelmeztük. A legnagyobb információértéket a páronkénti összehasonlítás képviselné, hiszen ezeknél a pozícióinformációt is figyelembe vettük, vagyis hogy a mondat megfelelő sorszámú tokenjei azonos szófajú címkével rendelkeznek-e. Jól látszik a Needleman-Wunsch algoritmus pozitív hatása, az adatsorok egymásra igazításával nőtt a Pearson-korreláció és a koszinusz-hasonlóság is. Ugyanakkor tudni kell, hogy a páronkénti összehasonlításon alapú értékek az átlagra és a szórásra érzékenyek, holott ez kategorikus adatoknál nem szabadna figyelembe venni, ezért az alábbi megközelítés matematikailag helytelen. Kijelenthető tehát, hogy jelenleg a mérőszámok közül a legoptimálisabb a gyakoriságon alapuló szózsák megközelítés a referenciában és a felismert szövegben előforduló POS tagek uniója felett.

3.2. Függőségi címkék hasonlósága

Táblázatba foglaltuk a függőségi címkékre vonatkozó hasonlósági értékeket is (2. táblázat). A szófaji címkéknél leírt, az összefüggés erősségét érintő megállapítások érvényesek a függőségi címkékre is. Ugyanakkor a nemzetközileg elterjedt LAS / UAS / LA mértékek a pozícióinformációt is figyelembe veszik, ezért ezek mutatják a legerősebb összefüggést (lásd 3. táblázat). A hátrányuk viszont az, hogy ezt a módszert megegyező mondatok „gold standard" szerinti elemzésének és függőségi elemző szerinti elemzésének összehasonlítására találták ki, amely magával vonja azt, hogy a függőségi címkék száma is egyenlő. Ezeknek a mértékeknek a különböző hosszúságú mondatokra történő adaptációja további megfontolásokat igényel.

2. táblázat. Pearson-korreláció és koszinusz-hasonlóság függőségi címkékre

Vizsgálattípus	Pearson-korreláció	Koszinusz-hasonlóság
Prezencia / abszencia jellegű (összes címke)	0,82	0,87
Páronkénti, sorrendi alapú	0,49	0,74
Gyakoriság jellegű (szózsákra)	0,63	0,89

3. táblázat. Függőségi címkék összehasonlítása

Vizsgálattípus	Hasonlóság
LAS	80,5 %
UAS	86,6 %
LA	84,3 %

3.3. Hibaarány jellegű jellemzők

Vegyük sorra először az elemzési egységre vonatkoztatott hibaarány alapú értékeket (4. táblázat). Látható, hogy az unigram megközelítésben a legszigorúbb a szóhibaarány alapú megközelítés, hiszen előfordulhat, hogy a szótövező ugyanazt rendeli a referenciában és a felismert szövegben előforduló tokenhez. Ha pedig a szótő sem egyezik, ettől még előfordulhat, hogy ugyanolyan szófajú tokenre téved a felismert szöveget feldolgozó nyelvi elemző, mint ami a referenciaszövegben van. Ugyanakkor látható, hogy a bigram megközelítésben is elviselhető hibaarány mutatható ki, közel esik a korábban említett, beszéd-szöveg átalakításból eredő 35% körüli szóhibaarányhoz. A **részletes** POS- és POS-bigram hibaarány meghatározására is lehetőségünk volt, mivel nemcsak a fő szófaji címkék álltak rendelkezésre, hanem „finomszemcsés" POS címkék is (pl. igék esetén az igeragozást is tartalmazza).

4. táblázat. Hibaarányok

Vizsgálattípus	Hibaarány értéke
Szóhibaarány	0,35
Szótőhibaarány	0,29
POS hibaarány	0,22
POS bigram hibaarány	0,34
Részletes POS hibaarány	0,29
Részletes POS bigram hibaarány	0,43
DEP hibaarány	0,25
DEP bigram hibaarány	0,39

Az 5. táblázatban néhány általános szövegszintű jellemzőt is megadunk a felhasznált korpuszra.

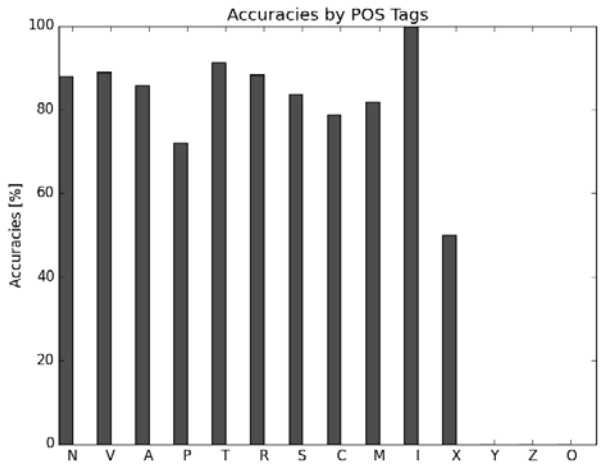
A 4. ábrán a szófaji címkék szerinti pontossági értéket ábrázoltuk, az egyező hosszúságú mondatokra, a „magyarlánc" elemző által használt szófaji kódokat megtartva [1]. Az indulatszó (I) kategória pontosságát figyelmen kívül hagyva - alacsony elemszáma miatt - a legjobb egyezéseket a főnév (N), ige (V) és a névelő (T) kategóriákban adta az elemző.

A függőségi viszonyokra vett kategóriánkénti megoszlás a 5. ábrán látható, az egyező hosszúságú mondatokra. Látható, hogy az elemző a legpontosabb a főnévhez tartozó névelő, a mondat gyökereként (ROOT) meghatározott igéhez tartozó igekötő (PREVERB), a helyhatározós (TLOCY) és a jelzős szerkezetek (ATT)

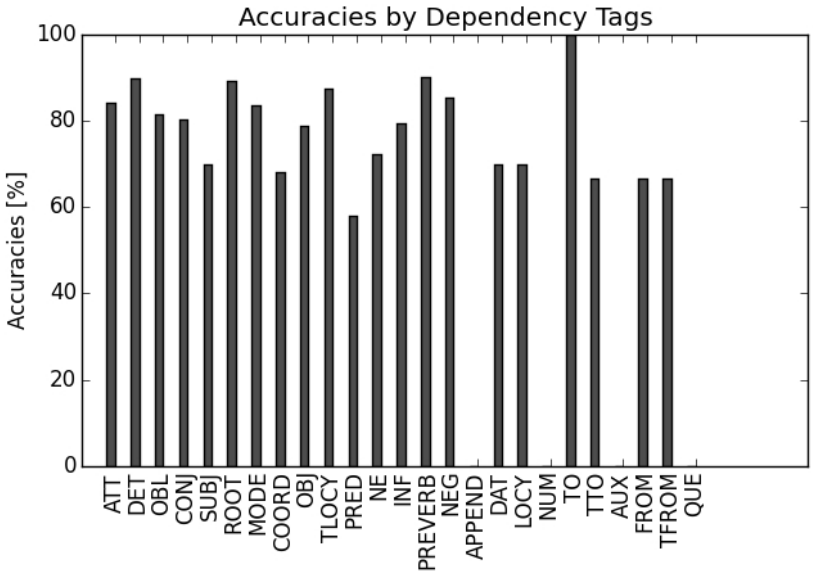
5. táblázat. Néhány szövegszintű jellemző

Vizsgálattípus	Érték
OOV-arány a referencia mondatokra	0,013
OOV-arány a felismert mondatokra	0,012
Referenciaszavak száma	2968
Felismert szövegben a szavak száma	3018
Mondatszintű Levenshtein-távolság	12
Mondatszintű Jaro-Winkler-távolság	0,92
Leghosszabb mondat (szóban mérve)	43
Átlagos mondathossz (szóban mérve)	12,17

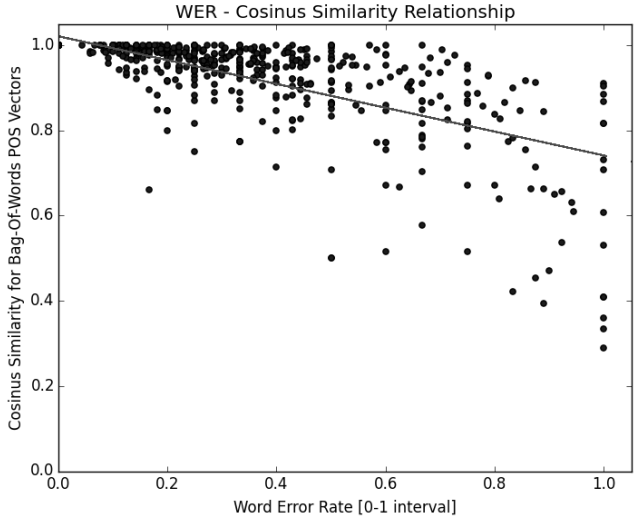
megállapításában volt. A szóhibaarány és a gyakoriság alapú, szófajokra tekintett koszinusz-hasonlóság közötti összefüggést a 6. ábra mutatja. Az összefüggés a várakozásoknak megfelelő, a szóhibaarány romlása a koszinusz-hasonlóság gyengülését vonja maga után, lineáris regressziót alkalmazva -0,279 meredekség és 1,02-es metszéspont adódott. Látható az adatokból, hogy az összefüggés közel lineáris, nem azonosítható tehát olyan szóhibaarány-küszöbérték, amelyet elérve a koszinusz-hasonlóság meredeken esni kezdene.



4. ábra. Az egyes szófaji kategóriák pontossága



5. ábra. Az egyes függőségi kategóriák pontossága



6. ábra. A szóhibaarány és a szófajokra vonatkozó, gyakoriság alapú koszinusz-hasonlóság összefüggése

4. Összegzés

A „magyarlánc” nyelvi elemzővel elért eredmények alapján elmondható, hogy a beszéd-szöveg átalakítással nyert szövegeken végzett elemzés nagyban korrelál a beszédfelismerési hibáktól mentes (referenciaátirat) szövegen végzettel. A kapott eredmények tanúsága szerint a vizsgált híranyag korpuszon a szófaj, illetve a függőségi viszony tévesztését (megváltozását) is eredményező beszédfelismerési hibák száma az összes felismerési hiba mintegy 2/3-ára tehető. Bigram kapcsolatokat tekintve a szófajtévesztés valószínűsége nagyon közel esett a beszéd-szöveg átalakítás szóhibaarányához. Az automatikus információkinyerés és tartalmi kivonatolás szempontjából leginkább releváns főnévi és igei szófajkategóriák esetében a tévesztési arányok még kedvezőbbek, kevesebb, mint a beszédfelismerési hibák felére tehetőek. Jóllehet a józan megfontolás alapján is szoros összefüggést várnánk a szóhibaarány és a felismerési hibákkal terhelt, valamint a hibátlan szövegek szintaktikai elemzéseinek hasonlósága között, kísérletileg is megerősítettük, hogy ez az összefüggés közel lineáris, amiből az következik, hogy automatikus beszéd-szöveg átalakítással nyert szövegek nyelvi elemzésekor nem található olyan kritikus szóhibaarány érték, amelyen túl a nyelvi elemzés drasztikus, fokozódó leromlásával kellene számolni, és így a szóhibaarány alapján jól előrejelezhető az elemzés pontossága is. Az eredmények alapján érdemes lenne a beszéd-szöveg átalakítást, majd nyelvi elemzést megvalósító feldolgozási láncot közvetlenül „gold standard” referenciával összevetve kiértékelni. Ígéretes jövőbeli kutatási irány lehet továbbá a beszéd elemzéséből (pl. hangsúlydetekcióból vagy általánosabban prozódiaalapú elemzésből) származó, szavakhoz rendelt tagek továbbvitele akár a nyelvi elemzésbe, akár közvetlenül az információkinyerésbe.

Hivatkozások

1. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP (2013) 763-771
2. Tarján, B., Fegyó, T., Mihajlik, P.: A Bilingual Study on the Prediction of Morph-based Improvement. In: 4th International Workshop on Spoken Languages Technologies for Under-Resourced Languages, Saint Petersburg, Russia (2014) 131-138
3. Lioma, C. and Blanco, R.: Part of speech based term weighting for information retrieval. In: M. Boughanem, C. Berrut, J. Mothe and C. Soule-Dupuy (Eds.), ECIR, LNCS Vol. 5478 (2009) 412-423
4. Beddoe, Marshall A.: Network protocol analysis using bioinformatics algorithms, http://www.4tphi.net/ awalters/PI/pi.pdf (2004)
5. Singh, S. P., Kumar, A., Darbari, H., Chauhan, S., Srivastava, N., Singh, P.: Evaluation of Similarity metrics for translation retrieval in the Hindi-English Translation Memory. Int. Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 8 (2015)
6. Choi, Jinho D., Tetreault, J., Stent, A.: It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL. (2015)
7. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string metrics for matching names and records. Kdd workshop on data cleaning and object consolidation. Vol. 3. (2003)

Nevetések automatikus felismerése mély neurális hálók használatával

Gosztolya Gábor[1,2], Beke András[3], Neuberger Tilda[3]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: ggabor@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
[3] MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33., e-mail: {beke.andras, neuberger.tilda}@nytud.mta.hu

Kivonat A nonverbális kommunikáció fontos szerepet játszik a beszéd megértésében. A beszédstílus függvényében a nonverbális jelzések típusa és előfordulása is változik. A spontán beszédben például az egyik leggyakoribb nonverbális jelzés a nevetés, amelynek számtalan kommunikációs funkciója van. A nevetések funkcióinak elemzése mellett megindultak a kutatások a nevetések automatikus felismerésére pusztán az akusztikai jelből [1,2,3,4,5,6]. Az utóbbi években a beszédfelismerés területén, a keretszintű fonémaosztályozás feladatában uralkodóvá vált a mély neurális hálók (DNN-ek) használata, melyek háttérbe szorították a korábban domináns GMM-eket [7,8,9]. Jelen kutatásban mély neurális hálókat alkalmazunk a nevetés keretszintű felismerésére. Kísérleteinket három jellemzőkészlettel folytatjuk: a GMM-ek esetében hagyományosnak számító MFCC és PLP jellemzők mellett alkalmazzuk az FBANK jellemzőkészletet, amely 40 Mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból áll. Vizsgáljuk továbbá, hogy az egyes frekvenciasávok milyen mértékben segítenek a mély neurális hálónak a nevetést tartalmazó keretek azonosításában. Ezért a dolgozat második részében kísérletileg rangsoroljuk, hogy az egyes sávok mennyire járulnak hozzá a mély neurális háló pontosságának eléréséhez.[1]

Kulcsszavak: nevetés, akusztikus modellezés, mély neurális hálók

1. Bevezetés

A nonverbális kommunikációnak nagy jelentősége van a beszédészlelés és a beszédmegértés során. Az üzenet átadásában fontos szerepet játszhatnak mind a vizuális elemek (pl. gesztusok, szemkontaktus), mind pedig a nem verbális hangjelenségek, mint amilyen a nevetés, a torokköszörülés vagy a hallható ki-, illetve

[1] Jelen kutatási eredmények megjelenését a „Telemedicina-fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken” című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatja. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatás továbbá a 108762 számú OTKA támogatásával jött létre.

belégzés. A természetes kommunikáció során a hallgatók egyidejűleg dekódolják a különböző modalitásokból érkező (vizuális és auditív) információkat, a multimodális percepció során a feldolgozási műveletek nem csupán a hallottakra, hanem a látottakra is kiterjednek.

A nevetések a spontán beszéd relatíve gyakori kísérőjelenségei, amelyek számos funkcióval rendelkezhetnek. Gyakoriságát tekintve spontán társalgásokban 1-3 előfordulást adatoltak percenként [10,11], de természetesen a nevetés előfordulását sok tényező befolyásolja, így például a beszédtéma, a kontextus, a társalgó partnerek ismeretsége, valamint a hierarchiaviszonyok. A nevetés funkcióját tekintve általában a beszélő érzelmi állapotának jelölője, az öröm fajspecifikus jelzője, a humor velejárója. Társas jelzés, társadalmilag kialakított, könnyen dekódolható jelenség. A kutatók számára azért lehet fontos a nevetések különböző szempontú vizsgálata, mert általa többet tudhatunk meg az emberi viselkedésről, a szociális interakció szerveződéséről. A nevetés a társalgásban számtalan funkciót tölthet be; gyakran kontextualizációs utasításként vagy értelmezési keretet pontosító jelzésként működik, utólag pontosíthatja, illetve eleve kijelölheti a társalgási stílust és a hozzátartozó értelmezési keretet [12,13].

A kutatások többsége elkülönít nevetéstípusokat. Günther [14] a diskurzus szerkezeti szerveződése szerint az alábbi típusokat különbözteti meg: csatlakozó/barátkozó, kontextualizáló, ellenkező/kihívó, reflexív, heterogén nevetés, valamint a be nem sorolható esetek (Günther [14]: 153-161; idézi Hámori [13]: 116). A perceptuális benyomás szerint Campbell és munkatársai [4] négy fonetikai típusra osztották a nevetéseket: zöngés nevetés, kuncogás, levegős és nazális nevetés. A produkciós oldalról vizsgálva az akusztikum alapján a fő megkülönböztető jegy, hogy a nevetés zöngével vagy a nélkül valósult meg (így megkülönböztethetők pl. zöngés énekszerű, zöngétlen horkantásszerzű vagy kevert típusok, vö. [15]). A nevetés produkcióját tekintve a társalgó partnerek részvételének szempontjából megkülönböztethetők az alábbi típusok: önálló nevetés, együttnevetés, nevetés a társalgó partner beszéde alatt (háttércsatorna-jelzésként), nevetős beszéd, kevert típus [11,16]. A korábbi akusztikai vizsgálataink alapján megállapítható volt, hogy ezek a típusok eltérő időtartamban, valamint eltérő harmonikus-zaj aránnyal (HNR) jelennek meg.

A nevetések akusztikai jellemzőit számos tanulmány elemezte a nemzetközi szakirodalomban [17,18,15,19], a magyar nyelvre vonatkozóan azonban alig akad ilyen jellegű munka [11,16,20]. A vizsgálatok szerint a nevetések akusztikai jellemzői (F0, formánsok, amplitúdó, zöngeminőség) a beszédhez hasonlatosak, azok hehezetes CV /hV/ szótagok sorozataként realizálódnak, bár a beszédhez képest hosszabb zöngétlen résszel valósulnak meg. A szöveges részektől való elkülönítésükben (a nevetések detektálásában) nagy szerepet játszik a zöngétlen-zöngés rész aránya. Két, amerikai angol beszélő (egy nő és egy férfi) nevetéseit vizsgálva kimutatták, hogy egy szótagnyi nevetés átlagos időtartama 204, illetve 224 ms, és a nevetések átlagosan 6,7, illetve 1,2 szótagból állnak [17]. A további eredmények szerint a nevetés produkciójában másodpercenként átlagosan 4,7 szótag adatolható, ami nagy hasonlóságot mutat az (angol, francia és svéd) olvasott mondatok másodpercenkénti szótagszámával. Német és olasz anyanyelvű beszélők esetében

azt találták, hogy a nevetések átlagos időtartama 798 ms nőknél és 601 ms férfiaknál, valamint hogy az alaphangmagasság átlagosan 472 Hz nőknél és 424 Hz férfiaknál, tehát a beszéd és a nevetések megkülönbözteté-sében az alapfrekvencia értékének is jelentős szerepe van [18]. Bachorowski és munkatársai [15] összehasonlítottak szakirodalmi adatokat a nevetések átlagos alapfrekvencia-értékeire vonatkozólag, amelyek nők esetében 160 és 502 Hz, férfiak esetében 126 és 424 Hz közötti értékkel jelentek meg a különböző kutatásokban. A magyar nevetéseket a BEA adatbázis [21] hanganyagaiban vizsgálták, és azt találták, hogy az átlagos időtartamuk 911 ms (átlagos eltérés: 605 ms), átlagos F0-értékük 207 Hz (átl. elt.: 49 Hz) férfiaknál, 247 Hz (átl. elt.: 40 Hz) nőknél. A nevetések számos akusztikai paraméterben (jitter, shimmer, jel-zaj viszony, F0-átlag) szignifikánsan különbséget mutattak a beszédszegmentumokhoz (jelen esetben szavakhoz) képest, ami megkönnyítheti az elkülönítésüket a szöveges részektől.

2. Nevetések felismerése

A nevetések automatikus osztályozása megközelítőleg egy évtizedes múltra tekint vissza. A nevetések leggyakrabban önállóan fordulnak elő, egy részük azonban nem önállóan, hanem a beszéddel egyidejűleg jelenik meg a társalgásokban [22,23], de gyakoriak az együttnevetések is, vagyis amikor két beszélő szimultán nevetése hangzik el. Kennedy és Ellis [2] tanulmányukban az átfedő nevetéseket is elemezték, a detektálásukhoz SVM-et használtak osztályozó algoritmusként (az MFCC-t, a spektrális ingadozást és a két asztali mikrofon jele közötti időbeni eltéréseket vizsgálva). A rendszerükkel 87%-os helyes osztályozási eredményt értek el. A Chist Era JOKER projekt keretein belül Tahon és Devillers a pozitív és a negatív hangulatú nevetések automatikus detektálását tűzték ki célul [24]. A nevetéseket az alaphangmagassággal, a formánsaival, intenzitásukkal és spektrális jellemzőikkel reprezentálták. A kutatásukban a WEKA szoftver SMO osztályozót használták RBF kernelfüggvénnyel. A tanító korpuszban a pozitív nevetések száma 140 db volt, míg a negatív nevetéseké 117 db. A tesztadatbázis 48 pozitív nevetést tartalmazott, és 27 negatívat. Az eredmények szerint a pozitív nevetések F-értéke $64,5\%$, míg a negatívaké $28,5\%$ volt.

Számos kutatás tűzte ki céljául a nevetések automatikus felismerését. Truong és van Leeuwen [1,3] Gauss-keverék modellel és perceptuális lineáris predikciós (PLP) akusztikai előfeldolgozási eljárással mutattak ki $87,6\%$-os helyes osztályozási eredményt. A PLP együtthatóinak száma befolyással lehet a nevetések osztályozási eredményére. Petridis és Pantic [25] kutatásukban kimutatták, hogy ha 13 együtthatót tartalmazó PLP-t használnak feed-forwald neurális hálózattal kombinálva, akkor az F-érték 64%, míg egy másik munkájukban 7 együtthatót tartalmazó PLP-vel 68% volt [26]. Reuderink és munkatársai [27] a PLP-RASTA jellemzőt használták a nevetések detektálására, illetve GMM és HMM osztályozót. Az eredmények azt mutatták, hogy a GMM osztályozó kicsivel jobb eredményt adott (átlagos AUC-ROC 0,825), mint a HMM (átlagos AUC-ROC 0,822). A PLP jellemző mellett igen népszerű az MFCC akusztikai reprezentáció. Ahogy a PLP esetén, úgy az MFCC együtthatóinak számának megválasztása sem

egyértelmű. Kennedy és Ellis [2] csak az első 6 MFCC együtthatót használták, és hasonló eredményt érték el, mint a 13 együtthatót tartalmazó MFCC-vel. Mindez arra utalhat, hogy a nevetések megkülönböztető akusztikai jellemzőit az alsóbb frekvenciasávokban érdemes keresni. A spektrális akusztikai jellemzők mellett vizsgálták a prozódia szerepét a nevetések detektálásában. Az eredmények azt mutatták, hogy a prozódia jól reprezentálja a nevetések dinamikus jellemzőit [1,28].

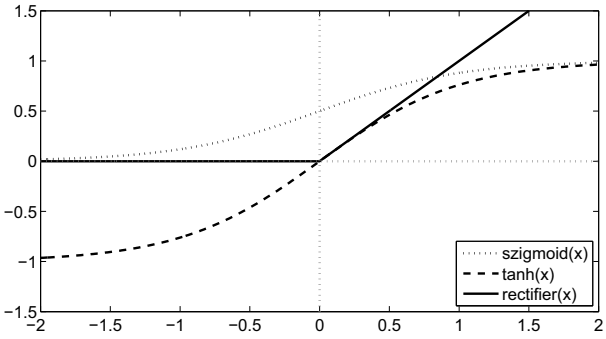
Számos kutatás nem egy-egy akusztikai jellemzőt alkalmazott, hanem több jellemző kombinációját. Knox és Mirghafori [6] neurális hálózatok alkalmazásával (MFCC és alaphangmagasság kombinált paraméterekkel) 90% fölötti teljesítményt értek el a nevetések detektálásában. Hasonlóan, a magyar spontán beszédből származó nevetések esetében is 90% fölötti eredményt mutatott a nevetések és a beszédszegmensek osztályozása GMM-SVM kevert módszerrel MFCC, PLP és akusztikai paramétereken tesztelve [16]. Egy másik vizsgálatunkban [29] különböző osztályozási technikákat és ezek kombinációit (GMM-ANN, GMM-SVM) alkalmaztuk a nevetések és a szavak osztályozásához. A legjobb eredményt (EER: $2,5\%$) az MFCC és az akusztikai paraméterek használatával a GMM és SVM kombinált osztályozóval értük el.

Az akusztikai jellemzők mellet fontos kérdés, hogy milyen osztályozó eljárást érdemes alkalmazni a nevetések detektálásában. A kutatások többsége vagy SVM-et [27], vagy GMM-et [1,27,30], vagy ANN-t [6,25,26,28] alkalmazott. Az utóbbi években a beszédfelismerés területén, a keretszintű fonémaosztályozás feladatában uralkodóvá vált a mély neurális hálók (DNN-ek) használata, melyek háttérbe szorították a korábban domináns GMM-eket. Korábbi kísérleteink során mi is kiemelkedő eredményeket értünk el a használatukkal mind magyar [7,8], mind angol [9] nyelven. Jelen dolgozatban mély neurális hálókat alkalmazunk a nevetés felismerésére. A korábbi megközelítéseinkkel ellentétben, ahol az egész szegmensre illesztettünk GMM-et, majd a modell paraméterei alapján végeztük el a döntést, most kizárólag keretszintű felismerést végzünk; a szegmensszintű döntést a keretszintű valószínűségek szorzata alapján hozzuk meg. Korábbi módszerünk kézenfekvőnek tűnő adaptálása már csak azért sem járható út, mert a GMM-mel szemben egy DNN nehezen interpretálható, és nagyságrendekkel több paraméterrel (súllyal) is rendelkezik.

3. Mély neurális hálók

A beszédfelismerés területén, a lokális valószínűség-eloszlások modellezésére hagyományosan GMM-eket volt szokás használni, azonban ezeket a mély neurális hálók az utóbbi pár évben szinte teljesen kiszorították. Ennek oka, hogy a mély neurális hálók jóval nagyobb pontosságot képesek elérni ebben a feladatban (l. pl. [9,31]), miközben elérhetővé váltak azok a hardverek, melyekkel a mély neurális hálók viszonylag gyorsan betaníthatók és kiértékelhetők.

A hagyományos hálózatok esetében egy vagy maximum két rejtett réteget szoktunk csak használni, és a neuronok számának növelésével próbáljuk javítani az osztályozási pontosságot. Az utóbbi idők kísérleti eredményei azonban amel-



1. ábra. A szigmoid, tanh és rectifier aktivációs függvények

lett szólnak, hogy – adott neuronszám mellett – több réteg hatékonyabb reprezentációt tesz lehetővé [32]. Az ilyen sok rejtett réteges, „mély" architektúrának azonban nem triviális a betanítása. A hagyományos neuronhálók tanítására általában az ún. backpropagation algoritmust szokás használni, ez azonban kettőnél több rejtett réteg esetében egyre kevésbé hatékony. Ennek egyik oka, hogy egyre mélyebbre hatolva a gradiensek egyre kisebbek, egyre inkább eltűnnek („vanishing gradient"), ezért az alsóbb rétegek nem fognak kellőképpen tanulni [32].

A mély neurális hálók tanítására először Hinton et al. javasolt egy módszert [33]. Ebben az eljárásban a tanítás két lépésben történik: egy felügyelet nélküli előtanítást egy felügyelt finomhangolási lépés követ. A felügyelt tanításhoz használhatjuk a backpropagation algoritmust, az előtanításhoz azonban egy új módszer, a DBN előtanítás szükséges. Ez az inicializálási lépés, habár megnöveli a végül betanított neurális háló pontosságát, elég körülményessé és időigényessé teszi a tanítási folyamatot.

Ennek egy alternatívájaként javasolta Seide et al. a diszkriminatív előtanítást [31]. Ebben az első lépésben hagyományos módon egy egyetlen rejtett réteget tartalmazó neurális hálót tanítanak be. Ezután minden lépésben eldobják a kimeneti réteget a hozzá tartozó súlyokkal együtt, majd a hálót kiegészítik egy újabb rejtett réteggel és egy kimeneti réteggel. Az új kapcsolatokat véletlen súlyokkal inicializálják, és ezt a hálót tanítják a hagyományos módon (általában backpropagation eljárással). Ezeket a lépéseket ismétlik mindaddig, míg a tervezett számú rejtett réteget el nem érik.

A harmadik megoldás, mellyel a mély neurális hálók taníthatóvá válnak, nem egy új tanítási módszer, hanem a rejtett rétegek neuronjainak aktivációs függvényének lecserélése. A hagyományos szigmoid (vagy ennek skálázott változata, a tanh) függvény helyett az ún. *lineáris rectifier* függvényt alkalmazva (l. 1. ábra) az összes rejtett réteg taníthatóvá válik szimplán backpropagation módszerrel [34], ugyanakkor szükség van a súlyok regularizációjára (pl. L1 vagy L2 norma használatával).

Korábbi kísérleteink során mindhárom megközelítést teszteltük mind angol, mind magyar nyelvű szöveg fonémafelismerése során [7,35]. Tapasztalataink szerint a legjobb eredményt a rectifier aktivációs függvény használata hozta, mi-

közben ez bizonyult a leggyorsabbnak is, ezért kísérleteinkben ezt a fajta hálót fogjuk alkalmazni.

4. Kísérletek és eredmények

A jelen kutatásban a nevetést tartalmazó és azt nem tartalmazó keretek elkülönítésére használtunk mély neurális hálót; a kimeneti réteg két neuronja felelt meg ennek a két osztálynak. Kísérleteinket magyar spontán beszédből vett nevetés- és beszédrészleteken végeztük, melyek a BEszélt nyelvi Adatbázisból (BEA, [21,36]) lettek kiválogatva. A BEA különböző típusú spontán beszédet tartalmaz: narratíva, véleménykifejtés, három fős társalgás. A jelen kutatáshoz 75 adatközlő felvételét választottuk ki, átlagosan 16 percet egy beszélőtől. A hanganyagokban megjelenő összes nevetést manuálisan címkéztük fel a Praat programban. Ugyanezektől a személyektől az elhangzott nevetések számával nagyságrendileg megegyező mennyiségű beszédszegmenst választottunk ki. Összesen 331 nevetés- és 320 beszédszegmenst használtunk, melyekből 463-ra tanítottunk, és 188 szegmens került a teszthalmazba. A tanító adatbázis 240 nevetést és 223 beszédszegmenst tartalmazott, míg a tesztadatbázis 91 nevetést és 97 beszédszegmenst. Mivel az adatmennyiség nem túl nagy, külön fejlesztési halmaz definiálása helyett tízszeres keresztvalidációt (cross-validation, CV) alkalmaztunk.

Saját neurálisháló-implementációnkat használtuk, mellyel korábban sok különböző területen értünk el jó eredményeket (pl. [9,37,38,39]). A neurális hálókat keretszinten tanítottuk. A GMM-ek esetében bevettnek számító MFCC és PLP jellemzők mellett kipróbáltuk az FBANK jellemzőkészletet, amely 40 Mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból áll [40]. Alkalmaztuk azt a fonémaosztályozás esetén bevett megoldást is, hogy a szomszédos keretek jellemzővektorait is felhasználtuk az egyes keretek osztályozása során. Az alkalmazott neurális hálók előzetes tesztek eredményei alapján három rejtett réteggel rendelkeztek, melyek mindegyikében 256 rectifier függvényt alkalmazó neuron volt, míg a kimeneti rétegben softmax függvényt használtunk. A súlyokat L2 regularizációval tartottuk kordában. A keretszintű valószínűségbecsléseket szegmensenként és osztályonként (nevetés, ill. beszéd) összeszoroztuk, és a szegmenst abba az osztályba soroltuk, amelyre ez az érték magasabb volt.

Mivel a neurális háló tanítása sztochasztikus folyamat (köszönhetően a súlyok véletlen inicializálásának), a tízszeres keresztvalidáció minden esetére öt-öt hálót tanítottunk. Ezekből öt hálót értékeltünk ki a teszthalmazra, majd az egyes kiértékelésekre kapott eredményeket összegeztük.

4.1. Eredmények

Az egyes jellemzőkészletekkel, valamint a szomszédos keretek számával elért keretszintű pontosságértékeket a 1. táblázat, míg a szegmensszintűeket a 2. táblázat tartalmazza. Egy-egy jellemzőkészleten belül a legjobb értékeket (0, 2% tűréssel) kiemeltük.

1. táblázat. A különböző jellemzőkészletek és szomszédszámok használatával elért keretszintű pontosságértékek

Jellemző-készlet	N	Keresztvalidáció				Teszthalmaz			
		Pr.	Re.	F_1	Acc.	Pr.	Re.	F_1	Acc.
MFCC	1	72,3%	79,9%	75,9%	84,3%	49,8%	**89,3%**	63,9%	72,1%
	5	76,4%	81,6%	79,1%	86,6%	54,2%	**89,1%**	67,4%	76,1%
	9	78,8%	**82,3%**	**80,5%**	**87,7%**	64,1%	85,8%	**73,4%**	**82,8%**
	13	**79,3%**	**82,1%**	**80,7%**	**87,8%**	63,7%	84,1%	72,5%	82,3%
	17	**79,1%**	81,5%	80,3%	**87,6%**	**64,9%**	82,3%	72,6%	**82,8%**
PLP	1	76,2%	78,8%	77,5%	85,8%	52,4%	84,1%	64,6%	74,5%
	5	80,7%	81,3%	81,0%	**88,2%**	63,1%	**87,1%**	73,2%	82,3%
	9	80,6%	**81,9%**	**81,3%**	**88,3%**	**65,1%**	86,7%	**74,4%**	**83,4%**
	13	**81,5%**	79,8%	80,6%	**88,1%**	**65,3%**	85,4%	74,1%	**83,4%**
	17	**81,3%**	78,1%	79,7%	87,7%	62,8%	85,3%	72,3%	81,9%
FBANK	1	**99,1%**	**99,8%**	**99,5%**	**99,7%**	**97,6%**	**99,3%**	**98,4%**	**99,1%**
	5	98,7%	**99,6%**	99,2%	**99,5%**	95,4%	**99,3%**	97,3%	98,5%
	9	98,0%	99,5%	98,7%	99,2%	95,2%	**99,2%**	97,1%	98,4%
	13	97,7%	99,2%	98,4%	99,0%	94,5%	98,9%	96,6%	98,1%
	17	97,4%	99,0%	98,2%	98,9%	93,0%	98,8%	95,8%	97,6%

Látható, hogy míg MFCC jellemzőkészlet esetén 9–17 szomszédos kereten együtt tanítva kapjuk a legjobb keretszintű értékeket, PLP esetén ez 9–13 keret, ennél nagyobb szomszédságot használva már romlanak az eredmények, FBANK esetén pedig szomszédok nélkül tanítva kapjuk meg az optimumot. Az MFCC és a PLP használatával kapott értékek esetén megfigyelhető egy eltolódás a pontosság (precision) és a fedés között a keresztvalidációval, illetve a teszthalmazon kapott értékek között; ennek oka valószínűleg az, hogy a két halmazon belül a két osztály példáinak aránya nem ugyanaz.

A szegmensszintű pontosságok lényegesen magasabbak a keretszintűeknél, ami érthető, hiszen egy szegmens pontos besorolásához elég, ha a benne található keretek többségét jól osztályozzuk. MFCC esetén ezúttal is 9–13 szomszédos kereten tanítani tűnik az optimális választásnak, míg PLP használatával ez 5–9. Ezekben az esetekben is megfigyelhető az eltolódás a fedés és a pontosság (precision) között a keresztvalidációs és a teszthalmazon mért értékek között. Az FBANK jellemzőkészletnél 1, illetve 5 szomszédos keretre is tökéletes osztályozást kapunk, több szomszédot használva ez keresztvalidáció során egy picit romlik, de mindig 99% fölött marad.

Az FBANK jellemzőkészlet kiugró eredménye valószínűleg annak köszönhető, hogy az eredeti beszédhanghoz közelebb álló jellemzőket tartalmaz. Valószínűleg a nevetés felismeréséhez néhány kitüntetett frekvenciasáv vizsgálata fontos, és ezek sokkal jobban detektálhatóak ebben a szűrősorokból álló jellemzőkészletben, mint akár az MFCC-ben, akár a PLP-ben. Ehhez azonban neurális háló használatára van szükségünk, mivel GMM-et csak dekorrelált jellemzőkészletre lehet tanítani (a tesztelt jellemzők közül ilyen a MFCC és a PLP).

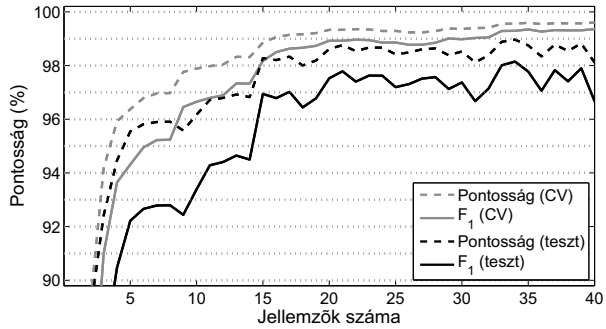
2. táblázat. A különböző jellemzőkészletek és szomszédszámok használatával elért szegmensszintű pontosságértékek

Jellemző-készlet	N	Keresztvalidáció				Teszthalmaz			
		Pr.	Re.	F_1	Acc.	Pr.	Re.	F_1	Acc.
MFCC	1	**98,8%**	85,2%	91,5%	92,4%	86,1%	94,9%	90,3%	89,5%
	5	**98,6%**	88,3%	93,1%	93,7%	90,4%	**96,7%**	93,4%	93,0%
	9	98,0%	89,1%	**93,3%**	**93,9%**	**97,5%**	95,5%	**96,5%**	**96,4%**
	13	97,4%	**90,0%**	**93,5%**	**94,0%**	96,2%	93,4%	94,8%	94,7%
	17	96,6%	88,1%	92,1%	92,7%	96,7%	89,3%	92,8%	92,9%
PLP	1	98,2%	85,1%	91,2%	92,1%	86,4%	94,2%	90,1%	89,4%
	5	**99,1%**	**88,3%**	**93,4%**	**94,0%**	93,8%	96,3%	95,0%	94,8%
	9	98,7%	**88,3%**	**93,2%**	**93,8%**	**94,4%**	**97,3%**	**95,8%**	**95,6%**
	13	98,7%	85,0%	91,3%	92,2%	**94,5%**	95,9%	95,2%	95,0%
	17	98,2%	81,9%	89,3%	90,5%	93,0%	93,2%	93,1%	92,9%
FBANK	1	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	5	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	9	**100,0%**	99,7%	**99,9%**	**99,9%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	13	99,6%	99,6%	99,6%	99,6%	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	17	99,4%	99,3%	99,3%	99,4%	**100,0%**	**100,0%**	**100,0%**	**100,0%**

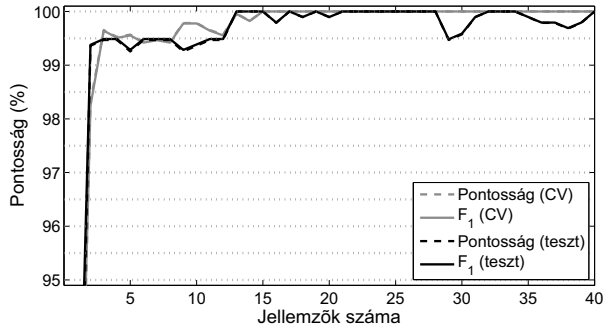
4.2. A felhasznált jellemzők vizsgálata

Tesztjeink során az FBANK jellemzőkészlettel meglepően jó eredményeket kaptunk. Ez a jellemzőkészlet az eredeti hanghoz közelebb álló, jobban interpretálható, mint akár az MFCC, akár a PLP. Az is nyilvánvaló, hogy az egyes frekvenciasávok nem ugyanolyan mértékben segítenek a mély neurális hálónak a nevetést tartalmazó keretek azonosításában. Ezért a következő részben kísérletileg rangsoroljuk, hogy az egyes sávok mennyire járulnak hozzá a mély neurális háló pontosságának eléréséhez.

A jellemzőkészletben 123 attribútum van; nyilvánvaló, hogy az összes lehetséges részhalmaz letesztelése aránytalanul sok időt emésztene fel. Ezért első lépésben sorba rendeztük a jellemzőket; a sok lehetséges mód közül mi ezt is a (már betanított) neurális hálók segítségével tettük meg. Egy neurális háló bemeneti rétegének neuronjai a jellemzőknek felelnek meg (amennyiben nem használjuk a szomszédos keretek jellemzőit), így az egyes neuronokból kimenő súlyok összessége (valamilyen mértékben) tükrözi az egyes jellemzők fontosságát. Mivel a súlyok valós számok, érdemes az egyes jellemzőkhöz tartozó súlyok négyzetösszegét venni: minél nagyobb ez az érték, annál fontosabbnak ítéli az adott neurális háló a jellemzőt. Ezután a jellemzőket ezen érték alapján csökkenő sorrendbe rendeztük, és mindig az első N db attribútum használatával tanítottunk neurális hálókat. (A kísérleti körülmények megegyeztek a 4. fejezetben bemutatottakkal.) Az elért pontosságértékek a 2. és 3. ábrán láthatóak. A keretszintű pontosságértékek már 15 jellemző használatával eljutnak $97-99\%$ közé, míg a tökéletes szegmensszintű osztályozáshoz 13 jellemző is elegendő. Ez igazolta azt a hipotézisünket, hogy csak néhány frekvenciasáv vizsgálata alapján is nagy pontossággal eldönt-



2. ábra. A mély neurális hálókkal elért keretszintű pontosságértékek a használt jellemzők számának függvényében

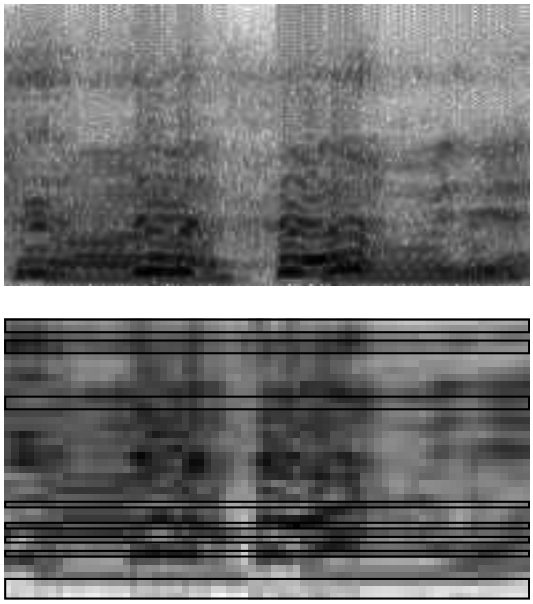


3. ábra. A mély neurális hálókkal elért szegmensszintű pontosságértékek a használt jellemzők számának függvényében.

hető, hogy az adott keret vagy szegmens nevetést tartalmaz-e. A kiválasztott frekvenciasáv-jellemzők a 4. ábrán láthatók. Az egyszerűség kedvéért nem ábrázoltuk az energiát, amely meglepő módon nem szerepelt a kiválasztott jellemzők között; illetve eltekintettünk az első- és másodrendű deriváltak megjelenítésétől is. A tizenöt jellemző közül egyébként tíz volt frekvenciasáv-szűrő, kettő első-, három pedig másodrendű derivált.

5. Összegzés

A jelen kutatásban mély neurális hálózatokat használtunk a nevetés keretszintű automatikus felismeréséhez. A kutatás során vizsgáltuk, hogy mely akusztikai jellemzők alkalmasabbak a nevetés azonosítására. Az akusztikai jellemzők közül a nevetésfelismerés szakirodalmában használt MFCC-t és PLP-t vettük górcső alá, illetve a mély neurális hálózatokhoz kiválóan alkalmas FBANK jellemzőt. Vizsgáltuk továbbá azt, hogy a keretszintű felismeréskor hány szomszédos kereten érdemes tanítani. A jellemzőkészlet mellett vizsgáltuk, hogy mely frekvenciasávok vesznek részt a mély neurális hálózatokkal történő nevetésfelismerésben.



4. ábra. Egy nevetésszegmens spektrogramja (fent), valamint az ebből kinyert BANK jellemzőkészlet és azon a 15 kiválasztott jellemzőhöz tartozó frekvenciatartományok (lent)

Ennek elemzéséhez a mély neurális hálózat kimeneti súlyait használtuk fel, illetve annak alapján rangsoroltuk az egyes frekvenciasávokat. Az eredmények azt mutatták, hogy a legjobb eredményt akkor kaptuk, ha az akusztikai jellemzők közül az FBANK reprezentációt használtuk a mély neurális hálózatok tanításához. A szomszédos keretek számának tekintetében az MFCC és a PLP használatakor érdemes nagyszámú szomszédos kereten tanítani, ugyanakkor az FBANK esetében szinte nincs is szükség szomszédos keretekre. A jellemzőválogatás során azt találtuk, hogy már 15 jellemző felhasználásakor is igen magas pontossági mutatót kapunk. Mindez azt bizonyítja, hogy az egyes frekvenciasávok nem egyenlő módon vesznek részt a nevetés azonosításában.

Összességében elmondható, hogy a mély neurális hálózatok jól alkalmazhatók a nevetések automatikus osztályozásában a korábbi kísérletekben használt GMM-SVM-mel összehasonlítva is. Az FBANK jellemzőkinyerés esetén a korábbi, ugyanezen korpuszra vonatkozó eredményeket felül is teljesíti. A kutatásunk egy újabb bizonyíték arra, hogy milyen kiválóan alkalmazható a mély neurális hálózat szinte bármely beszédtechnológiai kihívásban.

Hivatkozások

1. Truong, K.P., van Leeuwen, D.A.: Automatic detection of laughter. In: Interspeech, Lisszabon, Portugália (2005) 485–488

2. Kennedy, L.S., Ellis, D.P.W.: Laughter detection in meetings. In: Proceedings of the NIST Meeting Recognition Workshop at ICASSP, Montreal, Kanada (2004) 118–121
3. Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication **49**(2) (2007) 144–158
4. Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Interspeech, Lisszabon, Portugália (2005) 465–468
5. Nick, C.: On the use of nonverbal speech sounds in human communication. In Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M., eds.: Verbal and nonverbal communication behaviours. Springer-Verlag, Berlin, Heidelberg (2004) 117–128
6. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: Interspeech. (2007) 2973–2976
7. Grósz, T., Kovács, Gy., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY, Szeged, Magyarország (2014) 3–13
8. Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel. In: MSZNY, Szeged, Magyarország (2015) 174–181
9. Tóth, L.: Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP Journal on Audio, Speech, and Music Processing **2015**(25) (2015) 707–710
10. Holmes, J., Marra, M.: Having a laugh at work: How humour contributes to workplace culture. Journal of Pragmatics **34**(12) (2002) 1683–1710
11. Neuberger, T.: Nonverbális hangjelenségek a spontán beszédben. In Gósy, M., ed.: Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest (2012) 215–235
12. Glenn, P.: Laughter in interaction. Cambridge University Press, Cambridge, UK (2003)
13. Hámori, A.: Nevetés a társalgásban. In Laczkó, K., Tátrai, S., eds.: Elmélet és módszer. ELTE Eötvös József Collegium, Budapest (2014) 105–129
14. Günther, U.: What's in a laugh? Humour, jokes, and laughter in the conversational corpus of the BNC. PhD thesis, Universität Freiburg (2002)
15. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. Journal of the Acoustical Society of America **110**(3) (2001) 1581–1597
16. Neuberger, T., Beke, A.: Automatic laughter detection in spontaneous speech using GMM-SVM method. In: TSD, Pilsen, Csehország (2013) 113–120
17. Bickley, C., Hunnicutt, S.: Acoustic analysis of laughter. In: ICSLP, Banff, Kanada (1992) 927–930
18. Rothgänger, H., Hauser, G., Cappellini, A.C., Guidotti, A.: Analysis of laughter and speech sounds in Italian and German students. Naturwissenschaften **85**(8) (1998) 394–402
19. Trouvain, J.: Segmenting phonetic units in laughter. In: ICPhS, Barcelona, Spanyolország (2003) 2793–2796
20. Bóna, J.: Nonverbális hangjelenségek fiatalok és idősek spontán beszédében. Beszédkutatás **23**(8) (2015) 106–119
21. Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T.E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: Beszélt nyelvi adatbázis. In Gósy, M., ed.: Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest (2012) 9–24
22. Provine, R.R.: Laughter. American Scientist **84**(1) (1993) 38–45
23. Nwokah, E.E., Davies, P., Islam, A., Hsu, H.C., Fogel, A.: Vocal affect in three-year-olds: a quantitative acoustic analysis of child laughter. Journal of the Acoustical Society of America **94**(6) (1993) 3076–3090

24. Tahon, M., Devillers, L.: Laughter detection for on-line human-robot interaction. Cough **85**(65) (2015) 1–77
25. Petridis, S., Pantic, M.: Audiovisual discrimination between laughter and speech. In: ICASSP. (2008) 5117–5120
26. Petridis, S., Pantic, M.: Fusion of audio and visual cues for laughter detection. In: CIVR. (2008) 329–337
27. Reuderink, B., Poel, M., Truong, K., Poppe, R., Pantic, M.: Decision-level fusion for audio-visual laughter detection. In: MLMI, Utrecht, Hollandia (2008) 137–148
28. Petridis, S., Pantic, M.: Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: ICME. (2009) 1444–1447
29. Neuberger, T., Beke, A., Gósy, M.: Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech. In: ISSP, Köln, Németország (2014) 285–287
30. Ito, A., Wang, X., Suzuki, M., Makino, S.: Smile and laughter recognition using speech processing and face recognition from conversation video. In: CW. (2005) 437–444
31. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: ASRU. (2011) 24–29
32. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. (2010) 249–256
33. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
34. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: AISTATS. (2011) 315–323
35. Grósz, T., Tóth, L.: A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition. In: TSD, Pilsen, Csehország (2013) 36–43
36. Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative hungarian language. In: TSD2014. (2014) 424–431
37. Gosztolya, G.: On evaluation metrics for social signal detection. In: Interspeech, Drezda, Németország (2015) 2504–2508
38. Grósz, T., Busa-Fekete, R., Gosztolya, G., Tóth, L.: Assessing the degree of nativeness and Parkinson's condition using Gaussian Processes and Deep Rectifier Neural Networks. In: Interspeech, Drezda, Németország (2015) 1339–1343
39. Kovács, Gy., Tóth, L.: Joint optimization of spectro-temporal features and Deep Neural Nets for robust automatic speech recognition. Acta Cybernetica **22**(1) (2015) 117–134
40. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)

Magyar nyelvű szövegek automatikus fonetikai átírása

Novák Attila[1,2], Siklósi Borbála[2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport ,
[2] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
e-mail:{novak.attila, siklosi.borbala}@itk.ppke.hu

Kivonat Az írott szövegből hangzó beszédet előállító rendszerek egyik alapvető komponensének feladata a szöveg fonetikai átírása. Bár a betűsor–fonémasor-leképezés komplexitása nyelvenként változó, kivételek a legtöbb nyelvben vannak. Cikkünkben egy magyar szövegek fonetikai átírására alkalmas programot mutatunk be, amelyben a fonéma–grafémaátírószabályok implementálása mellett morfológiai elemző is szerepet kapott a morfémahatárok és szóösszetételek meghatározásához. Az így létrejött rendszer olyan szövegek átírása során is jól teljesít, melyek sok idegen szót tartalmaznak.

1. Bevezetés

A bemutatott rendszer célja írott szövegek automatikus és egyértelmű átírása azok fonetikus reprezentációjára. A rendszert egy magyar földrajzi neveket tartalmazó adatbázis átírására használtuk.

Bár az írott ábécé egységei általában megfeleltethetőek a beszélt nyelv fonetikai egységeinek, ennek a leképezésnek a bonyolultsága nyelvenként változó lehet. Ha csak a latin ábécét használó nyelveket vizsgáljuk ebből a szempontból, akkor is jelentős különbségeket találunk. Ezért egy fonetikai átírórendszer mindig nyelvspecifikus, és a különböző módszerek alkalmazhatósága az adott nyelv morfoszintaktikai és fonológiai tulajdonságaitól függ.

Az angol helyesírás szabályait viszonylag korán rögzítették, azonban a kiejtési rendszer tovább fejlődött [9]. Ezért gyakran elég nehéz megjósolni az írott és kiejtett alakok közötti megfeleltetést. Mivel azonban a szóalakok száma véges, ezért akár kézzel, akár géppel készített szótár alkalmas lehet arra, hogy az angol nyelv összes szavát – azok írott és a kiejtés szerint átírt alakjaival – tároljuk. Problémát csak az új szavak és nevek megjelenése jelenthet.

Más nyelvek, például a magyar esetében az írott és kiejtett alakok általában sokkal közelebb állnak egymáshoz. A legtöbb esetben a kiejtés egyértelműen megjósolható az írott alak alapján. Mégis adódnak kivételes esetek és fonetikai korlátokból fakadó megkötések. Továbbá a magyar nyelv agglutináló jellegéből adódóan a szóalakok nem felsorolható volta miatt azok szótárban való tárolása nem lehetséges [8].

Ezért, olyan automatikus módszerre van szükség a fonetikai átirat meghatározására, ami ráadásul figyelembe veszi a technikai követelményeket, azaz a nagy mennyiségű offline adat tárolása helyett a feldolgozásra helyezi a hangsúlyt. Célunk egy ilyen rendszer megvalósítása volt.

2. Kapcsolódó munkák

A fonetikus átírást megvalósító módszerek három fő csoportba sorolhatók: [4]:

- szótárban való megfeleltetés,
- szabályalapú módszerek,
- adatvezérelt módszerek.

A szótárban való megfeleltetést akkor alkalmazhatjuk, ha az írott és a fonetikus reprezentáció csak konvenciókon alapul, nincsenek alkalmazható szabályszerűségek vagy általánosítások. Ezeknek a módszereknek az az előnye, hogy a szótárban kiegészítő információkat (hangsúly, szófaj) is tárolhatunk. Hátrányuk viszont, hogy a szótárak kézzel való létrehozása nagyon költséges és nehéz feladat.

Agglutináló nyelvek esetén azonban, függetlenül az agglutináció korlátozottságától, mindig vannak olyan szóalakok, amik egy előre létrehozott szótárból hiányoznak. A szabályalapú módszerek ezt a problémát előre meghatározott átírószabályok alkalmazásával oldják meg. Ezeket a nyelvspecifikus szabályokat nyelvészek definiálják, majd valamilyen keretrendszerben (pl. véges fordítóautomaták [7]) formalizálva alkalmazhatóak. Az ilyen szabályalapú megoldásoknak is szükségük van egy kivételszótárra, amiben a rendhagyó kiejtésű szóalakok átiratát rögzítik.

Gépi tanulási módszerek alkalmazására is láthatunk példát fonetikai átírás megoldására. [5] szerzői megmutatták, hogy az ilyen módszerek általánosító képessége jobb, mint a szabályalapú rendszereké (az angol nyelv esetén). Az egyik legismertebb ilyen implementáció a *Pronunciation by Analogy* (PbA), azaz a hasonló szavak átiratán alapuló átírás elvén működik [6]. Emögött az elmélet mögött az a pszicholingvisztikai modell áll, ami egy szó kiejtését a hozzá hasonló alakú, ismert kiejtésű szavak alapján határozza meg. Egy másik megközelítés, az ún. *Joint-sequence model* egy írott szóalak legvalószínűbb kiejtését a Bayes-szabály alapján határozza meg. Minden ilyen adatvezérelt megközelítés esetén szükség van azonban egy szótárra, vagy egy átírt korpuszra amiből a statisztikák meghatározhatók, illetve a gépi tanulási módszerek számára tanítóanyagként szolgál.

Magyar nyelvre létezik egy online kiejtésszótár, ami 1,5 millió szóalakot és azok fonetikai átiratát tartalmazza [2]. A szótár létrehozása több lépésből állt. Először, egy nagy írott korpusz szavait gyűjtötték össze és ebből a listából eltávolították az idegen, illetve a helyesírási hibát tartalmazó szavakat. Ezután átírószabályokat alkalmaztak. Végül, a kivételes esetek meghatározása után ezeket kézzel javították. A szótár szerzői szerint így a szótár referenciaként használható, hiszen ez magyar nyelvre a legnagyobb, IPA átiratot tartalmazó és helyesnek tekinthető szótár. Azonban, a szótár korlátja, hogy csak az eredeti korpuszban

szereplő szavakat ismeri, más szóalakok vagy új szavak átírására nem alkalmazható. Emellett az online felhasználói felületben implementált korlátozások miatt tömeges lekérdezésre nem használható.

3. Módszer

Fonémikus nyelvek esetén (pl. finn, észt, magyar, stb.) az írott alakok kiejtés szerinti átírása majdhogynem egyértelmű. Például az *ablak* szó kiejtése [ɔblɔk]. (Az 1. táblázat tartalmazza az egyes magyar betűkhöz tartozó kiejtés Nemzetközi Fonetikai Ábécé szerinti átírását.) Van azonban két olyan jelenség, ami az átírást megnehezíti: bizonyos hangkapcsolódások, illetve a régies, vagy idegen szavak. További problémát jelent a nyelv szemiotikus rendszerének (számok, rövidítések) normalizálása.

1. táblázat. A magyar betűknek megfelelő fonémák IPA jelöléssel

betű	**IPA**	**betű**	**IPA**	**betű**	**IPA**	**betű**	**IPA**
á	aː	b	b	n	n	zs	ʒ
a	ɔ	p	p	ny	ɲ	s	ʃ
o	o	d	d	j	j	cs	t͡ʃ
u	u	t	t	h	h	l	l
ü	y	g	g	v	v	r	r
i	i	k	k	f	f	dz	d͡z
é	eː	gy	ɟ	z	z	dzs	d͡ʒ
ö	ø	ty	c	sz	s		
e	ɛ	m	m	c	t͡s		

Rendszerünk három komponensből áll: (1) a Humor morfológiai elemzőből [10,11], (2) egy a rendhagyó szótöveket tartalmazó lexikonból és (3) a fonológiai szabályok XFST-ben (Xerox Finite-State Tool) [3] való implementációjából.

3.1. Morfológiai elemzés

A feldolgozás során először a szóalakok morfológiai szerkezetét határozzuk meg, hogy a megfelelő morfémahatárokon érvénybe lépő morfofonológiai szabályok érvényességi körét megállapítsuk (pl. lexikai palatalizáció, hosszú mássalhangzók vs. szóösszetételek, stb.). Továbbá, a magyarban igen gyakori összetett szavak esetén is előfordulhatnak rendhagyó elemek, amelyek felismeréséhez az összetételi határoknál történő szegmentálásra van szükség. Ráadásul, több fonémát bigráfokkal ír le a magyar nyelv (*cs, gy, ty, ny, sz, zs, dz, dzs*, és ezek hosszú alakjai). De ha morfémahatáron szerepelnek egymás mellett ezek a mássalhangzók, akkor külön ejtjük őket (más szabályok miatt előfordulhat részleges-, vagy teljes hasonulás). Például az *eszközsáv* helyes átirata [ɛskøsʃaːv], nem pedig [ɛskøʒaːv].

Lehetnek továbbá olyan összetett szavak, amelyeknek az egyes összetevői rendhagyó kiejtésűek. Ezek felismeréséhez is a morfológiai elemzőre van szükség annak elkerülése érdekében, hogy az általános fonológiai szabályok szerint írjuk át őket.

3.2. Rendhagyó szótövek

Minden nyelvben vannak rendhagyó kiejtésű szavak. Ezek általában tulajdonnevek és idegen szavak, amik azonban valamilyen mértékben idomulnak a befogadó nyelv fonológiai rendszeréhez. Ilyen példa az angol *file* szó, ami a magyar helyesírás szerint használható az eredeti formájában is, de a kiejtéshez alkalmazkodó formában is írható (*fájl*). Mindkét esetben a [faːjl] a szó fonetikai átirata. Ezzel szemben a *New York* kifejezést csak ebben az alakban írhatjuk, kiejtése pedig [ɲuːjork]. Az idegen szavak mellett a hagyomány szerinti írásmóddal rendelkező szavak is hasonlóan viselkednek, így sok családnév, földrajzi név is ebbe a kategóriába tartozik.

További rendhagyó esetek az olyan szóalakok, amelyekben a helyesírási norma nem a köznyelvi kiejtésnek megfelelően jelöli valamely hang hosszúságát (pl. az *egyesület* szóalak kiejtésének az írott alak szerinti [ɛɟɛʃylet] helyett az [ɛɟːɛʃylet] felel meg).

Szintén a lexikonban vannak felsorolva az olyan rövidített, illetve különböző szimbólumokat tartalmazó kifejezések, melyek kiejtése nem az írott alak megfelelője (pl. számok, képletek, dátumok, mértékegységek, stb.). Ide sorolható továbbá a rövidítések különböző típusainak kezelése is. Ahhoz, hogy ezekhez a szóalakokhoz a kiejtés hozzárendelhető legyen, ezeket egy előfeldolgozási lépésben normalizálni kell. A nyelv szemiotikus rendszerének normalizálása számos alfeladatot tartalmaz, ezek részletes tárgyalása megtalálható [13]-ban. A rövidítések kérdéséről viszont itt is ejtünk néhány szót. A rövidített alakok kiejtése során több esettel állhatunk szemben:

- a rövidítést úgy ejtjük ki, mintha egy valódi (itt: idegen) szó lenne (pl. *NATO* [naːtoː]),
- a rövidített alak helyére szóban annak kifejtett változatát helyettesítjük be (pl. *du.* [deːlutaːn]),
- a rövidített alakot betűzve ejtjük ki (pl. *USB* [uːɛʃbeː]).

Rendszerünkben a rövidítésként felismert alakokat először egy olyan listával egyeztetjük, amiben a szóként kiejtett rövidített alakok vannak felsorolva. Ha ebben nem szerepel, akkor az az alapértelmezett szabály érvényesül, hogy betűzve ejtjük ki a rövidítést.

3.3. Fonológiai szabályok

A harmadik komponens a helyesírás által nem jelölt fonológiai szabályok XFST-beli implementációja. A szabályok [12] leírását követik. A szabályok alkalmazásának sorrendjét több tényező határozza meg (a teljes sorrendet l. a 2. táblázatban). A mássalhangzók helyesírási sajátosságainak kezelése megelőzi a többi

szabályt. A lexikai szabályokat a posztlexikális folyamatok végrehajtása előtt kell alkalmazni. Ezen kívül számos további jelenség egyedi kezelésére volt szükség. Ez szintén hatással volt a szabályok alkalmazási sorrendjére, ezeket lejjebb részletezzük.

2. táblázat. Fonológiai szabályok az alkalmazásuk sorrendjében

#	**szabály**
1.	hosszú digráfok átírása, x, w, qu, y, ly
2.	lexikai h-törlés
3.	lexikai palatalizáció
4.	lexikai palatális összeolvadás (a lex. palatalizáció megelőzi)
5.	a több szótagú tövek végén álló magas magánhangzók rövidülése (opcionális)
6.	az intervokalikus és szóvégi *dzs* és *dz* nyúlása
7.	minden szó első szótagja hangsúlyt kap
8.	zöngésségi hasonulás (regresszív, a jobb kontextust a kimeneten kell ellenőrizni)
9.	adaffrikáció (a zöngésségi hasonulás megelőzi)
10.	nazális hasonulás
11.	degemináció
12.	*j*: a fon. frázis végén: zöngétlen réshang zöngétlen obstruensek után; réshang lesz zöngés mássalhangzók után a fon. frázis végén
13.	a *h* posztlexikális váltakozása (szonoránsok után zöngésedik; kódában palatalizáció és velarizáció)
14.	posztlexikális palatalizáció
15.	zár- és réshangok, nazálisok, likvidák: gemináció minden határon keresztül
16.	affrikáták: gemináció csak a toldalékhatárokon
17.	magánhangzók átírása

Helyesírási sajátosságok kezelése

1. Ez a szabály egyrészt a digráffal jelölt palatális és szibiláns mássalhangzók (pl. *ty, sz, zs* stb.) gemináta (hosszú) alakjának kezelésére (pl. *tty, ssz, zzs* stb.) szolgál. Az ilyen alakú betűkapcsolatok valódi mássalhangzókapcsolatokat is jelölhetnek, pl. a *ssz* betűsorozat *s+sz* [Ss] kapcsolat is lehet,

ez azonban csak akkor lehetséges, ha a két mássalhangzó között morfémahatár húzódik. Emellett a szabály a csak idegen szavakban és nevekben előforduló *q, w, x* és a *ty, gy, ny, ly* betűkapcsolatokon kívüli *y* betűk (részben kontextusfüggő) kiejtését adja meg.

Lexikális folyamatok

2. A *h*-végű szavak egy részénél (pl. a *méh* szó esetében) a tővégi *h*-t általában nem ejtjük, ha nem követi magánhangzó kezdetű toldalék.
3. Az inflexiós toldalékok elején álló *j* palatalizálja az előtte álló tővégi zárhangokat és a *l*-et. Ez a lexikális palatalizációs szabály csak inflexióstoldalékhatárokon működik.
4. Az inflexiós toldalékok elején álló *j* palatális geminátává olvad össze az előtte álló tővégi palatális mássalhangzókkal. A lexikai palatalizáció megelőzi (táplálja) ezt a szabályt.
5. A hosszú felső nyelvállású magánhangzóra (*í, ú, ű*) végződő több szótagú töveket a köznyelvben általában rövid tővégi magánhangzóval ejtjük, kivéve a rendkívül igényes „színpadi" kiejtést. Ezt az opcionális rövidülést implementáltuk a rendszerben.
6. Az intervokális és a szóvégi *dzs* és *dz* ejtése hosszú. Ez alól csak néhány lexikai kivétel van, pl. a *fridzsider* [fridʒider] szó.

Hangsúly

7. A hangsúlykijelölés meglehetősen egyszerű a magyarban: a hangsúly mindig az első szótagra esik. Kizárólag a hangsúlytalan elemek: a determinánsok és klitikumok, illetve a tágabb tagmondat- vagy fráziskontextusban fellépő hangsúlytörlések okoznak problémát, ezeket azonban az xfst-szabályrendszeren kívül kezeltük.

Posztlexikális szabályok

8. A magyarban regresszív (jobbról balra ható) zöngésségi hasonulás érinti az obstruenseket. Két kivételes hang van ebből a szempontból: a *v* zöngétlenedik, de nem zöngétlenít, a *h* pedig zöngétlenít, de nem zöngésedik. Az a folyamat megelőzi az adaffrikációt.
9. Adaffrikáció: bizonyos zárhang-réshang és zárhang-affrikáta kapcsolatok a megfelelő affrikátává olvadnak össze. Nem implementáltuk a szabályrendszerben azokat az adaffrikációs jelenségeket, amelyek csak lezser, gyors beszédben érvényesülnek, mint a szóhatárokon átívelő zárhang-réshang adaffrikáció vagy a palatális-zárhang, illetve palatális-affrikáta adaffrikáció.
10. A nazális *n* képzési helye hasonul az azt követő zárhang vagy nazális képzési helyéhez, illetve az *n* és *m* labiodentális nazálisként [ɱ] realizálódik, ha labiodentális frikatíva követi.
11. Számos degeminációs folyamat van, amelyek különböző környezetekben játszódnak le. A monomorfemikus gemináták bármely más mássalhangzó környezetében degeminálódnak (rövidülnek): CC-X → C-X, X-CC → X-C (ahol

a - bármilyen morfémahatár lehet, illetve nincs is szükség morfémahatár jelenlétére). A morfémahatáron keresztül történő degemináció XC-C → X-C, C-CX → C-X csak akkor kötelező, ha az X obstruens (és nazális környezetben is implementáltuk a folyamatot, mert ebben a környezetben is gyakori). A C-CX → C-X degemináció csak az obstruensek egy részhalmazát érinti. A likvidákat követő LC=C → L=C degemináció csak inflexióstoldalék-határokon következik be.

12. A szóvégi *j* zöngétlen [ç], illetve zöngés frikatívaként [j] realizálódik, ha zöngés, illetve zöngétlen mássalhangzót követ.
13. A *h* posztlexikális váltakozást mutat. Intervokális, illetve magánhangzó és szonoráns közötti helyzetben zöngésedik. Elöl képzett magánhangzót követő kódában [ç]-vé palatalizálódik, egyéb esetben kódában [x]-vá velarizálódik.
14. Posztlexikális palatalizáció: a dentális *t, d, n* palatalizálódik a palatális *ty, gy, ny* előtt.
15. A zár- és réshangok, a nazálisok és a likvidák minden morfémahatáron keresztül geminálódnak.
16. A nem túl lezser beszédben az affrikáták csak a toldalékhatárokon geminálódnak.
17. Végül a hosszú magánhangzók reprezentációját is a V**:** jelölésre konvertáljuk.

4. Kiértékelés

A rendszer kiértékeléséhez George Orwell 1984 című regényét használtuk. A sok egyedi és idegen szóösszetétel és szóalak miatt esett a választás erre a szövegre. A rendszerünk által létrehozott átiratot a több nyelvre is elérhető eSpeak rendszer [1] magyar implementációjával létrehozott átirattal hasonlítottuk össze. Az eSpeak használható úgy, hogy kimeneteként IPA átiratot kapjunk. Megjegyzendő, hogy az online elérhető[3] magyar elektronikus kiejtési szótár [2] használata a kiértékelés során felmerült, azonban a hozzáférési felület nem volt alkalmas arra, hogy automatikus kiértékelésre használjuk. Ebben a szótárban 1,5 millió szóalak található, ragozott alakokkal együtt, ami jól reprezentálja a magyar nyelvet és 99%-ban helyes. Azonban az adatbázis nem letölthető és a leírásában szereplő első 1000 találatot elmentő funkció sem elérhető, ezért nem tudtuk felhasználni.

A valódi összehasonlításhoz az eSpeak rendszert is kiegészítettük, mert a posztlexikális hasonulások (zöngésségi hasonulás, palatalizáció, nazális hasonulás, a /h/ és /j/ hasonulása) nincsenek jelölve az eSpeak kimenetén, ezen kívül nem különbözteti meg az affrikátákat a zárhang-réshang kapcsolatoktól (pl. /t͡ʃ/ vs. /tʃ/), illetve a gemináta mássalhangzók reprezentációja gyakran hibás (pl. /t:/ helyett /tt/). Ezeket a hibákat a kimenet utófeldolgozásával javítottuk, hogy helyes és a saját rendszerünkkel összehasonlítható átiratok keletkezzenek. Egy másik eltérés a két rendszer között a tővégi hosszú felső nyelvállású magánhangzók opcionális rövidülése, amit a saját rendszerünkben implementáltunk, az eSpeak viszont ezeket a kicsit mesterkélten finomkodó hosszú alakjukban írja át. Ezeknek a hangoknak a rövid kiejtése jellemzőbb a köznyelvi magyarban.

[3] http://beszedmuhely.tmit.bme.hu/mksz/

A kiértékelés során hibaarányt mértünk (szó szinten) a teljes korpuszon vizsgálva. Azokban az esetekben, ahol több helyes átirat is helyes, bármelyik változatot elfogadtuk. A két rendszer eredményeit a 3. táblázat tartalmazza. Az eredeti eSpeak rendszer szóhibaaránya (WER) 14,81%, a kiegészített eSpeak hibaaránya 0,98%, míg a saját rendszerünk által elért hibaarány csupán 0,35% volt. Látható tehát, hogy a tesztszövegben előforduló idegen, illetve kitalált szavak sem okoztak gondot az átíróprogram számára.

3. táblázat. Kiértékelés. u/i: a rövidülő tővégi hosszú felső nyelvállású magánhangzókat tartalmazó szavak aránya; hason/h/j/N/zöng: azon szavak aránya, amelyekben a zöngésségi/palatális/nazális/j/h-hasonulás hibásan nincs jelölve, de ettől eltekintve helyesek; WER: a maradék szóhiba-arány.

rendszer	**WER**
a mi rendszerünk WER	0,35%
eSpeak u/i	0,98%
eSpeak WER	2,26%
eSpeak hason/h/j/N/zöng	14,81%

Az eSpeak kimenetében tapasztalt korábban nem említett hibák elsősorban a következő okokra vezethetők vissza: Lexikai hiányok (ide értve a szövegben szereplő számos angol név kiejtését), gyakori rövidítések nem megfelelő feloldása, a gemináta /r/ és a *ch* digráf kiejtésével kapcsolatos hibák, néhány szó kiejtésének ábrázolásával kapcsolatos idioszinkratikus hibák, és a lexikai palatalizáció túlalkalmazása olyan helyeken, ahol nem lenne szabad megjelennie. Az utóbi hibát a morfológiai elemzés hiánya okozza: a lexikai palatalizációt mintaillesztéses módszerrel kezelik az eSpeakben, és a minta ott is illeszkedik, ahol nem kéne.

Az általunk implementált rendszernek sokkal jobban megy az angol nevek kiejtése, hibáit elsősorban (az eSpeakétől különböző) lexikai hiányok, egyes rövidítések hibás feloldása, és egyes álösszetételek túlelemzése okozza. Az Orwell által kreált az 1984-ben szereplő újbeszél szavak egyik rendszernek sem okoztak komoly fejtörést, mert a kiejtésük szabályos, és mindkét rendszer algoritmikus átírókomponenst tartalmaz ahelyett, hogy pusztán szótárra támaszkodna.

5. Konklúzió

Bemutattunk egy magyar nyelvű szövegek automatikus fonetikai átírására alkalmas automatikus eszközt. A rendszer nem csak egyes szavakat képez le azok egy szótárban található átiratára, hanem teljes mondatok átírására is alkalmas, mivel figyelembe veszi a szóhatárokon előforduló hasonulásokat. Ezt egy, a morféma-, és

összetételi határok meghatározására képes morfológiai elemző és fonetikai átírószabályok alkalmazásával valósítottuk meg. Bemutattuk továbbá, hogy nagyméretű lexikon nélkül is jó minőségű fonetikai átírás állítható elő, hiszen a rendszer nem korlátozódik egy előre létrehozott lexikonban eltárolt szavak kezelésére. Ez a funkció egy olyan nyelv esetén, mint a magyar, ahol újabb és újabb szóalakok fordulhatnak elő, kiemelkedő fontosságú. Megmutattuk, hogy egy sok idegen szót tartalmazó korpuszon való kiértékelés során a rendszerünk jóval alacsonyabb hibaaránnyal teljesít, mint egy kereskedelmi eszköz, aminek a kimenetét ráadásul sokkal kevésbé szigorúan kezeltük.

Hivatkozások

1. eSpeak. http://espeak.sourceforge.net/, accessed: 2015-04-10
2. Abari, K., Olaszy, G., Zainkó, Cs., Kiss, G.: Magyar kiejtési szótár az interneten. In: IV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 223–230. SZTE, Szeged (2006)
3. Beesley, K., Karttunen, L.: Finite State Morphology. No. 1 in CSLI studies in computational linguistics: Center for the Study of Language and Information, CSLI Publications (2003), http://books.google.hu/books?id=59RoAAAAIAAJ
4. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. 50(5), 434–451 (May 2008), http://dx.doi.org/10.1016/j.specom.2008.01.002
5. Damper, R., Marchand, Y., Adamson, M., Gustafson, K.: Evaluating the pronunciation component of text-to-speech systems for english: a performance comparison of different approaches. Computer Speech and Language 13(2), 155 – 176 (1999), http://www.sciencedirect.com/science/article/pii/S0885230898901176
6. Dedina, M.J., Nusbaum, H.C.: Pronounce: a program for pronunciation by analogy. Computer Speech and Language 5(1), 55 – 64 (1991), http://www.sciencedirect.com/science/article/pii/088523089190017K
7. Kaplan, R.M., Kay, M.: Regular models of phonological rule systems. Comput. Linguist. 20(3), 331–378 (Sep 1994), http://dl.acm.org/citation.cfm?id=204915.204917
8. Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Saraclar, M.: Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 487–494. HLT-NAACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dx.doi.org/10.3115/1220835.1220897
9. Németh, G., Olaszy, G.: A magyar beszéd. Akadémiai Kiadó, Budapest, Hungary (2010)
10. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
11. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)

12. Siptár, P.: A magánhangzók. In: Kiefer, F., Bánréti, Z., Ács, P. (eds.) Fonológia. No. 2 in Strukturális magyar nyelvtan, Akadémiai Kiadó (1994), http://books.google.hu/books?id=j6xiAAAAMAAJ
13. Taylor, P.A.: Text-to-speech synthesis. Cambridge University Press, Cambridge, UK, New York (2009), http://opac.inria.fr/record=b1129276

Gépi beszéd természetességének növelése automatikus, beszédjel alapú hangsúlycímkéző algoritmussal

Szaszák György[1], Beke András[2], Olaszy Gábor[1], Tóth Bálint Pál[1]

1 Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{szaszak,olaszy,toth.b}@tmit.bme.hu
2 MTA Nyelvtudományi Intézet, Fonetikai Osztály

Kivonat A minél természetesebb hangzás elérése a géppel előállított beszédben napjainkban is igen fontos kutatási terület. A hangzás természetességét számos más tényező mellett a prozódia is nagyban befolyásolja, ezért alapvető követelmény egy olyan, precízen annotált korpusz megléte, amely alapján gépi tanulással pontos generatív modelleket állíthatunk elő. A korpusz kézi címkézése költséges és hosszadalmas, még a prozódiai egységekre, hangsúlyokra vonatkozóan is, ráadásul nemzetközi tapasztalatok is igazolják, hogy a szakértő címkézők ítélete is szubjektív, hiszen a különböző szakértők által előállított hangsúlyozásra vonatkozó annotációk közötti átfedés ritkán haladja meg a 80%-ot. A fentiek miatt gyakran használnak automatikus címkéző eljárásokat. A hangsúlycímkézést leggyakrabban a szöveges átirat alapján végzik el, ami azonban szerényebb pontosságot szolgáltat az emberi annotáláshoz képest. Alternatívaként jelen munkában egy beszédjel alapú hangsúlycímkéző algoritmust valósítunk meg. Az így nyert hangsúlycímkézés ellenőrzésére hat (3-3 férfi és női) HMM-TTS rendszert tanítunk, majd szubjektív lehallgatási tesztekkel (CMOS) hasonlítjuk össze a rendszereket.

Kulcsszavak: gépi beszédfelismerés, nyelvi elemzés, információkinyerés

1. Bevezetés

A gépi beszédelőállítás célját szolgáló beszédkorpuszok tervezése, rögzítése, és különösen precíz címkézése fontos feladat, amely a szöveg-beszéd átalakítás (Text-to-Speech, TTS) minőségét is alapvetően meghatározza. A címkézést kézzel vagy automatikusan végezhetjük. A kézi címkézés általában pontos, de nagyon időigényes, és nem küszöbölhető ki maradéktalanul a szubjektivitás sem. Szakértő címkézők által készített prozódiai annotációban például 70 és 80% között találták az alapfrekvencia-változások jelölésének egyezőségét egy angol nyelvű korpusz ToBI szerinti annotációjában [1]. Saját tapasztalataink is azt támasztják alá, hogy a humán címkéző nem tud a jelentéstől elvonatkoztatni, és lehallgatás alapú címkézés során percepciójában nem tudja például elkülöníteni az akusztikailag (pl. alapfrekvencia-csúcs), illetve a nyelvileg (szintaxis és szemantika) jelölt hangsúlyokat, amelyek az emberben gyakran egységes hangsúlyérzetként jelentkeznek.

Emellett korábbi kísérleti eredmények is arra utalnak, hogy ha a hangsúly a szintaxisból következik, akkor annak az akusztikai megjelölése elmaradhat [2]. A korpuszok címkézésekor jó lenne, ha szelektíven, kizárólag az akusztikai evidencia alapján tudnánk megjelölni, hol található olyan marker, amely a hangsúlyozással kapcsolatba hozható.

A kézi hangsúlycímkézés alternatívája az automatikus módozat, amelyet tipikusan a beszéd szöveges átiratán végzett szövegelemzés alapján végeznek szabály alapon vagy esetleg adatvezérelten. Az automatikus eljárások sem mentesek azonban a hibáktól, ami ismét az akusztikailag és nyelvileg jelölt hangsúlyok különbözőségéből, valamint az egyéni variabilitásból, vagy szövegen felüli kommunikációs szándékból fakad. A szabályalapú megközelítések egyelőre elterjedtebbek, pedig az általánosítóképességük korlátai miatt eleve nem hibátlan a szintaktikailag jelzett hangsúlyos pozíciók azonosítása sem. Ez utóbbi kivételkezeléssel javítható, de a szintaktikai és az akusztikai jelzések közötti különbségek ily módon nem kezelhetők.

Cikkünkben egy akusztikai elemzésen alapuló automatikus hangsúlycímkéző eljárást mutatunk be és értékelünk ki. Meglátásunk szerint a gépi szövegfelolvasáshoz az akusztikailag jelzett hangsúlyok jelölése a fontos a tanítókorpuszban, a szövegszinten kikövetkeztethető, de legalábbis percepciósan megjelenő „hangsúlyokat" a természetes beszédben sem jelezzük külön. A nemzetközi irodalomban számos hasonló kísérletről számoltak be [3], de ezek tipikusan a ToBI címkézés automatikus elkészítésére vonatkoztak [4]. Az eljárások közös pontja, hogy szegmentális, legfeljebb szótagszintű elemzésre támaszkodnak, de a szupraszegmentális vetületet korlátozottan képesek figyelembe venni. Bár a hangsúly valóban leginkább a szótaghoz köthető, véleményünk szerint hatékonyabb a szupraszegmentális oldalról, felülről lefelé haladva megközelíteni (vö. napjaink leginkább elfogadott beszédprodukciós modelljével [5], amelyben a végső prozódiai struktúra felülről lefelé egyre finomodik a mélyebb szintek hozzáadódó befolyása révén).

A bemutatásra kerülő beszédjel alapú hangsúlycímkéző eljárás fonológiai frázisok automatikus felismerésén alapul [6], ennek hátteréről korábban az MSzNy konferenciákon is részletesen beszámoltunk [7]. Mivel a fonológiai frázis definíció szerint egyetlen hangsúlyos szótagot tartalmaz (magyarban ez az első szótagon kötött hangsúly miatt a fonológiai frázis legelső szótagja), az eljárással automatikus hangsúlycímkézés valósítható meg. A hangsúlycímkézés többszintűvé is tehető, mivel a detektálni kívánt fonológiai frázisok egyes típusai között is éppen a hangsúly jellege, erőssége az egyik elkülönítő kritérium (az intonációs kontúr mellett).

Cikkünk felépítése az alábbiak szerint alakul: elsőként bemutatjuk a szöveg, és a beszéd alapján végzett automatikus hangsúlycímkézési eljárásokat. A címkézés nélküli, valamint a két különféle eljárással címkézett korpuszokon egy-egy TTS rendszert tanítunk férfi és női hangra is, amelyeket szubjektív lehallgatási tesztekkel hasonlítunk össze.

2. Automatikus hangsúlycímkézés a szöveg alapján

A szövegből történő hangsúlycímkézés szabályalapon történik, amelyeket kivétellisták egészítenek ki. A Profivox TTS rendszerben alkalmazott hangsúlycímkézés (és -generálás) teljes körű leírása a [8] irodalomban található, ehelyütt ennek egy rövid áttekintésére szorítkozunk. A szöveg alapú hangsúlycímkézés négy szintet különböztet meg:

- Nagyon erős hangsúly: általában valamilyen kontrasztivitásban, tagadásban jelenik meg, lista alapján határozzuk meg;
- Erős hangsúly: szintén szólista alapján határozza meg az algoritmus;
- Hangsúlyos: szövegszintű szabályok alapján adódik;
- Hangsúlytalan: a fennmaradó, vagy az irtó szabály miatt hangsúlytalanná vált szótagokon.

Ezen belül a szabályok elsősorban a hangsúlyos szótagok meghatározásában működnek közre. A főbb szabályok az alábbiak:

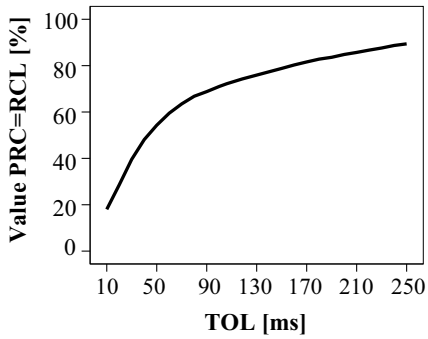
- A mondatkezdő szavak hangsúlyosak;
- Névelő és az *és* kötőszó után álló szavak hangsúlyosak;
- Vessző után hangsúlyos szó következik (figyelembe véve egy erre a célra kialakított kivétellistát);
- A mondat utolsó szava sosem hangsúlyos;
- Névelők és erősen hangsúlyos szó után álló szavak sosem hangsúlyosak.

A Profivox TTS applikáció jelenleg használt változatában háromszintű hangsúlymodellezés van: erősen hangsúlyos (nagyon erős és erős hangsúly összevontan), hangsúlyos és hangsúlytalan szótagcímkéket használunk. Cikkünk hátralévő részében a szöveg alapú hangsúlycímkézésre angol elnevezése után a **TBSM** (Text Based Stress Modelling) rövidítéssel utalunk.

3. Automatikus hangsúlycímkézés a beszédjel alapján

Az automatikus hangsúlycímkézés fonológiai frázisok detektálásán alapul. A fonológiai frázisokat prozódiai jellemzők alapján Viterbi-algoritmussal illesztünk a beszédjelre. A fonológiai frázis [9] egyetlen hangsúlyos pozícióval rendelkezik, ez magyar nyelv esetén az első szótagon kötött hangsúlyozás miatt a frázis első szótagja. A szótagláncot és a szótagok kezdő- és végidőpontját ismerjük a korpuszból, így a fonológiai frázishatárok ismeretében már csak a hangsúlyos szótagok azonosítása van hátra közvetlenül a fonológiai frázishatár utáni szótagon.

A fonológiai frázisok detektálását végző algoritmust részletesen bemutattuk a [6] irodalomban, illetve korábban az MSzNy konferencián [7], így ehelyütt részleteiben nem ismertetjük, csak az algoritmusban a [6] forrásban dokumentálthoz képest végzett változtatásokat emeljük ki: az alapfrekvencia-követőt lecseréltük a Kaldi toolkit *compute-kaldi-pitch* eszközére, amely zöngétlen keretekre is szolgáltat értéket (a pontos algoritmust lásd: [10]). Ez az alapfrekvencia-követő nagyon



1. ábra. A fonológiai frázisszegmentáló pontossága (és hatékonysága) a TOL toleranciaérték függvényében, $PRC = RCL$ munkapontokra.

kedvező viselkedésű, a Viterbi-algoritmuson és néhány paraméterezhető költségfüggvényen keresztül könnyen elérhető, hogy a szolgáltatott alapfrekvenciakontúr oktávugrásoktól lényegében mentes, konzisztens, simított görbe legyen, amely további utófeldolgozást már nem igényel. Használatával jelentős pontosságnövekedést értünk el.

3.1. A fonológiai frázisszegmentáló kiértékelése

A [6] irodalomban megadott tanítókorpuszon (BABEL) és feltételekkel, de a Kaldi alapfrekvencia-követőjével kinyert jellemzőkön tanítottuk a fonológiai frázisszegmentáláshoz használt modelleket. A tanított HMM/GMM modelleket tízszeres keresztvalidációban ki is értékeltük, kézi fráziscímkézést használva referenciaként. Egy frázis detektálását akkor tekintettük helyesnek, ha a két frázishatár közötti eltérés egy toleranciaértéken (TOL) belüli volt. A detektált frázishatárokra ezután hatékonyság (recall, RCL), pontosság (precision, PRC) és átlagos eltérés (average time deviation, ATD) értékeket számítottunk. Az 1. ábrán látható a frázisszegmentáló frázishatár-detektálásra vonatkozó hatékonysága és pontossága TOL függvényében azokra a munkapontokra, ahol $RCL = PRC$. Ha $TOL = 100ms$, akkor ez a munkapont $PRC = RCL = 71,0\%$, ahol $ATD = 31,9ms$. $TOL = 200ms$ toleranciaértékre $PRC = RCL = 84,8\%$, $ATD = 54,3ms$.

3.2. Hangsúlyok szótagra illesztése

A beszédjel alapú hangsúlycímkézés is háromszintű, az egyes szinteket a fonológiai frázis típusa alapján különítjük el. Mivel a fonológiai frázisok típusainak elkülönítésében éppen a hangsúly erőssége az egyik alkalmazott kritérium, ez nem okoz különösebb nehézséget (lásd az 1. táblázatot). A fonológiai frázisok (FF) hangsúlyának erősségét az intonációs frázison (IF) belüli pozíció (IF kezdetre eső FF erősen hangsúlyos), illetve a szintaktikai, szemantikai és pragmatikai

viszonyok alakítják (pl. a mondathangsúlyt tartalmazó FF is erősen hangsúlyos lesz).

1. táblázat. A fonológiai frázisokhoz tartozó szótaghangsúly erőssége (az első szótagon)

FF típusa	Hangsúly	Jellemzés
me	erős	Intonációs frázis kezdete
fe	erős	Erősen hangsúlyos FF
fs	normál	Normál FF
mv	normál	IF végén ereszkedő kontúrú
fv	normál	IF végén emelkedő kontúrú
s	nincs	Hangsúlytalan(ná vált) FF
sil	nincs	Csend

Az így kapott hangsúlycímkézésre angol elnevezése után (Audio Based Stress Modelling) **ABSM** rövidítéssel hivatkozunk a továbbiakban.

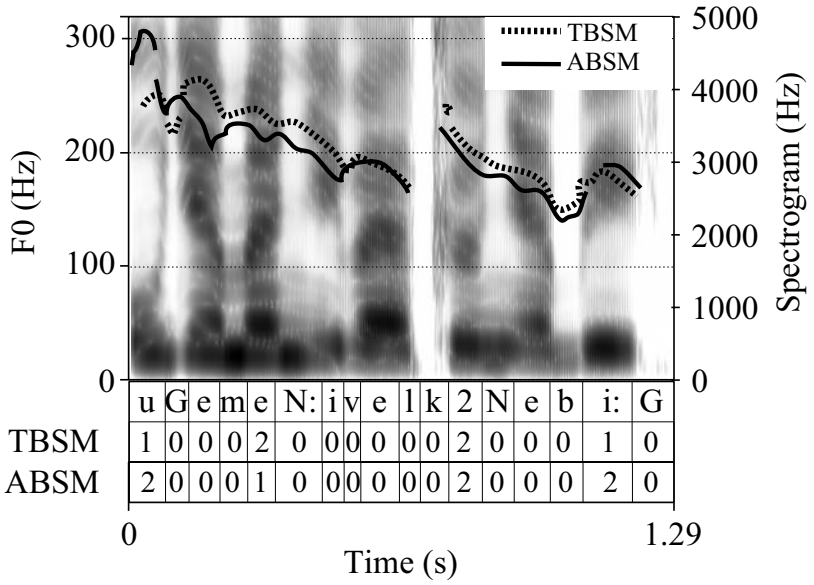
4. A gépi szövegfelolvasó tanítókorpusza

A TTS betanításához használt beszédkorpusz a Magyar Párhuzamos Precíziós Beszédadatbázis, amely 1984 mondatot tartalmaz 14 beszélő felolvasásában [11]. A precíziós címkézés a fonetikai átiratra és a beszédhangszintű címkézésre utal, a kézi hangsúlycímkézés egyelőre még hiányzik az adatbázisból.

A korpuszt a bemutatott két eljárással (TBSM és ABSM) is felcímkéztük hangsúlyokra, majd a címkézést összevetettük hasonlóságuk tekintetében, illetve TTS rendszerekben is.

4.1. A szöveg és a beszédjel alapú hangsúlycímkézés összevetése

A 2. ábrán látható egy rövid példamondatra vonatkozóan a kétféle eljárással generált hangsúlycímkesor. Általánosan elmondható, hogy mind a 14 beszélőt figyelembe véve, ABSM módszerrel az összes szó 48,4%-a, TBSM módszerrel pedig 33,1%-a kapott valamilyen hangsúlyt, tehát a beszédjel alapján másfélszer gyakrabban ítéltünk valamely szótagot hangsúlyosnak. A két módszer közötti fedést vizsgálva meglepő jelenséget tapasztaltunk (lásd 3. ábra): csak hangsúlyos és hangsúlytalan szótagokat megkülönböztetve a két eljárás legalább valamelyike által hangsúlyosnak címkézett szavakra a szavak kevesebb mint 1/3-át jelöli mindkét módszer egységesen hangsúlyosnak. Ennek a viszonylag gyenge átfedésnek a mélyebb vizsgálata kívül esik a cikk jelenlegi témáján, így csak annyit jegyzünk meg, hogy ebben egyrészt vélhetően a TBSM módszer heurisztikus jellege, általánosítóképességének korlátai játszhatnak közre, másrészt befolyásolhatja az eredményt az is, hogy a szintaktikailag kikövetkeztethető hangsúly nem feltétlenül realizálódik akusztikailag is (vö. [12]), de ezt a jelenséget magyar nyelvre tudtunkkal még nem vizsgálták, jóllehet részben [2] eredményei is ebbe az irányba is engednek következtetni.

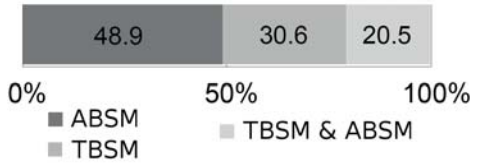


2. ábra. Az „Ugye, mennyivel könnyebb így." mondat ABSM és TBSM címkéi. A beszédhangokat SAMPA kódjukkal adtuk meg, a szótagok hangsúly szerinti címkézésében 0=hangsúlytalan, 1=hangsúlyos, 2=erősen hangsúlyos.

4.2. Kísérleti TTS mintarendszerek

A hullámforma alapján készített hangsúlymodell hatásait magyar nyelvű rejtett Markov-modell alapú szövegfelolvasó rendszerben (Hidden Markov Model based Text-to-Speech, HMM-TTS) [13] vizsgáltuk meg. A HMM-TTS tanítókorpuszaként a magyar nyelvű, párhuzamos, precíziós beszédadatbázis egy női és egy férfi beszédhangját használtuk. A tanító adatbázis mindkét beszélő esetén a teljes, 1984 mondatból álló halmazt tartalmazta. A mondatok 44 kHz-en, 16 biten lettek rögzítve. A döntési fák építéséhez az MDL (Minimum Description Length) kritériumot használtuk. Mind a női, mind pedig a férfi beszélő esetén háromhárom különböző szövegfelolvasó rendszert készítettünk el az alábbiak szerint:

– Az első rendszer döntési fái nem tartalmaztak hangsúllyal kapcsolatos jellemzőket, tehát a tanítás során explicit módon nem adtunk meg hangsúlyozásra vonatkozó információt. Ezt úgy értük el, hogy a tanítás során a döntési fák építéséhez szükség összes hangsúllyal kapcsolatos kérdést eltávolítottuk korábbi szövegfelolvasó rendszerünkből [13]. A továbbiakban erre a rendszerre **NOSM** rövidítéssel (NO Stress Model) hivatkozunk.
– A második rendszer minden hangsúllyal kapcsolatos kérdést tartalmazott, továbbá a tanító adatbázisban a hangsúlyos szótagokat szabály alapon becsültük. Ez a rendszer megegyezik a korábban bemutatott HMM-TTS rend-



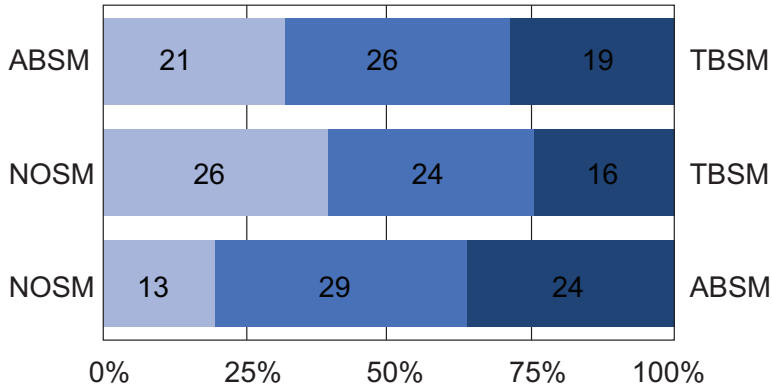
3. ábra. Az ABSM és TBSM hangsúlycímkézések hasonlósága (fedése).

szerünkkel [13]. A cikkben **TBSM** rövidítéssel hivatkozunk erre a megoldásra (Text Based Stress Model).

– A harmadik rendszer szintén minden hangsúllyal kapcsolatos kérdést tartalmazott. Ez esetben azonban a tanító adatbázisban a hangsúlyos szótagokat a jelen cikkben ismertetett módon, statisztikai módszerrel, pusztán a hullámforma alapján határoztuk meg. Szintézis során ez esetben is szabály alapon becsültük a hangsúlyokat. Továbbra is **ABSM** (Audio Based Stress Model) rövidítéssel jelöljük ezen rendszerünket.

5. Kiértékelés

A jelen cikkünkben bemutatott módszer érzeti hatásait szövegfelolvasó rendszerekben párösszehasonlításos meghallgatásos teszttel (Comparison Mean Opinion Score, CMOS) értékeltük ki. A teszt során egymástól függetlenül vizsgáltuk meg a férfi és női beszélőket. A meghallgatásos tesztben a korábban bemutatott három-három rendszer vett részt: NOSM, TBSM és az ABSM. A tesztalanyoknak az egyes rendszerek által generált mondatokat páronként kellett összehasonlítaniuk, aszerint, hogy mennyire találják természetesnek azok prozódiáját. Három lehetőség közül lehetett választani: (1) az első mondat természetesebb hangzású; (2) azonos a két mondat hangzása; (3) a második mondat természetesebb hangzású. Minden mondatpárban a két mondat két különböző rendszerrel lett elkészítve (NOSM vs. TBSM, NOSM vs. ABSM és TBSM vs. ABSM). Egy tesztalany összesen 18 mintapárt hasonlított össze. A mintapárok sorrendjét, és a mintán belül a rendszerek sorrendjét álvéletlen módon alakítottuk ki az esetleges memóriahatások elkerülése céljából. Összesen 21 alany (9 férfi, 12 nő) vett részt a meghallgatásos tesztben, akik összesen 378 mintapárt értékeltek. Minden alany magyar anyanyelvű volt. A legfiatalabb tesztelő 22, a legidősebb 70 éves volt. A tesztalanyok átlagéletkora 34 év volt. A meghallgatásos tesztet az interneten keresztül lehetett kitölteni. A meghallgatásos teszt eredményeit a 4. és az 5. ábra mutatja be. Az eredményeket megvizsgálva a hangsúly-információt nem tartalmazó rendszer (NOSM) mindkét beszélő esetében jobban teljesített, mint a fonetikus átirat alapú hangsúlymodell (TBSM). Bár elsőre meglepő ez az eredmény, a 3. ábrán látottak fényében egybecseng korábbi megállapításainkkal, hogy a beszédkorpuszban ténylegesen megjelenő hangsúlyok és a szöveg alapján becsült hangsúlyok között kevés átfedés lehet. A beszédjel alapú hangsúlymodell (ABSM) férfi beszélő esetén több szavazatot kapott, mint a NOSM, valamint

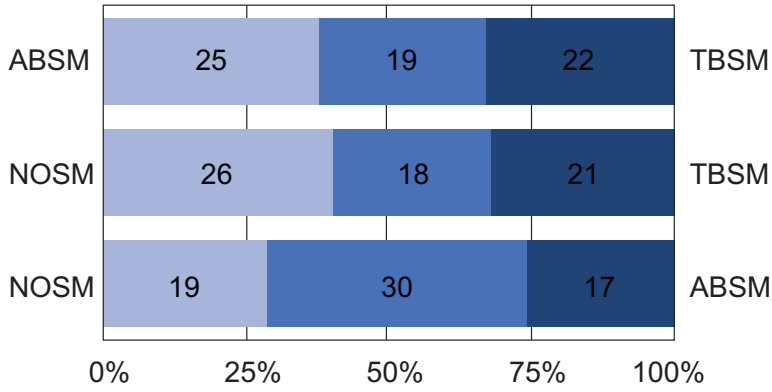


4. ábra. A meghallgatásos teszt eredményei férfi beszélő esetén.

mindkét beszélő esetén jobban teljesített, mint a TBSM. A szignifikanciát egymintás t-teszttel vizsgáltuk $\alpha = 0,05$ mellett. Szignifikáns eltérést találtunk a férfi beszélő esetén a NOSM (hangsúly-információ nélküli) és az ABSM (beszédjel alapú hangsúlymodell) rendszer összehasonlítása során, az utóbbi javára. A női beszélőnél nem sikerült szignifikáns eltérést igazolni, de a szavazatok megoszlásából látható, hogy a hangsúlymodell nélküli rendszer és a beszédjel alapú hangsúlymodell szinte egyenlő szavazatokat kapott a két rendszer prozódiáját azonosnak értékelő hallgatók magas aránya mellett.

6. Összegzés

Cikkünkben automatikus hangsúlycímkézést, illetve hangsúlymodellezést vizsgáltunk a szöveg, valamint a beszédjel alapján magyar nyelvű HMM-TTS rendszerben. A két eljárást az explicit hangsúlyjelölés nélküli esettel és egymással is összehasonlítottuk, páronkénti szubjektív meghallgatásos teszttel. A hangsúlymodellezés hatása csak a tanult HMM-TTS modelleken keresztül érvényesülhet, szintézisidőben ugyanis mindig a szöveg alapján becsültük a hangsúlyokat. Az eredményekből fontos következtetéseket vonhatunk le: a korpuszon végzett, fonetikus átirat alapú hangsúlycímkézésnél előnyösebb, a meghallgatásos teszt alapján történő hangsúlymodellezés nélküli eset, hiszen jobb eredményt ad. A beszédjel alapú hangsúlycímkézés, illetve az ezen a címkézésen végzett modellezés a férfi beszélő esetén szignifikáns javulást eredményezett a beszéd természetességének szubjektív megítélésében, míg a női beszélőnél nem volt szignifikáns különbség a hangsúlymodellezés nélküli esethez képest ($\alpha = 0,05$ mellett). Fontos megjegyezni, hogy a tesztalanyoknak kizárólag a prozódia természetességének megítélése volt a feladatuk, de eközben elkerülhetetlenül befolyásolta döntésüket az érzeti általános beszédminőség is. Az eredmények, beleértve a szöveg és a beszédjel alapján generált hangsúlyok közötti csekélynek mondható átlapolást



5. ábra. A meghallgatásos teszt eredményei női beszélő esetén.

is, felvetik annak a lehetőségét, hogy az emberi percepció a hangsúlyozásban nem a prozódia szintaxist megerősítő szerepét várja, hanem bizonyos tűréshatárral „megengedi" a hangsúlyos helyek váltakozását ugyanazon közlésben, és a hangsúlyra járulékos információforrásként tekint. Ezt a felvetést jelen munkában azonban nem vizsgáltuk, a jelentésbeli percepciós eltérések és a hangsúlyozás kapcsolatáról tehát nem tudunk ennél biztosabb következtetést levonni a rendelkezésünkre álló adatokból. Eredményeink alapján fontosnak találjuk a téma további vizsgálatát, a beszédjel és a hangsúlyok kapcsolatának egzaktabb meghatározását, és a hullámformán alapuló, pontosabb hangsúlymodell gépi beszéd természetességére gyakorolt hatásának elemzését.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki Bartalis István Mátyásnak, a meghallgatásos teszt megtervezésében és kialakításában nyújtott segítségéért;
a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely a PD-112598 projekt keretében a kutatást támogatta;
a Swiss National Science Foundationnak (Svájci Államszövetség), amely az „*SP2: SCOPES project on speech prosody*" (SNSF N^o IZ73Z0-152495/1) számú projekt keretében a kutatásunkat támogatta.

Hivatkozások

1. Pitrelli, J.F., Beckman, M.E., Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of the 1994 International Conference on Spoken Language Processing. Volume 1. (1994) 123–126
2. Beke, A., Szaszák, Gy.: Combining NLP techniques and acoustic analysis for semantic focus detection in speech. In: Proceedings of the 5th IEEE International Conference on Cognitive Infocommunications. (2012) 493–497

3. Heggtveit, P.O., Natvig, J.E.: Automatic prosody labelling of read Norwegian. In: Proceedings of Interspeech. (2004) 2741–2744
4. Wightman, C., Syrdal, A., Stemmer, G., Conkie, A., Beutnagel, M.: Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative speech synthesis. In: Proceedings of International Conference on Spoken Language Processing. Volume 2. (2000) 71–74
5. Levelt, W.J.M.: Speaking: From Intention to Articulation. MIT Press, Cambridge (1989)
6. Szaszák, Gy., Beke, A.: Exploiting prosody for syntactic analysis in automatic speech understanding. Journal of Language Modelling **0**(1) (2012) 143–172
7. Vicsi, K., Szaszák, Gy.: Folyamatos beszéd szó- és frázisszintü automatikus szegmentálása szupraszegmentális jegyek alapján: II. rész: Statisztikai eljárás, finn-magyar nyelvű összehasonlító vizsgálat. In: III. Magyar Számítógépes Nyelvészeti Konferencia. (2005) 360–370
8. 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 **3-4** (2000) 201–215
9. Selkirk, E.: The syntax-phonology interface. In: International Encyclopaedia of the Social and Behavioural Sciences. Oxford: Pergamon (2001) 15407–15412
10. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. (2014) 2494–2498
11. Olaszy, G.: Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai,. Beszédkutatás (2013) 261–270
12. Ananthakrishnan, S., Narayanan, S.: Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Transactions on Audio Speech and Language Processing **16**(1) (2008) 216–228
13. Tóth, B., Németh, G.: Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis. Acta Cybernetica **19**(4) (2010) 715–31

Mély neuronhálós akusztikus modellek gyors adaptációja multi-taszk tanítással

Tóth László, Gosztolya Gábor*

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
e-mail: {tothl, ggabor}@inf.u-szeged.hu

Kivonat A környezetfüggő mély neuronhálós akusztikus modellek gyors adaptációja különösen nehéz kihívás, mivel egy kis méretű adaptációs mintában a környezetfüggő állapotok többségére nincs tanítópélda. Nemrégiben egy olyan új mély neuronhálós tanítási séma bukkant fel, amely a hálózatot egyszerre tanítja környezetfüggő és környezetfüggetlen példákon. Ez az ún. multi-taszk technológia felveti annak a nagyon egyszerű adaptációs módszernek a lehetőségét, hogy az adaptáció során csak környezetfüggetlen címkéken tanítsunk. Jelen cikkben ezt a módszert próbáljuk ki, kombinálva egy KL-divergencia alapú regularizációs technikával. Kísérleteinkben a multi-taszk tanítási séma már önmagában 3%-os hibacsökkenést hoz egy híradós beszédfelismerési feladaton. A kombinált adaptációs módszert is bevetve további 2-5% hibaredukciót sikerült elérnünk az adaptációs minta méretének függvényében, ami 20-tól 100 másodpercig terjedt.

Kulcsszavak: mély neuronháló, akusztikus modellezés, beszédfelismerés, adaptáció

1. Bevezetés

Az utóbbi években a rejtett Markov-modellek (hidden Markov model, HMM) hagyományos Gauss-keverékmodellje (Gaussian mixture model, GMM) helyett egyre inkább a mély neuronhálókat (deep neural network, DNN) kezdik alkalmazni. Az évtizedek alatt azonban a GMM-alapú modellezésnek számos olyan finomítását találták ki, amelyek nem vihetők át triviális módon a HMM/GMM rendszerekből a HMM/DNN rendszerekbe. Az egyik ilyen finomítás a környezetfüggő (context-dependent, CD) modellek készítése és betanítása. Jelen pillanatban a HMM/DNN rendszerek környezetfüggő állapotait ugyanazzal a jól bevált technológiával szokás előállítani, mint a HMM/GMM rendszerekben. Ez azt jelenti, hogy egy mély neuronhálós felismerő készítésének első lépéseként lényegében be kell tanítani egy hagyományos GMM-alapú felismerőt [3,7,12]. Habár születtek javaslatok arra nézve, hogy a GMM-eket hogyan lehetne kihagyni a folyamatból, ezek egyelőre inkább csak kísérleti próbálkozások [1,5,14,20]. Ami a mély neuronhálók környezetfüggő állapotokkal való betanítását illeti, Bell és

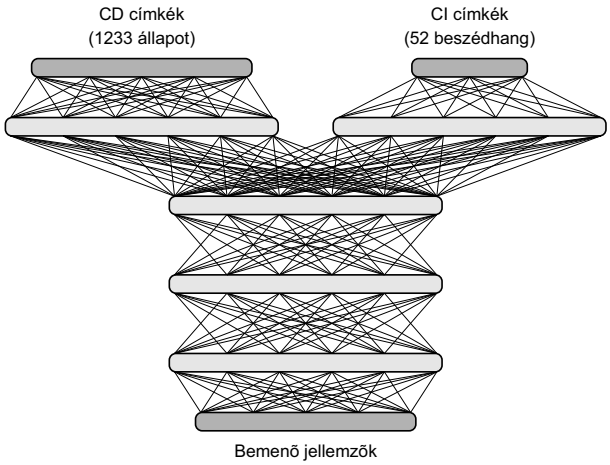
* A jelen kutatás során használt TITAN X grafikus kártyát az NVIDIA Corporation ajándékozta csoportunknak.

társai nemrégiben bemutattak egy új megoldást. Az ún. multi-taszk tanítás lényege, hogy a környezetfüggő címkékkel párhuzamosan környezetfüggetlen (context-independent, CI) címkékkel is tanítjuk a hálózatot [2]. Technikailag ezt úgy lehet megvalósítani, hogy a hálózatba két kimenő réteget veszünk fel, ahol egyikük a CD, másikuk pedig a CI címkék megtanulására törekszik [13]. A CI címkék párhuzamos tanítása egyfajta regularizációs hatást fejt ki a CD címkék tanulása során. Bell és tsai. módszerét mi is kipróbáljuk hamarosan, ami 3% hibacsökkenéshez fog vezetni a szószintű hibában.

A DNN akusztikus modellek adaptálása során a modell regularizációja kiemelt fontossággal bír. Mivel a mély neuronhálók jellemzően sok paraméterrel (réteg, ill. neuron) rendelkeznek, nagyon hajlamosak a túltanulásra, kiváltképp ha az adaptációs minta mérete kicsi. Talán a legelterjedtebb megoldás a túltanulás ellen, amikor a hálózatot kiegészítik egy lineáris réteggel, és az adaptáció során csak ezt a lineáris réteget engedik tanulni [4,16]. Hasonló megoldás a (túl)tanulás korlátozására, ha az adaptáció során csak a rétegek és/vagy súlyok csak egy kis részét engedjük tanulni [9,10]. Egy további megoldási lehetőség, ha csak a neuronok bias értékeit [17], vagy a rejtett neuronok aktivációs amplitúdóját [15] engedjük adaptálódni. A megoldások egy másik csoportja a túltanulás kockázatát valamilyen regularizációs megszorítás alkalmazásával csökkenti. Li és tsai. olyan L2-regularizáció alkalmazását javasolták, amely bünteti az adaptáció előtti és utáni hálózati súlyértékek nagy eltérését [8]. Gemello az ún. 'konzervatív tanítást' javasolta, melynek lényege, hogy az adaptációs mintában nem szereplő osztályokra az adaptálatlan hálózat kimeneteit használjuk a tanítás során célértékként [4]. Yu és tsai. egy olyan megoldást vetettek fel, amelyben a tanulási célértékek az adaptálatlan modell kimenete és az adaptációs minta címkéi közötti lineáris interpolációval állnak elő. Matematikailag ez a megoldás a Kullback-Leibler divergencia regularizációjaként formalizálható [18].

A környezetfüggő modellek használata jelentősen megnöveli a túltanulás kockázatát az adaptáció során, hiszen az állapotszám megnövelése lecsökkenti az egy állapotra eső tanítópéldák számát. Price és tsai. erre egy olyan hálózati struktúrát javasoltak, amelyben két kimeneti réteg épül egymásra, ahol az alsó a CD, a felső pedig a CI címkéknek felel meg [11]. Ezzel a megoldással betanítás és felismerés során a CD címkéket lehet használni, míg a CI kimeneti réteggel dolgozunk az adaptáció során, amikor kevés a címkézett tanítóadat.

Ebben a cikkben egy olyan megoldást javasolunk, amely alapötletében hasonlít Price és tsai. megoldásához, de az alkalmazott hálózati topológia teljesen más. Míg ők a CD és CI címkéknek megfelelő kimeneti rétegeket egymás fölé helyezték, mi egymás mellé rakjuk azokat, hasonlóan a multi-taszk tanítás során alkalmazott elrendezéshez. Ezzel a struktúrával az adaptáció módja triviálisan adódik: míg a (multi-taszk) betanítás során mind a CD, mind a CI kimeneti réteg kap mintákat, adaptáció során csak a CI kimenetet tanítjuk. Hogy tovább csökkentsük a túltanulás kockázatát, a tanítás során a Yu-féle KL-regularizációs technikát is alkalmazni fogjuk [18]. Kísérleteink azt mutatják, hogy ennek a regularizációnak kritikus szerepe van, főleg amikor az adaptációs mintahalmaz nagyon kicsi. A kombinált módszert egy felügyelet nélküli adaptációs feladaton



1. ábra. A multi-taszk neuronháló struktúrája.

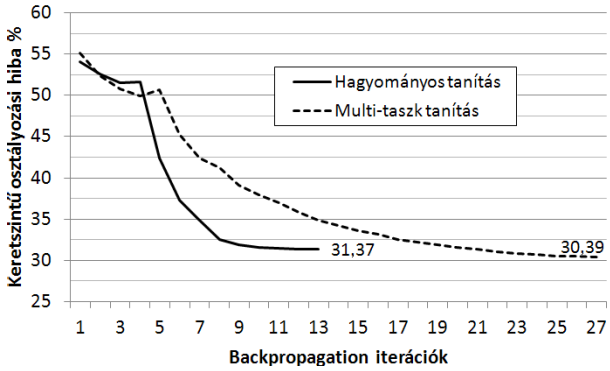
fogjuk kipróbálni, ahol az adaptációs minta mérete 20 és 100 másodperc között ingadozik. E hossz függvényében 2% és 5% közötti relatív hibacsökkenést sikerült elérni.

2. Multi-taszk tanítás

A multi-taszk tanítás lényege, hogy egy gépi tanuló algoritmusnak több megtanulandó feladatot adunk párhuzamosan, amitől jobb általánosítási képesség elérését reméljük. Tudomásunk szerint a multi-taszk tanítást DNN akusztikus modellek készítése során Seltzer és Droppo alkalmazták először. A TIMIT beszédhang-felismerési feladaton kísérletezve azt tapasztalták, hogy az aktuális adatvektor felismerési pontossága megnő, ha a hálózatnak másodlagos feladatként a fonetikai környezetet is meg kell tanulnia [13]. Bell és tsai kísérletében a CD címkéket tanuló neuronháló a CI címkék felismerését kapta másodlagos feladatként, ami 3%-10% relatív csökkentést hozott a szószintű hibában a hagyományos tanításhoz képest [2].

Az 1. ábra mutatja az általunk alkalmazott hálózati topológiát. Mint látható, a hálónak két kimeneti rétege van, egy a CD címkék és egy a CI címkék számára. A két kimeneti réteghez igazítva a legfölső rejtett réteget is kettéosztottuk, ami eltérést jelent Bell és mtsai. megoldásához képest [2]. Ezzel a struktúrával némileg jobb eredményeket kaptunk, habár a javulás nem volt szignifikáns.

Szintén a Bell-féle cikket követve a CI címkék esetén a beszédhangokat nem szedtük szét a szokásos 3 állapotra, azaz a CI címkék megfeleltek a monofón beszédhang-címkéknek. Tanítás előtt a CD állapotokat átkonvertáltuk a megfelelő monofón címkékre, és a tanítás során a hálózat mindkét fajta címkét megkapta. A tanítás folyamán minden egyes adatköteget (batch) véletlenszerűen a CD



2. ábra. A keretszintű CD hiba alakulása a validációs halmazon hagyományos és multi-taszk tanítás esetén.

vagy a CI kimeneti réteghez rendeltünk, és a hiba visszapropagálását csak a hálózat adott felén hajtottuk végre. A közös rétegek súlyait természetesen minden esetben frissítettük, míg a megosztott rejtett réteg és kimeneti réteg esetében csak az aktuális hozzárendelt hálózati ág súlyai tanultak. A következő szekcióban megmutatjuk, hogy ez a tanulási technika hogyan befolyásolja a modell konvergenciáját.

3. Multi-taszk tanítási kísérletek

A kísérletek során a „Szeged" magyar híradós beszédadatbázist használtuk [6]. A korpusz 28 órányi híradófelvételt tartalmaz nyolc tévécsatornáról. Az adatok tanító-tesztelő felosztása és a nyelvi modell ugyanaz volt, mint a korábbi munkáinkban [6]. A CD állapotok előállítására az ICASSP 2015 konferencián bemutatott módszerünket használtuk, ami 1233 trifón állapotot eredményezett [5]. A CI címkék száma 52 volt.

A kiindulási modellként alkalmazott mély neuronháló 4 rejtett réteget tartalmazott, rétegenként 2000 egyenirányított lineáris (rectified linear, ReLU) neuronnal. A multi-taszk tanítás céljaira a hálót az 1. ábrán látható módon alakítottuk át. A multi-taszk hálónak két kimenetei rétege volt, egy a CD és egy a CI címkék tanulásához, valamint a legfölső rejtett rétegnek is két változata volt, rendre 2000-2000 neuronnal. A tanítás a backpropagation algoritmussal történt, melynek során a szokványos keretenkénti keresztentrópia hibafüggvényt minimalizáltuk, az adatkötegeket a korábban leírt módon hol az egyik, hogy a másik kimeneti réteghez rendelve. A backpropagation előtt előtanítási módszert nem használtunk, mivel a korábbi eredmények azt mutatták, hogy ReLU neuronok használata esetén az előtanulásnak semmi vagy minimális haszna van csak [6,19].

A kísérletezés során megpróbáltuk belőni a CD, illetve a CI ágra irányított adatcsomagok optimális arányát. Míg Bell és tsai. az 50-50%-ot találták optimálisnak, esetünkben a 75%-25% arány (a CD kimenet javára) kicsit alacsonyabb

1. táblázat. A keretszintű (FER) és a szószintű (WER) hiba alakulása a tanító, validáló és teszthalmazokon.

Tanítási mód	FER %		WER %	
	Tanító h.	Val. h.	Val. h.	Teszth.
Hagyományos	25.9%	31.4%	17.7%	17.0%
Multi-taszk	23.5%	30.4%	17.4%	16.5%

hibaarányt adott az adatkeretek szintjén, habár ez a szószintű hibaarányt nem befolyásolta számottevően.

Ha a hálózatnak két dolgot kell egyidejűleg tanulnia, az értelemszerűen megnehezíti a tanulás konvergenciáját. Esetünkben a tanítás a backpropagation algoritmussal történt, ahol a tanulási rátát exponenciálisan felezgettük. Azt tapasztaltuk, hogy a multi-taszk tanítás során nem lehet olyan gyors léptékben csökkenteni a tanulási rátát, mint a hagyományos tanulás esetén: a 0,5-es szorzó helyett 0,8-del kaptuk a legjobb eredményt. Ugyanazt a megállási feltételt használva a multi-taszk tanulásnak körülbelül kétszer annyi iterációra volt szüksége a konvergenciához. A 2. ábrán egy példát láthatunk arra, hogy a tanítás során hogyan csökken a CD kimeneten számolt hiba a hagyományos és a multi-taszk tanítás esetén.

A 1. táblázatban összehasonlíthatjuk a kétféle tanítási móddal kapott végső hibaarányokat. Mint láthatjuk, a multi-taszk tanítással kb. 3% szószintű hibaarány-csökkenést sikerült elérni, ami nagyságrendileg megegyezik Bell és tsai. eredményeivel [2]. Azonban, míg ők azt találták, hogy a kisebb szószintű hiba ellenére a keretszintű CD hiba *nőtt*, mi ilyen ellentmondást nem tapasztaltunk. Ennek oka az lehet, hogy mi nagyobb arányban mutattunk CD példákat a hálónak, így a tanulás során nagyobb hangsúlyt kapott a CD kimeneten mért hiba.

4. Akusztikus adaptáció a multi-taszk modellel

Az állapotkapcsolt környezetfüggő modelleket előállító algoritmus zsenialitása abban rejlik, hogy a CD állapotok számát hozzá tudjuk igazítani a rendelkezésre álló tanító adatok mennyiségéhez. A gyakorlatban mindig arra törekszünk, hogy a CD állapotok számát olyan nagyra válasszuk, amennyi állapotot még biztonságosan be tudunk tanítani a túltanulás veszélye nélkül. Ha azonban a betanított modelljeinket egy új környezethez vagy beszélőhöz kell adaptálni, akkor az adaptációs tanításhoz rendelkezésünkre álló példák száma rendszerint nagyságrendekkel kisebb a teljes tanítóhalmaznál. Emiatt a CD modellek adaptációs mintára való tanítása szükségszerűen magában hordozza a túltanulás veszélyét. Azonban a multi-taszk modell egy kézenfekvő megoldást kínál a túltanulás esélyének csökkentésére: az adaptáció során a modell CD ágának nem mutatunk példákat, mivel a legtöbb CD címkére úgysem lenne példa a kis méretű adaptá-

ciós halmazban. Ehelyett az adaptáció során csakis a CI ágnak adunk példákat, amely ágat az adathiány problémája jóval kevésbé sújtja.

A mély neuronhálós akusztikus modelleknek rengeteg paraméterük (azaz súlyuk) van, ami nagyfokú rugalmasságot biztosít nagy tanítóhalmaz esetén, viszont növeli a túltanulás kockázatát egy kicsi adaptációs halmazon. Erre a legegyszerűbb megoldás, ha az adaptáció során csak a paraméterek egy részét – például egyetlen rejtett réteget – engedünk tanulni [10]. Ezzel ráadásul az adaptáció időigényét is csökkentjük. Mi a tanulást úgy korlátoztuk az adaptáció során, hogy az csak a CD és CI ágak legfelső közös rejtett rétegének súlyait frissítse (l. 1. ábra). Ezzel a megszorítással együtt is azt tapasztaltuk, hogy felügyelet nélkül adaptáció esetén nagyon nehéz megtalálni az optimális tanulási rátát. Míg kis értékek mellett stabil, de szerény hibacsökkenést kaptunk a fájlok legtöbbjére, nagyobb értékek egyes fájlokra nagy javulást adtak, másokra pedig hatalmas romlást. Arra gyanakodtunk, hogy az adaptációt a hibásan becsült adaptációs címkék viszik félre, és ezért a Yu-féle regularizációs megoldás [18] kipróbálása mellett döntöttünk. A módszer lényege, hogy a tanulás során büntetjük, ha az adaptált modell kimenete nagyon eltérne az adaptálatlan modell kimenetétől. Mivel a neuronhálók kimenete diszkrét valószínűségi eloszlásként értelmezhető, az eltérés mérésére a Kullback-Leibler divergencia adódik természetes megoldásként. Némi levezetés után (l. [18]) azt kapjuk, hogy az adaptáció során az adaptációs mintán kapott becsült címkéket simítani kell az adaptálatlan modell kimenő valószínűségeivel. Formálisan, a tanulási célok előállításához az alábbi lineáris interpolációt kell alkalmaznunk:

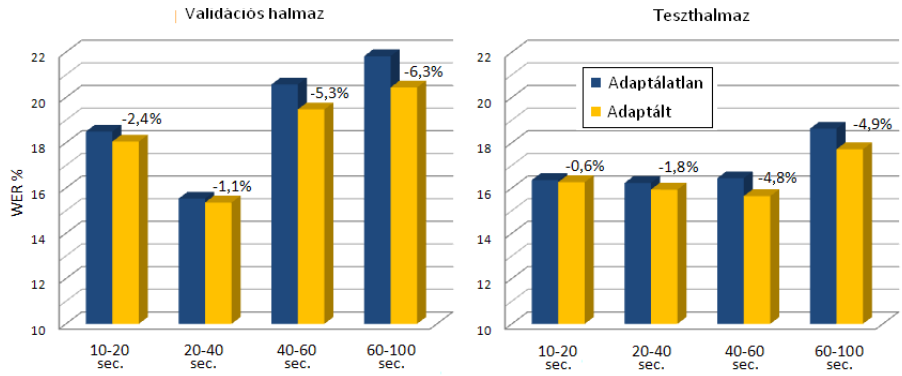
$$(1-\alpha)p(y|x) + \alpha p_{un}(y|x),$$

ahol $p(y|x)$ jelöli a 0-1 jellegű adaptációs címkézést (ez felügyelet nélküli esetben felismeréssel, felügyelt esetben kényszerített illesztéssel áll elő), $p_{un}(y|x)$ az adaptálatlan modell kimenete, az α paraméter segítségével pedig a simítás erősségét lehet állítani.

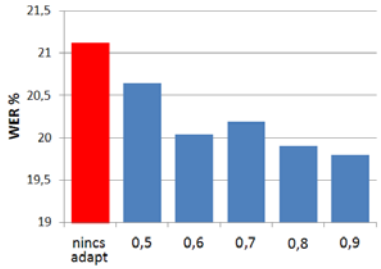
5. Kísérletek felügyelet nélküli adaptációval

A híradós tanítókorpuszunk validációs része 448 felvételt tartalmazott (kb. 2 óra összhosszban), míg a teszthalmaz 724 fájlból állt (4 óra összhosszal). Az egyes fájlok hossza egy mondat (pár másodperc) és kb. 100 másodperc között szórt. Az adaptációs kísérlet során a 10 másodpercnél rövidebb fájlokat nem használtuk. Azt biztosan lehetett tudni, hogy egy fájlon belül a beszélő személye és az akusztikai viszonyok nem változnak, tehát az adaptációnak van értelme, de ezen túl más, a beszélőre vonatkozó információ nem állt rendelkezésre. A kísérletek minden esetben felügyelet nélkül adaptációra törekedtek. Ez abból állt, hogy az adott fájlt felismertettük az adaptálatlan modellel, és a kapott becsült szöveges átiratot használtuk a modell adaptációs tanítására. Ezután a felismerést megismételtük, ezúttal már a fájlhoz adaptált modellel.

Az adaptációnak több paramétere volt, amelyeket a validációs halmazon kellett belőnünk. Ilyen paraméter volt a tanulási ráta, a tanulási iterációk száma, valamint a KL-regularizációs módszer α paramétere. A kezdeti, KL-regularizációt



3. ábra. A szószintű hiba (WER) csökkenése az adaptáció során az adaptációs minta méretének függvényében.



4. ábra. A KL-divergencián alapuló regularizációs módszer α paraméterének hatása a szószintű hibára (WER).

nem alkalmazó kísérleteinkben az optimális tanulási ráta fájlonként nagyon nagy eltéréseket mutatott, a regularizáció bevezetése után azonban az eredmények jóval stabilabbá váltak. A végső tesztekben öt tanítási iterációt mentünk minden fájlon, a normál tanítási rátához hasonló nagyságrendű rátával indulva.

A 3. ábra mutatja a szószintű hiba alakulását adaptáció előtt és után. A fájlokat a hosszuk függvényében négy csoportra osztottuk. Mint látható, a teszthalmazon 10 és 20 másodperc közti hossz esetén a hiba csak minimálisan csökkent, és még a 20-40 másodperc közti hossztartományban is csupán 2%-ot esett. Azonban a 40 másodpercnél hosszabb fájlok esetén a relatív hibacsökkenés 5-6%-ra ment fel a validációs halmazon és 5%-ra a teszthalmazon. Sajnálatos módon adatbázisunk nem tartalmazott 100 másodpercnél hosszabb fájlokat, így algoritmusunk tesztelését nem tudtuk hosszabb fájlokra is kiterjeszteni.

A 4. ábra érzékelteti a KL-regularizációs módszer hozzájárulását a jó eredményekhez. A kiértékelést a validációs készlet 40 másodpercnél hosszabb fájljain végeztük. Az ábrából nyilvánvalóan látszik, hogy a regularizációnak kulcsszerepe volt az adaptáció hatékonyságában. A legjobb eredményt mindig elég erős regu-

larizációval, $0,8 - 0,9$ közötti α értékekkel kaptuk, még a leghosszabb fájlméret (60-100 mp.) esetén is.

6. Konklúzió

A DNN akusztikus modellek adaptációja jelenleg nagyon aktív kutatási terület. A környezetfüggő DNN modellek adaptálása az adatelégtelenségi probléma miatt különösen nagy kihívás. A nemrégiben javasolt multi-taszk tanítási modell környezetfüggetlen címkéken is tanít, így kézenfekvő megoldást kínál az adaptációra. Kísérleteinkben azt tapasztaltuk, hogy mindemellett a Yu-féle regularizációs trükköt is be kellett vetnünk ahhoz, hogy stabilan viselkedő adaptációs eljárást kapjunk. Ezzel a megoldással egy híradós felismerési feladaton 3% relatív szóhiba-csökkenést értünk el csupán a multi-taszk tanítással, majd további 2%-5% hibacsökkenést az általunk javasolt adaptációs technikával.

Hivatkozások

1. Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In: Proc. of ICASSP. pp. 230–234 (2014)
2. Bell, P., Renals, S.: Regularization of deep neural networks with context-independent multi-task training. In: Proc. ICASSP. pp. 4290–4294 (2015)
3. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. ASLP 20(1), 30–42 (2012)
4. Gemello, R., Mana, F., Scanzio, S., Laface, P., de Mori, R.: Linear hidden transformations for adaptation of hybrid ANN/HMM models. Speech Communication 49(10-11), 827–835 (2007)
5. Gosztolya, G., Grósz, T., Tóth, L., Imseng, D.: Building context-dependent DNN acoustic models using Kullback-Leibler divergence-based state tying. In: Proc. ICASSP. pp. 4570 – 4574 (2015)
6. Grósz, T., Tóth, L.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: Proc. TSD. pp. 36–43 (2013)
7. Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V.: Application of pretrained deep neural networks to large vocabulary speech recognition. In: Proc. Interspeech (2012)
8. Li, X., Bilmes, J.: Regularized adaptation of discriminative classifiers. In: Proc. of ICASSP. Toulouse, France (2006)
9. Liao, H.: Speaker adaptation of context dependent Deep Neural Networks. In: Proc. of ICASSP. pp. 7947–7951. Vancouver, Canada (2013)
10. Ochiai, T., Matsuda, S., Lu, X., Hori, C., Katagiri, S.: Speaker adaptive training using deep neural networks. In: Proc. ICASSP. pp. 6399–6403 (2014)
11. Price, R., Iso, K., Shinoda, K.: Speaker adaptation of deep neural networks using a hierarchy of output layers. In: Proc. SLT. pp. 153–158 (2014)
12. Seide, F., Li, G., Chen, L., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. ASRU. pp. 24–29 (2011)
13. Seltzer, M., Droppo, J.: Multi-task learning in deep neural networks for improved phoneme recognition. In: Proc. ICASSP. pp. 6965–6969 (2013)

14. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN training. In: Proc. of ICASSP. pp. 307–312 (2014)
15. Swietojanski, P., Renals, S.: Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In: Proc. SLT. 171-176 (2014)
16. Trmal, J., Zelinka, J., Müller, L.: Adaptation of feedforward artificial neural network using a linear transform. In: Proc. TSD. pp. 423–430 (2010)
17. Yao, K., Yu, D., Seide, F., Su, H., Deng, L., Gong, Y.: Adaptation of context-dependent Deep Neural Networks for Automatic Speech Recognition. In: Proc. of SLT. pp. 366–369. Miami, Florida, USA (2012)
18. Yu, D., Yao, K., Su, H., Li, G., Seide, F.: KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proc. ICASSP. pp. 7893–7897 (2013)
19. Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G.E.: On rectified linear units for speech processing. In: Proc. ICASSP. pp. 3517–3521 (2013)
20. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: Proc. of ICASSP. pp. 5597–5601 (2014)

IV. Szemantika, szentimentelemzés

Angol és magyar nyelvű kérdések a számítógépes nyelvészetben

Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat A cikkben korpuszalapú vizsgálatok segítségével bemutatjuk a magyar és angol nyelvű kérdések sajátságait, különös figyelmet fordítva a közösségi médiában előforduló kérdésekre. Emellett a kérdések számítógépes nyelvészeti hasznosíthatóságára is rámutatunk, egyrészt többszavas kifejezések azonosításában, másrészt eldöntendő kérdésekre felajánlott automatikus válaszlehetőségek továbbfejlesztésében.

Kulcsszavak: kérdések, szemantika, pragmatika, többszavas kifejezések, alkalmazás

1. Bevezetés

A kérdések szerepe a természetes nyelvekben főként az, hogy információt kérjünk másoktól, így központi fontossággal bírnak az emberi kommunikációban. A kérdések a legtöbb nyelvben eltérően viselkednek az állításoktól, és sajátos nyelvi (logikai, interperszonális és szintaktikai-szemantikai) jellemzőkkel bírnak, amire több nyelvészeti tanulmány is rámutatott (pl. [1,2,3]), a számítógépes nyelvészetben viszont – a kérdések megválaszolása (question answering) témakört leszámítva – kevésbé vizsgált témának számít.

Jelen cikk célja, hogy korpuszalapú vizsgálatok segítségével megvizsgáljuk a magyar és angol nyelvű kérdések sajátságait, különös figyelmet fordítva a közösségi médiában előforduló kérdésekre. Emellett a kérdések számítógépes nyelvészeti hasznosíthatóságára is rámutatunk, egyrészt angol nyelvű többszavas kifejezések azonosításában, másrészt magyar nyelvű eldöntendő kérdésekre felajánlott automatikus válaszlehetőségek továbbfejlesztésében.

2. Angol nyelvű adatok

Angol nyelvű adataink két forrásból származnak: az angol Univerzális Dependencia Treebank (UD) [4] és a QuestionBank (QB) [5]. Utóbbi 4000 kérdést tartalmaz, az UD treebankben pedig a mondatok 6,76%-a (1124 darab) kérdés. A kérdőszavak részletes elemzéséből (l. 1. táblázat) kiderül többek között, hogy például a *how* és *why* szavak használata sokkal gyakoribb az UD treebankben,

amelyben szerepelnek közösségi médiából származó szövegek is: a felhasználók gyakran kérnek egymástól segítséget a weben különféle ügyekkel kapcsolatban, ugyanakkor ritkábban tesznek fel személyekre vagy helyekre, időpontokra irányuló tényszerű kérdéseket, ellentétben a QB kérdéseivel.

1. táblázat. Az angol kérdőszavak eloszlása.

	UD	**%**	**QB**	**%**	**Összesen**	**%**
what	173	43,14	2333	60,63	2506	58,98
who	21	5,24	445	11,56	466	10,97
how	72	17,96	233	6,06	305	7,18
where	25	6,23	243	6,31	268	6,31
when	5	1,25	199	5,17	204	4,80
how many	6	1,50	173	4,50	179	4,21
which	15	3,74	87	2,26	102	2,40
why	48	11,97	49	1,27	97	2,28
how much	12	2,99	49	1,27	61	1,44
how long	5	1,25	26	0,68	31	0,73
how about	11	2,74	0	0,00	11	0,26
what about	8	2,00	0	0,00	8	0,19
whose	0	0,00	8	0,21	8	0,19
whom	0	0,00	3	0,08	3	0,07
Összesen	401	100,00	3848	100,00	4249	100,00

3. Kérdések a nem sztenderd szövegekben

A webes nyelvhasználat, különös tekintettel a felhasználók által írt szövegekre, számos olyan sajátossággal rendelkezik, melyek eltérnek a sztenderd nyelvtől [6,7,8]. Ide sorolhatjuk többek között a központozás hiányát vagy következetlen használatát, a rövidítések gyakori használatát, az újonnan alkotott szavakat, a gyakori elírásokat és emotikonokat. Bizonyos esetekben pedig a mondatok szintaktikai szerkezete is eltérhet a sztenderdtől. Például az angolban gyakran elmaradnak az alanyi funkciót betöltő első személyű névmások:

Can't believe you left last night.

A mondatok sokszor ellipszist tartalmaznak, és pusztán egy frázisból állnak:

Very professional.
Reasonable rate.

Az elliptikus kérdések számos társalgási funkciót tölthetnek be, céljuk lehet a visszakérdezés vagy a tényleges kérdésfeltevés. Utóbbi esetben egyes, a sztenderd nyelvben kötelező mondatrészek hiányozhatnak[1].

Az alábbiakban közlünk néhány példát a korpuszból, kiegészítve a teljes, sztenderd nyelvnek megfelelő kérdésváltozattal:

Any feedback from Rick Buy? vs. **Is there** any feedback from Rick Buy?
Sushi tonight? vs. **How about eating** sushi tonight?
Any help? vs. **Could you please give me** any help?
Weather in december in Tremblant? vs. **What is the** weather **like** in December in Tremblant?
Paris or England while studying aboard? vs. **Should I choose** Paris or England while studying abroad?
Dwarf Hamster Making Too Much Noise On Wheel at Night? vs. **What should I do if my** dwarf hamster **is** making too much noise on **its** wheel at night?

A fentihez hasonló kérdések automatikus feldolgozását több tényező is nehezíti. Egyrészt a kontextus ismerete nélkül nem könnyű azonosítani a kérdések szándékolt jelentését, másrészt ehhez gyakran világtudás is szükséges. Gyakran kérésként értelmezendők (*Any help?*), máskor ajánlatként vagy javaslatként (*Sushi tonight?*), érdeklődésként (*Any feedback from Rick Buy?* és *Weather in december in Tremblant?*) vagy pusztán problémaleírásként és közvetett segítségkérésként (*Dwarf Hamster Making Too Much Noise On Wheel at Night?*). A hasonló példák részletes elemzése igen kívánatos lenne a közösségi média szövegeinek automatikus feldolgozása céljából, jelen cikkben részletesebben azonban a nem elliptikus kérdésekkel foglalkozunk.

4. Magyar nyelvű adatok

A magyar adatok a Szeged Dependencia Treebankből [10] származnak, kiegészítve 2000 mondatnyi, webről származó szöveggel [11]. A szövegek összesen 5668 kérdést tartalmaznak. Arányaiban a legtöbb kérdés a webes szövegekben szerepel, hiszen a szövegek egy része a gyakorikerdesek.hu weboldalról származik, ahol felhasználók válaszolnak egymás kérdéseire. Az adatokból az is kiderül, hogy a legtöbb eldöntendő kérdés a webes és az irodalmi szövegekben fordul elő, ahol főleg interperszonális funkcióval bírnak: a hallgató egyetértésének vagy beleegyezésének kiváltása. Ezzel ellentétben a jogi szövegeket, illetve gazdasági híreket tartalmazó alkorpuszokban alig-alig találhatunk kérdéseket, természetesen ez az adott szövegtípusok leíró jellegéből fakad.

A kérdőszavak eloszlása doménenként változó, azonban leginkább a *mi*, *milyen* és *miért* kérdőszavak fordulnak elő az adatbázisban. Számítástechnikai szövegekben gyakori még a *hogyan* használata, ami azzal magyarázható, hogy a

[1] Hasonló jelenséget figyelhetünk meg az angol médiában használt címekben, melyek szintén sajátos szintaktikai szerkezeteket tartalmaznak [9].

2. táblázat. Statisztikai adatok a magyar korpuszban található kérdésekről.

Alkorpusz	**mondat**	**kérdés**	**%**	**kérdőszó**	**k.szó/kérdés**	**eldönt.**	**eldönt./kérdés**
iskolás	24720	1435	5,81	1352	94,22	83	5,78
irodalom	18558	2408	12,98	1660	68,94	748	31,06
számítástechnika	9627	532	5,53	498	93,61	34	6,39
újság	10210	689	6,75	601	87,23	88	12,77
rövidhír	9574	71	0,74	70	98,59	1	1,41
jog	9278	243	2,62	243	100,00	0	0,00
web	1935	290	14,99	217	74,83	73	25,17
Összesen	83902	5668	6,76	4641	81,88	1027	18,12

3. táblázat. A leggyakoribb magyar kérdőszavak eloszlása.

kérdőszó	**iskolás**		**irodalom**		**sz.tech.**		**jog**		**újság**		**hír**		**web**		**Összesen**	
mi	397	28,75	605	36,01	109	19,96	22	12,43	155	26,32	8	13,33	77	33,77	1373	29,46
milyen	263	19,04	141	8,39	109	19,96	65	36,72	70	11,88	13	21,67	12	5,26	673	14,44
miért	167	12,09	211	12,56	33	6,04	4	2,26	48	8,15	2	3,33	35	15,35	500	10,73
ki	105	7,60	105	6,25	32	5,86	19	10,73	60	10,19	6	10,00	25	10,96	352	7,55
hogyan	54	3,91	72	4,29	62	11,36	8	4,52	47	7,98	1	1,67	5	2,19	249	5,34
mennyi	60	4,34	67	3,99	39	7,14	15	8,47	38	6,45	5	8,33	9	3,95	233	5,00
hogy	72	5,21	114	6,79	2	0,37	0	0,00	6	1,02	0	0,00	13	5,70	207	4,44
hol	53	3,84	77	4,58	10	1,83	2	1,13	25	4,24	0	0,00	16	7,02	183	3,93
mikor	47	3,40	35	2,08	11	2,01	9	5,08	26	4,41	6	10,00	6	2,63	140	3,00
melyik	29	2,10	23	1,37	35	6,41	13	7,34	18	3,06	5	8,33	5	2,19	128	2,75

számítástechnikai kézikönyv gyakran tartalmaz technikai jellegű leírásokat arra nézve, hogy mit hogyan kell beállítani a számítógépen. Érdekességképpen megfigyelhetjük azt is, hogy informális nyelvhasználatban (webes, iskolás és irodalmi szövegek) jóval gyakrabban fordul elő a *hogy* kérdőszó, mint a hosszabb, sztenderd változata (*hogyan*).

A magyar webes szövegekben is találunk arra példákat, hogy a hallgató szándéka nem információkérésre irányul, hanem például javasol vagy kér valamit (*Nem lehetne háromnegyed 9kor?*), esetleg az aktuális témával kapcsolatos ellenszenvét fejezi ki (*Szalagavatón táncolni én?*). E típusú kérdések természetesen pragmatikai jelentéstartalommal bírnak, így az elvárt válasz sem mindig egyszerűsíthető le az igen/nem válaszok egyikére. A későbbiekben szeretnénk e kérdéstípusokat részletes vizsgálatnak is alávetni mind elméleti, mind számítógépes nyelvészeti szempontból.

5. Angol prepozíciós igék azonosítása

A többszavas kifejezések több tokenből állnak össze, melyekre jellemző, hogy a teljes egység jelentése (részben) különbözik az egységek saját jelentésétől [12]. A többszavas kifejezések közé tartoznak az úgynevezett angol prepozíciós igék (VPC-k), melyek egy ige és egy (vagy több) prepozíció kombinációjából állnak (*set up* vagy *come in*). A VPC-k felszíni szintaktikai szerkezete gyakran hasonlít más kompozicionális szintaktikai frázisokra: például a *to set up the rules* és a *to run up the road* kifejezések első látásra hasonló felépítésűnek tűnnek, azonban míg az elsőben VPC-t találunk, addig a második példa pusztán egy ige + prepozíciós frázis kombinációja.

A fentiek alapján tehát a felszíni szintaktikai szerkezet nem feltétlenül bír megkülönböztető erővel a VPC-k automatikus azonosításában. Azonban még létezik néhány olyan szintaktikai és szemantikai teszt, melyek segíthetnek abban, hogy a VPC-ket elkülönítsük más hasonló felépítésű, ám kompozicionális egységektől, melyek egyike az aktuális mondat kérdéssé alakítása. A VPC-t tartalmazó mondatokból képzett kérdések gyakran tartalmazzák a *who* és *what* kérdőszavakat, míg a prepozíció a mondat végén helyezkedik el (*What did you set up?*). Ezzel szemben a kompozicionális egységekből létrehozott kérdésekben többnyire a *where* és *when* kérdőszavak fordulnak elő (*Where did you run?*). Mindemellett fontos hangsúlyozni, hogy a **Where did you set?* és a **What did you run up?* kérdések nyelvtanilag nem elfogadhatók.

A következőkben bemutatjuk, hogy az angol prepozíciós igék (VPC-k) automatikus azonosítására pozitív hatással bírnak a kérdésekre irányuló nyelvi jellemzők.

5.1. Gépi tanulási módszerek

Gépi tanulási kísérleteinkben nagyrészt [13] eredményeire támaszkodtunk. Rendszerünk kiértékeléséhez a Tu & Roth korpuszt [14] választottuk. Hogy eredményeink teljes egészében összevethetők legyenek a korpuszon elért korábbi eredményekkel [14,13], egy SVM modell tanítottunk [15], ötszörös keresztvalidációval, a Weka csomag [16] alapbeállításait használva.

A kiértékelésben a pontosság (accuracy) metrikát használtuk.

5.2. Felhasznált jellemzők

Méréseink során a következő egyszerű jellemzőket használtuk fel:

1. **Kérdésekre vonatkozó jellemzők**:
   (a) kérdőszó;
   (b) a kérdőszó szófaji elemzése;
   (c) a kérdőszó mondatbeli helye (a mondat élén áll-e vagy sem);
   (d) a kérdőszó távolsága a megelőző igétől;
   (e) a kérdőszó távolsága a megelőző főnévtől;
   (f) a kérdőszó szintaktikai szerepe.

2. **Igei jellemzők**:
   (a) megvizsgáltuk, hogy az ige lemmája megegyezik-e a leggyakoribb angol igék egyikével, mivel általában a leggyakoribb igék szerepelnek VPC-kben;
   (b) megvizsgáltuk, hogy az ige mozgást fejez-e ki, mivel a VPC-kben sokszor mozgást jelentő igét találhatunk (pl. *come*, *go*).
3. **A prepozícióra vonatkozó jellemzők**:
   (a) megvizsgáltuk, hogy a prepozíció egyike-e az angol nyelv leggyakoribb prepozícióinak;
   (b) megvizsgáltuk, hogy a prepozíció irányt jelöl-e;
   (c) megvizsgáltuk, hogy a prepozíció a-val kezdődik-e, mivel etimológiailag az *a* prefixum mozgást jelöl (pl. *across*);
   (d) a prepozíció mondatbeli helye;
   (e) a prepozíció nyelvtani szerepe;
   (f) megvizsgáltuk, hogy a prepozíciónak van-e gyermek csomópontja a függőségi fában, és amennyiben volt, felvettük annak a nyelvtani szerepét is.
4. **Mondatszintű jellemzők**:
   (a) a mondat hossza;
   (b) külön jellemzőként jelöltük, ha az ige és a prepozíció egyaránt mozgást és irányt jelölt, mivel ezen kombinációk gyakran kompozicionális jelentésűek (pl. *go out*);
   (c) van-e tárgya az igének;
   (d) van-e névmási tárgya az igének;
   (e) van-e névmási alanya az igének.

Megjegyezzük, hogy a kérdésekre vonatkozó jellemzők és a prepozícióra vonatkozó jellemzők közül az utolsó három új, azaz tudomásunk szerint a VPC-k azonosítására vonatkozó hatásukat ezidáig még nem vizsgálták.

5.3. Eredmények

Az eredmények a 4. táblázatban láthatók.

4. táblázat. Gépi tanulási eredmények a Tu& Roth korpuszon.

SVM	**SVM kérdések nélkül**	**Tu & Roth**	**Nagy T. & Vincze**
80,05	77,46%	78,6	81,92

Tu és Roth eredeti cikkükben [14] 78,6%-os pontosságot értek el, Nagy T. és Vincze [13] pedig 81,92%-ot ugyanezen a korpuszon mérve. A jelen cikkben közölt eredmények meghaladják [14] eredményeit, azonban [13] eredményeinél valamivel

alacsonyabbak, itt elsődlegesen azonban a kérdésekre épülő jellemzők hozzáadott értékére voltunk kíváncsiak.

Az új jellemzők bevezetése hozzájárult a rendszer jó teljesítményéhez. Modellünket újratanítottuk pusztán a szakirodalomban már korábban is használt jellemzőket hasznosítva, azaz az általunk bevezetett új jellemzőket mellőztük. Így 77,46%-os pontosságot, azaz 3,81 százalékponttal alacsonyabb teljesítményt értünk el. E kísérletünk is igazolja a kérdésekre épülő jellemzők hozzáadott értékét egy számítógépes nyelvészeti feladatban.

6. Automatikus válaszadás kérdésekre

A Yako alkalmazás fő célja, hogy egységes felületet biztosítson a telefonra beérkező üzeneteknek, legyen azok formája SMS, e-mail vagy pedig Messenger-üzenet [17]. Az érkező üzenetekre a felhasználó természetesen választ is írhat, szintén egy egységes felület segítségével. A felhasználó kényelmét biztosítandó az alkalmazás az üzenetben érkezett kérdésekre automatikus válaszlehetőségeket ajánl fel, melyek közül a felhasználó egy mozdulattal kiválaszthatja a szándékolt választ. E funkció eredetileg egyszerű eldöntendő kérdésekre működött, melyekre igennel vagy nemmel lehet válaszolni, továbbá olyan kérdésekre, ahol két lehetőség közül lehet választani.

E cikkben célunk, hogy a Yako által kezelhető kérdések körét bővítsük, azokhoz megfelelő nyelvi reprezentációt nyújtva. Ehhez megvizsgáltuk a magyar korpuszban előforduló eldöntendő kérdéseket, és a leggyakoribb szintaktikai és morfológiai mintázatokra építve felállítottunk néhány újabb lehetséges sémát az eldöntendő kérdésekre. E sémákat az 5. táblázatban foglaltuk össze, ahol a morfológiai információkat MSD-kódok formájában jelenítjük meg.

5. táblázat. Eldöntendő kérdések sémái és lehetséges válaszok.

Séma	**Válaszjavaslat**	**Példa**
Rp + Va* + Vmn	igen/Rp + nem	El akarsz jönni? Igen/El. Nem.
Rm + Vm*	de + nem	Nem jössz velünk? De. Nem.
Rm + Va* + Vmn	de + nem	Nem akarsz eljönni? De. Nem.
ugye	igen + nem	Ugye eljössz az MSZNY-re? Igen. Nem.
vajon	igen + nem	Vajon eljön a karácsony? Igen. Nem.
-e	igen + nem	Eljön-e az MSZNY-re? Igen. Nem.
, nem? (mondat végén)	de + nem	Eljössz az MSZNY-re, nem? De. Nem.
Vm* (mondat elején)	igen + nem	Eljössz az MSZNY-re? Igen. Nem.
$N*_1$ vagy $N*_2$	$N*_1$ + $N*_2$	Sört vagy bort? Sört. Bort.
$N*_1$, $N*_2$ vagy $N*_3$	$N*_1$ + $N*_2$ + $N*_3$	Sört, bort vagy kólát? Sört. Bort. Kólát.

Reményeink szerint ezek az újonnan felállított sémák hozzájárulnak a Yako alkalmazás további tökéletesítéséhez.

7. Összegzés

Ebben a cikkben angol és magyar nyelvű kérdések számítógépes nyelvészeti elemzése és felhasználása felé tettük meg az első lépéseket. Először korpuszalapú vizsgálatok segítségével feltérképeztük az egyes kérdéstípusok és kérdőszavak gyakoriságát, majd a gyakorlati hasznosítás felől közelítettük meg a kérdéskört. Gépi tanulási kísérletekkel kimutattuk, hogy az angol prepozíciós igék automatikus azonosítására pozitív hatással bírnak a kérdésekre irányuló nyelvi jellemzők: az alaprendszerhez képest 3,8 százalékpontos javulást értünk el pontosság terén. Emellett a Yako alkalmazás egyik funkciójának – az automatikus válaszadási lehetőségeknek – továbbfejlesztésére tettünk javaslatot, a korpuszban található eldöntendő kérdések nyelvi jellemzőinek feltérképezése segítségével, ezzel is igazolva a kérdések szerepének fontosságát számítógépes nyelvészeti alkalmazásokban.

A jövőben célunk, hogy a közösségi médiában előforduló kérdések interperszonális és pragmatikai funkcióit részletesebben is feltérképezzük, illetve azok számítógépes nyelvészeti hasznosíthatóságát is megvizsgáljuk. Mindemellett szeretnénk a kérdések nyelvészeti jellemzőit más számítógépes nyelvészeti alkalmazásokba is beépíteni és azok hasznosságát megvizsgálni.

Köszönetnyilvánítás

A jelen kutatás a PARSEME COST Action (IC1207) projekt keretében az Európai Unió támogatásával valósult meg.

Hivatkozások

1. Groenendijk, J., Stokhof, M.: Questions. In van Benthem, J., ter Meulen, A., eds.: Handbook of Logic and Language, Amsterdam/Cambridge, MA, Elsevier/MIT Press (1997) 1055–124
2. Graesser, A.C., Person, N.K., Huber, J.D.: Mechanisms that generate questions. In Lauer, T.E., Peacock, E., Graesser, A.C., eds.: Questions and information systems, Hillsdale, NJ, Lawrence Erlbaum Associates (1992) 167–187
3. Dikken, M.d.: On the morphosyntax of wh-movement. In Boeckx, C., Grohmann, K., eds.: Multiple wh-fronting, Amsterdam, John Benjamins (2003) 77–98
4. Nivre, J., Bosco, C., Choi, J., de Marneffe, M.C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., Zeman, D.: Universal dependencies 1.0 (2015)
5. Judge, J., Cahill, A., Genabith, J.V.: Questionbank: Creating a corpus of parse-annotated questions. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06). (2006) 497–504

6. Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V.: The French Social Media Bank: a treebank of noisy user generated content. In: Proceedings of COLING 2012, Mumbai, India, The COLING 2012 Organizing Committee (2012) 2441–2458
7. Mott, J., Bies, A., Laury, J., Warner, C.: Bracketing Webtext: An Addendum to Penn Treebank II Guidelines. Linguistic Data Consortium (2012)
8. Bies, A., Mott, J., Warner, C., Kulick, S.: English Web Treebank. Technical report, Linguistic Data Consortium, Philadelphia (2012) LDC2012T13.
9. Bell, A.: The language of the News Media. Blackwell, Oxford (1991)
10. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
11. Vincze, V., Varga, V., Papp, P.A., Simkó, K.I., Zsibrita, J., Farkas, R.: Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, Szegedi Tudományegyetem (2015) 122–132
12. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002, Mexico City, Mexico (2002) 1–15
13. Nagy T., I., Vincze, V.: VPCTagger: Detecting Verb-Particle Constructions With Syntax-Based Methods. In: Proceedings of the 10th Workshop on Multiword Expressions (MWE), Gothenburg, Sweden, Association for Computational Linguistics (2014) 17–25
14. Tu, Y., Roth, D.: Sorting out the Most Confusing English Phrasal Verbs. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. SemEval '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 65–69
15. Cortes, C., Vapnik, V.: Support-vector networks. Volume 20. Kluwer Academic Publishers (1995)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
17. Farkas, R., Kojedzinszky, T., Zsibrita, J., Wieszner, V.: Yako: egy intelligens üzenetváltó alkalmazás nyelvtechnológiai kihívásai. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2015) 323–325

Aspektusszintű annotáció és szentimentet módosító elemek egy magyar nyelvű szentimentkorpuszban

Szabó Martina Katalin[1,2], Vincze Veronika[3], Hangya Viktor[4]

[1]Precognox Informatikai kft.
[2]Szegedi Tudományegyetem, Orosz Filológiai Tanszék
mszabo@precognox.com; szabo.martina@lit.u-szeged.hu
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu
[4]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
hangyav@inf.u-szeged.hu

Kivonat: A dolgozat egy magyar nyelvű kézzel annotált szentimentkorpusz annotálásának tapasztalatait, valamint hasznosításának első eredményeit mutatja be. A dolgozat célja kettős: Egyrészt, bemutatjuk mindazokat a módosításokat, amelyeket a korpusz második annotálási fázisában alkalmaztunk az első annotálási fázis megoldásaihoz képest, a későbbi hasznosíthatóság javítása érdekében. Másrészt, megmutatjuk a korpusz első vizsgálati és felhasználási eredményeit. A vizsgálat keretében elsősorban a különböző annotált elemek fragmentumbeli, valamint szófajok közötti megoszlását tártuk fel. A dolgozatban beszámolunk a korpusz kvantitatív, illetve kvalitatív sajátságainak néhány figyelemre méltó összefüggéséről. Mindemellett az annotációra támaszkodva bizonyos elemcsoportokból lexikonokat generáltunk, amelyek reményeink szerint támogatni tudják majd az automatikus szentimentelemzés eredményességét. A dolgozat a generált lexikonok alapvető adatairól is számot ad.

1. Bevezetés

A dolgozatban a magyar nyelvű kézzel annotált szentimentkorpuszunk annotálásának tapasztalatait, valamint hasznosításának első eredményeit mutatjuk be.

Egy korábbi dolgozatunkban beszámoltunk egy magyar nyelvű szentimentkorpuszról, amelyet termékvélemény-szövegekből hoztunk létre kutatási és fejlesztési céllal [1]. A korpusz szöveganyagát a [http://divany.hu/] honlap termékvéleményeiből állítottuk össze. A honlap készítői időközönként bizonyos termékcsoportokat tesztelnek, s közzéteszik a tesztelők véleményét. (1) alatt közlünk két példát a nyers korpuszból.

(1) a. Elképesztően kellemetlen íze van, előfordult, hogy valaki nem is ismerte fel benne a csokoládéízt. Legtöbbször az hangzott el, hogy mellékíze van, édes és bóvli termék. Sajnáljuk azokat a gyerekeket, akik ezt kapják. Igaz, a csomagolása elég kecsegtető, de ez ne tévesszen meg senkit!
[http://divany.hu/2013/karacsony/adventinaptarteszt]

b. Nem maradt kellemetlen, zsíros érzés a bőrömön használat után. Az illata nagyon kellemes, nem kell nagy mennyiség belőle és úgy érzem, jól tisztít. Nem szárítja a bőröm, így használata után nem is kell hidratálót felkennem, üde, jó érzés. Ráadásul nagyon olcsó, az 50 milliliteres mini kiszerelés csak 199 forint!
[http://divany.hu/tejbenvajban/2014/08/28/arclemoso_teszt_tej_vagy_gel]

A korpusz második annotálási fázisában átdolgoztuk az annotálási elveket, majd a korpusz teljes szöveganyagát feldolgoztuk kézzel.

A korpusz hiánypótló magyar nyelvű kutatási és fejlesztési eszköz. Tudomásunk szerint ez az egyetlen olyan magyar nyelvű szentimentkorpusz, amely fragmentum- és aspektusszintű annotációval rendelkezik, emellett az annotáció keretében számos, a szentimentelemzés szempontjából releváns nyelvi elem is önálló taget kapott (az annotációról, valamint annak elméleti alapvetéseiről részletesebben l. később, 2).

Az elkészült adatbázis összesen 154 véleményszöveget, 17 059 mondatot és 251 202 tokent (központozással) tartalmaz.

2. Az annotálási rendszer felülvizsgálata

2.1. Annotációs alapelvek

A korpusz első annotálási fázisában a teljes értékelő kifejezést (másképpen: szentimentfragmentumot), azon belül a pozitív és a negatív polaritású szentimentkifejezéseket, azok targetjeit, valamint esetleges siftereit jelöltük be a korpuszban [1].

A szentimentelemzésben a feldolgozási egységek terjedelme szerint alapvetően a következő három szintet különböztetjük meg: a dokumentum-, a mondat-, valamint az entitás- és aspektusszintű annotációt [2]. A dokumentumszintű elemzés (*Document-level Sentiment Classification*) célja, hogy a teljes szöveg viszonylatában megadja a benne megfogalmazott értékelés polaritását, tehát azt, hogy az adott szöveg összességében pozitív vagy negatív értékelést fejez-e ki. A mondatszintű elemzés (*Sentence-level Sentiment Classification*) a szöveg egyes mondatainak vonatkozásában kívánja meghatározni azok polaritását. A mondatszintű elemzés egy altípusának tekinthetjük az ún. tagmondat-szintű elemzést (*Clause-level Sentiment Classification*), amely az értékelés polaritását az egyes tagmondatok vonatkozásában határozza meg. Megállapítható, hogy bár a mondat-, illetve tagmondatszintű osztályozás kétségtelenül árnyaltabb feldolgozást tesz lehetővé a dokumentumszintű osztályozásnál, igazán hatékony szentimentelemzést egyik megoldással sem érhetünk el. Tudniillik, egyik fentebbi eljárásmóddal sem tárhatjuk fel azt, hogy a szövegben megfogalmazott értékelés pontosan mire, azaz mely targetre irányul, emellett további problémákat eredményez, ha az adott, egyetlen egységként kezelt szövegrész (dokumentum vagy mondat) több targetet és/vagy több szentimentet is tartalmaz.

A fentebb elmondottakkal összefüggésben, a leghatékonyabb szentimentelemzési megoldást az ún. entitás- és aspektusszintű feldolgozás (*Entity and Aspect-level Sentiment Classification*) kínálja. Ez a megoldás ahelyett, hogy a szöveg valamely szerkezeti egységét (dokumentum, bekezdés, mondat vagy tagmondat) venné az elemzés alapjául, kifejezetten magukra az értékelésekre koncentrál, és a feldolgozás alap-

egységét egy target, valamint az annak vonatkozásában kifejezett szentiment kapcsolatában határozza meg [2].

A bemutatott megközelítési módok hatékonyságát összevetve úgy döntöttünk, hogy korpuszunk annotálásában entitás- és aspektusszintű feldolgozást alkalmazunk. Ez a gyakorlatban azt jelentette, hogy egy egységként, azaz szentimentfragmentumként annotáltunk minden olyan szövegrészt, amely egyetlen meghatározott polaritású értékelést fejezett ki egyetlen (esetleg több, egymással mellérendelő szintaktikai viszonyban álló) target vonatkozásában. E megoldás következtében a fragmentumok az esetek túlnyomó többségében mondatnyi, vagy a mondatnál kisebb egységek, azonban ritkán az is előfordul, hogy egy fragmentum átível a mondat határán.

A szentimentfragmentum annotálását követően, azon belül több, a szentimentelemzés szempontjából releváns nyelvi elemet eltérő taggel láttunk el. Szentimentkifejezésként kezeltük azokat az egy szóból álló, vagy állandósult többszavas szókapcsolatokat, amelyek lexikai szinten értékítéletet hordoznak valamely target vonatkozásában [1]. Azokat a nyelvi elemeket, amelyek valamilyen módon hatást gyakorolnak a szövegekben megfogalmazott értékelő tartalmakra, az angol nyelvű terminológia alapján szentimentshiftereknek neveztük, és külön taggel láttuk el a korpuszban [2]. Ily módon a negáló, az irreáló, valamint a növelő és a csökkentő értelmű intenzifikáló elemek önálló jelölést kaptak az annotációban.

2.2. A második annotálási fázisban alkalmazott módosítások

A második annotálási szakaszban bizonyos mértékben eltérő annotálási megoldásokat alkalmaztunk a korábbi feldolgozási rendszerhez képest. Ennek oka az volt, hogy a megelőző munkafázis, valamint az első korpuszvizsgálatok tapasztalatai alapján úgy véltük, a targetek kezelésében kidolgozottabb, még részletesebb elemzést lehetővé tevő annotációra van szükség.

Amint azt a megelőző részben ismertettük (l. fentebb, 2.1), a korpusz első feldolgozási fázisában, a fragmentumszintű feldolgozásnak köszönhetően már a targetekkel összefüggésben kezeltük a szentimenteket. Ugyanakkor az entitás- és aspektusszintű annotációt nem tudtuk maradéktalanul megvalósítani, tekintettel arra, hogy az entitásokat, valamint azok egyes aspektusait nem különböztettük meg az annotáció szintjén egymástól. Az újabb annotációs alapelv szerint a korpuszban előforduló entitásokat, valamint azok különböző aspektusait eltérő annotációs taggel (target 1-20) láttuk el, következetesen alkalmazva azokat egy adott dokumentumon belül. A <target 1> címkével rendre az adott termék, azaz maga az entitás szövegbeli előfordulását jelöltük, míg a többi target-címkével annak különböző aspektusait annotáltuk. A címszerűen előforduló terméknevek jelölésére, akárcsak a korpusz első annotálási fázisában, a topic-címkét alkalmaztuk.

A bemutatott feldolgozási megoldás kettős haszonnal bír. Egyrészt, a módosítás eredményeképpen olyan annotáció jött létre, amely lehetőséget ad az entitás–aspektus-összefüggések gépi feltárására, segítve ezzel egy hatékonyabb automatikus szentimentelemző rendszer kidolgozását. Az entitások és az aspektusok megkülönböztetése a szentimentelemzés szempontjából kardinális probléma. Nagy jelentőséggel bír ugyanis, hogy egy értékelés az adott entitáshoz, vagy csupán annak valamely aspektu-

sához kapcsolódik-e, illetve, hogy a szövegben megjelenő aspektusok közül pontosan melyiket minősíti [2]. Tekintsük az alábbi példákat!

(2) a. Bár a *kiszolgálás* nem olyan jó, imádom ezt az *éttermet*.
b. A *kiszolgálás* nem olyan jó, a *berendezés* viszont barátságos.

Mindkét példa két ellentétes polaritású értékelést fogalmaz meg. Azonban a (2a) alatti példában a pozitív szentimentérték jelentősebb a negatív szentimentértéknél, tekintve, hogy az előbbi a teljes entitáshoz, az utóbbi pedig csupán annak egy aspektusához kapcsolódik. Ugyanakkor a (2b) alatti példában mind a pozitív, mind a negatív értékelés egy entitás egy-egy aspektusára vonatkozik.

A fentebb ismertetett hasznon túl, az új annotációs megoldás még a targeteket érintő koreferencia-viszonyok kezelését is biztosítja, hiszen egy dokumentumon belül egy adott entitást, valamint egy adott aspektust (és az esetlegesen azokat helyettesítő névmásokat) rendre ugyanazzal a target-taggel láttunk el.

3. A korpusz hasznosításának első eredményei

3.1. Vizsgálati eredmények

A korpusz első vizsgálatának keretében a különböző annotált elemek pozitív és negatív fragmentum-beli, valamint szófajok közötti megoszlását tártuk fel. A korpusz vizsgálatára azért volt szükség, hogy pontos képet kaphassunk a szentimentkifejezések használati sajátságairól a magyar nyelvű szövegek vonatkozásában.

Az alábbi táblázat összefoglalja a korpusz alapvető statisztikai adatait:

1. táblázat. Az annotáció alapvető statisztikai adatai

annotált elem	**összes előfordulás**	**pozitív fragmentumban**	**negatív fragmentumban**
PosSentiment	6693		
NegSentiment	8053		
SentiWordPos	7554	5800	1754
SentiWordNeg	7698	1272	6426
Topic	1365		
Target	7827	3731	4096
Negation	3342	1360	1982
IntensifierPlus	4973	2405	2568
IntensifierMinus	1080	301	779
Irreal	991	296	695
OtherShifter	1134	642	492
ÖSSZES:	**50747**	**15807**	**18792**

A fentebbi adatok alapján a következő legfontosabb megállapításokat tehetjük: A pozitív (SentiWordPos) és a negatív szentimentkifejezések (SentiWordNeg) majdhogynem egyenlő arányban képviseltetik magukat a korpuszban, ugyanakkor ez nem eredményezi azt, hogy a pozitív (PosSentiment) és a negatív fragmentumok (NegSentiment) is hasonló megoszlási arányt mutatnának. A negatív fragmentumok ugyanis többségben vannak a pozitív fragmentumokkal szemben. Az eredmény a shifterek szentimentelemzés-beni szerepére mutat rá, amelyek képesek a szentimentkifejezések lexikai szintű polaritásának akár teljes mértékű megváltoztatására.

A fentebbi eredményekből következően, a pozitív és a negatív szentimentkifejezések megoszlása a velük azonos polaritású fragmentumokban nem egyforma. Amíg ugyanis a negatív szentimentkifejezések 83,47%-a azonos polaritású fragmentumban szerepel, azaz csupán 16,52%-uk található pozitív fragmentumban, addig a pozitív szentimentkifejezések magasabb megoszlási aránnyal (23,21%) fordulnak elő a velük ellentétes polaritású fragmentumokban. A kapott eredmények alapján, egy negatív értékelést pozitív szentimentkifejezéssel (pl. *nem **jó***) gyakrabban fogalmazunk meg, mint egy pozitív értékelést negatív szentimentkifejezéssel (pl. *nem **rossz***). A tapasztalatok illeszkednek az ún. Pollyanna-hipotézishez, amely nyelvi univerzáléként tételezi a pozitív töltetű kifejezések magasabb használati arányát a negatív töltetű nyelvi elemekkel szemben [3].

Kíváncsiak voltunk, vajon mutatkozik-e valamilyen megoszlási eltérés a növelő és csökkentő értelmű intenzifikáló elemeket illetően a pozitív és a negatív fragmentumok között. Megvizsgálva a korpuszban annotált intenzifikáló elemeket megállapítottuk, hogy a növelő értelmű intenzifikáló elemek arányaikban közel azonos gyakorisággal fordulnak elő a pozitív (6693:2706) és a negatív (8053:3347) polaritású fragmentumokban (pl. ***nagyon** jó; **nagyon** rossz*). Ugyanakkor, a csökkentő értelmű intenzifikálók előfordulási gyakorisága nem azonos a két fragmentumtípusban: a negatív fragmentumokban (8053:779) jóval gyakrabban fordulnak elő a pozitív fragmentumoknál (6693:301). A jelenségnek valószínűleg az az oka, hogy amíg a növelő értelmű intenzifikáló elemek nem képesek a szentimentérték megváltoztatására, addig a csökkentő értelműek igen, és jelenlétük nem azonos szemantikai változást eredményez a pozitív és a negatív polaritású szentimentkifejezések mellett. Amíg ugyanis a pozitív polaritású szentimentkifejezések módosítójaként rendre negatív értékítéletet eredményez (pl. ***kevésbé** jó*), addig a negatív polaritású szentimentkifejezések mellett nem idéz elő feltétlenül polaritásváltást (pl. ***kevésbé** rossz*).

A csökkentő értelmű intenzifikáló elemeknél mutatkozó tendencia a negáló és az irreáló elemek esetében éppúgy megfigyelhető. Ez utóbbi elemek ugyanis szintén gyakoribb előfordulásúak a korpuszban a negatív polaritású fragmentumokban (pl. ***nem** jó; jó **volna***).

A kapott eredmények arra mutatnak, hogy a magyar nyelvű szövegek automatikus szentimentelemzése során különös figyelmet kell majd fordítanunk a pozitív szentimentkifejezések mellett megjelenő negáló, irreáló, valamint csökkentő értelmű intenzifikáló elemekre, hiszen azok gyakorta okoznak az esetükben polaritásváltást.

A szentimentértéket befolyásoló egyik leggyakoribb, ugyanakkor az egyik nehezebben kezelhető nyelvi jelenség a negáció. Automatikus feldolgozását két alapvető sajátság is nehezíti: Egyrészt, a negáló elemekkel alkotott szerkezetek gyakorta nem

kompozicionálisak [4, 5]. Bár gyakran, de nem minden esetben okoznak polaritásváltást, csupán polaritásváltozást. A polaritásváltás az a jelenség, amelyben a teljes, értékelést megfogalmazó fragmentum polaritása és a bennfoglalt puszta szentimentkifejezés lexikai szintű polaritása egymással ellentétes. Ezzel szemben polaritásváltozásnak nevezzük azt az esetet, amikor a fragmentum polaritása csupán bizonyos mértékű elmozdulást mutat a szentimentkifejezés lexikai szintű polaritásához képest, vagy neutralizálódik [6, 7]. A negáció másik sajátsága, amely megnehezíti annak kezelését, az az, hogy a negáló elem funkcióját a szövegekben számos kifejezés betöltheti (erről a későbbiekben részletesebben is lesz szó, l. lentebb).

Annak céljából, hogy pontosabb képet kaphassunk a negáló elemek szerepéről, megvizsgáltuk, hogy hány esetben okozzák ténylegesen a polaritás ellentétére változását. Megvizsgáltuk tehát minden negatív és pozitív szentimentkifejezés fragmentumbeli környezetét, a következő eredménnyel: A korpuszban előforduló 7554 pozitív polaritású szentimentkifejezés mellett összesen 1728 esetben fordul elő negáló elem, amely az esetek túlnyomó többségében, összesen 1554 kifejezésnél a polaritás megváltozását okozta. Hasonló eredményt kaptunk a negatív szentimentkifejezések vizsgálatával is. A negatív polaritású kifejezések közül összesen 1512 mellett jelenik meg a negáló elem, amely 1287 esetben a polaritás megváltozását idézte elő. Mindez arra mutat, hogy bár elméletileg a negáló elemek vagy teljes polaritásváltást, vagy csupán polaritásbeli elmozdulást idéznek elő, a gyakorlatban – legalábbis a korpuszunk adatai alapján – az előbbit tekinthetjük tendenciának.

A szófaji megoszlási arányok tekintetében különösen figyelemre méltónak találtuk a negáló elemek sajátságait. Az annotált korpusz szófaji szempontú vizsgálata előtt azt feltételeztük, hogy a negáló elemként alapvetően tagadószókat (pl. *ne, sem, dehogy*) és a létige tagadó alakjait (*nincs, nincsen, sincs, sincsen*), emellett kisebb számban tagadó névutókat (pl. *hiányában, nélkül*), valamint néhány egyéb módosítószót (pl. *aligha, látszatra*) fogunk annotálni [8, 9]. A korpuszvizsgálat eredményei azonban arra mutattak, hogy a negáló elemek nagyobb változatossággal szerepelnek kezdeti hipotézisünknél. A korpuszban annotált összesen 3516 negációs token legnagyobb részét határozószók (2587), igék (468), névmások (145), valamint mellérendelő kötőszók (93) adták.

Annak céljából, hogy érzékeltessük a negáló elemek változatosságát, a negátorként annotált nyelvi elemek közül (3) alatt közlünk néhány példát.

(3) hiánya, elillant, nélkülözi, bizarr lenne azt állítani, helyett, hiányoltam, semmi köze sincs, nulla, nem érezni benne, lespórolták

Megvizsgálva a csökkentő és a növelő értelmű intenzifikáló elemek szófaji megoszlását megállapítottuk, hogy amíg az előbbi funkcióját a legtöbbször melléknév tölti be mind a pozitív (31,44%), mind a negatív fragmentumokban (52,08%), addig az utóbbi funkciójában az esetek többségében (52,32% és 47,69%) határozószót találunk. Ugyanakkor, amint arra a fentebbi megoszlási arányok rámutatnak, a csökkentő értelmű intenzifikáló elemek esetében a szófaji megoszlási arányok nem azonosak a pozitív és a negatív polaritású fragmentumok között. A pozitív fragmentumokban ugyanis a különböző szófajok előfordulási arányai kiegyenlítettebbek a negatív fragmentu-

mokhoz képest, és bennük például a határozószók és a főnevek is jelentősebb szerephez jutnak a vizsgált funkcióban.

Az irreálók részletesebb vizsgálata során kiderült, hogy körülbelül kétharmaduk ige vagy határozószó. A határozószók esetében elsődlegesen a szójelentés hordozza az irreáló tartalmat (pl. *állítólag, valószínűleg, talán*), melyek párhuzamot mutatnak az úgynevezett episztemikus bizonytalanságot jelölő lexikai elemekkel [5]. Az igéknél azonban kettősséget figyelhetünk meg: egyrészt magának az igének a jelentése hordozza a bizonytalanságot (pl. *tűnik, hasonlít, imitál*) vagy szubjektivitást (pl. *érez, gondol*), másrészt morfológiai eszközök segítségével érhető el az irreáló tartalom. Ilyen például a lehetőséget kifejező -hAt toldalék, mely 115 esetben (11%) fordult elő a korpusz irreáló elemei között, és a feltételes mód, mely 291-szer (28%) jelent meg. A feltételes módhoz gyakran kapcsolódó *ha* és *mintha* kötőszavak is gyakori irreáló elemek a korpuszban, összesen 228 esetben (22%) találkozhattunk velük. A korpuszunkban megjelenő, nem a valóságot leíró nyelvi elemek szoros átfedést mutatnak az [5]-ben felvázolt, nyelvi bizonytalanságot jelölő elemekkel, így ezek összevető elemzésére a későbbiekben mindenképpen szeretnénk sort keríteni.

3.2. Felhasználási eredmények

Az annotációra támaszkodva bizonyos elemcsoportokból lexikonokat generáltunk, amelyek reményeink szerint javítani tudnak majd az automatikus szentimentelemzés eredményességén. A munka keretében pozitív és negatív szentimentlexikont, entitás- és aspektusszótárat, továbbá a különböző szentimentshifterek szótárait generáltuk, melyek alapvető statisztikai adatait az alábbi táblázat mutatja be:

2. táblázat. A korpuszból generált lexikonok alapvető statisztikai adatai

lexikon	**elemszám**
Pozitív szentimentkifejezések	2568
Negatív szentimentkifejezések	3343
Entitások szótára	2773
Aspektusok szótára	4781
Negáló elemek	95
Növelő értelmű intenzifikálók	744
Csökkentő értelmű intenzifikálók	199
Irreálók	195

A szentimentkifejezésekből generált lexikonok eredményességét szeretnénk összevetni a korábban, nem korpuszalapon készített szótáraink hatékonyságával is a jövőben.

Az entitások és az aspektusok szótárait a második annotálási szakaszhoz kidolgozott target-annotációnak köszönhetően volt lehetőségünk generálni, hiszen – ahogyan azt már korábban ismertettük (l. 2.2) – a korpuszban előforduló entitásokat, valamint azok különböző aspektusait eltérő annotációs taggel (target 1-20) láttuk el. Úgy véljük, ezek a szótárak különleges segítséget nyújthatnak majd hasonló, termékekkel kapcsolatban megfogalmazott véleményszövegek automatikus szentimentelemzése során.

4. Az eredmények felhasználási lehetőségei

A dolgozatban a magyar nyelvű kézzel annotált szentimentkorpuszunk annotálásának tapasztalatait, valamint hasznosításának első eredményeit mutattuk be.

A korpusz második annotálási fázisában átdolgoztuk az annotálási elveket, majd a korpusz teljes szöveganyagát feldolgoztuk kézzel. A dolgozatban bemutattuk mindazokat a módosításokat, amelyeket a korpusz második annotálási fázisában alkalmaztunk az első annotálási fázis megoldásaihoz képest, a későbbi hasznosíthatóság javítása érdekében. Ezt követően részletesen tárgyaltuk a korpusz első vizsgálati és felhasználási eredményeit. A vizsgálat keretében elsősorban a különböző annotált elemek fragmentum-beli, valamint szófajok közötti megoszlását tártuk fel. Emellett az annotációra támaszkodva bizonyos elemcsoportokból lexikonokat generáltunk, amelyek reményeink szerint támogatni tudják majd az automatikus feldolgozói munka eredményességét. A dolgozatban a generált lexikonok alapvető statisztikai adatairól is számot adtunk.

A korpusz hozadékaira támaszkodva terveink között szerepel egy olyan automatikus elemző rendszer létrehozása, amely képes a szentimentkifejezéseket azok targetjeivel és shiftereivel összefüggésben hatékonyan kezelni.

Hivatkozások

1. Szabó M.K., Vincze V.: Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács A., Varga V., Vincze V. (szerk.).: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015), Szeged, Szegedi Tudományegyetem (2015) 219–226
2. Liu, B.: Sentiment Analysis and Opinion Mining. Draft (2012) Elérhető: http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
3. Boucher, J., Osgood, C.: The Pollyanna hypothesis. Journal of Verbal and Learning Behavior 8(1) (1969) 1–8
4. Israel, M.: The pragmatics of polarity. In: Horn, L., Ward, G. (szerk.).: The Handbook of Pragmatics, Oxford, Blackwell (2004) 701–723
5. Feldman, R., Rozenfeld, B., Breakstone, M.Y. SSA: A Hybrid Approach to Sentiment Analysis of Stocks (2010) Elérhető: http://web.mit.edu/michab/www/ISCOL_Paper.pdf

6. Szabó M. K.: A polaritásváltás és -változás kezelési lehetőségei a szentimentelemzésben. In: Tavaszi Szél Konferencia, Eger (2015a) Megjelenés előtt
7. Szabó M. K.: A nyelvi értékelés mibenléte a számítógépes értékeléselemzés (szentimentelemzés) szempontjából. In: Lingdok – Nyelvész-doktoranduszok dolgozatai, Szeged (2015b) Megjelenés előtt
8. Szabó M. K., Vincze V.: A negáló szentimentshifterek kezelési kérdései a magyar nyelvű szövegek szentimentelemzésében. In: XXV. Magyar Alkalmazott Nyelvészeti Kongresszus előadásai (2015) Publikálásra benyújtva
9. Pete, I.: Az állító és tagadó mondatok szinonimiája a magyarban. Magyar Nyelv 95/3. (1999) 305–312
10. Vincze, V.: Uncertainty Detection in Hungarian Texts. In: Proceedings of COLING 2014, Dublin (2014) 1844–1853

Forrás

http://divany.hu/

Az érzelmek beszédre gyakorolt hatása, azaz a spontán beszéd szintaxisának érzelmekkel való kapcsolata a HuComTech Korpuszban

Kiss Hermina

Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
4032 Debrecen, Egyetem tér. 1.
kissh3@gmail.com

Kivonat: A HuComTech multimodális adatbázis 55 órányi spontán beszéd korpusza ötféle nyelvi szinten, hétféle annotációs sémában ad lehetőséget az elemzésekre. Az első MSZNY Konferencia előadásomban (2011) a szintaktikai annotációs szint szabályrendszerét mutattam be. A második alkalommal (2014) ennek a szintnek a felhasználásával és a multimodalitás segítségével részleteztem az addigi feltevéseinket és eredményeinket a gesztusok (deiktikus gesztusok) és a mondattan (azon belül a tárgy megjelenése a tagmondatokban) vonatkozásában.
Ezúttal egy másik szempontrendszer felhasználásával olyan felvetéseket szeretnék bemutatni, amelyek az alapvető érzelmek [1] és a beszéd közötti kapcsolatot mutatják be. A korpusz informális beszélgetéseinek egyik fontos jellemzője az, hogy szándékosan olyan kérdések hangzanak el és olyan sorrendben, amelyek érzelmi hatást váltanak ki, elgondolkodtatják és visszaemlékezésre ösztönzik a válaszoló riportalanyt. A beszélők a kérdésekre válaszolva történeteket mesélnek el az életükről, a világról. Az érzelmek és a gondolkodás kapcsolatának témakörében már számos eredmény született [2] [3], de konkrét kísérletekkel alátámasztott kutatás az érzelmek és a szintaktikai szerkezetek vonatkozásában tudomásunk szerint eddig még nem jelent meg. A vizsgálatainkhoz ennek megfelelően az érzelemkutatásnak azt az ágát tudjuk felhasználni, ami illeszkedik a korpusz jellemzőihez, azaz beemelhetjük az emlékezetkutatás bizonyos szempontjait az elemzéseinkbe. Tehát a meglévő korpusz tulajdonságait használjuk fel arra, hogy igazoljuk vagy cáfoljuk a feltevéseinket.

1. Bevezetés

A HuComTech Korpusz a multimodális jegyei alapján kiválóan alkalmas az érzelmek és a beszéd számítógépes nyelvészeti megközelítéséhez. A multimodális korpuszban megfigyelhetjük a különböző annotálási kategóriák együttjárását, elvégezhetjük az érzelmek és a beszéd együttes előfordulásának mondattani vizsgálatát [4]. Köztudott, hogy az érzelmek és a lelkiállapot megjelenik a nyelvhasználatban és a beszédben. A gesztusok, a kéztartás, a testtartás a testbeszéd részeként szorosan köthető az érzelmekhez. A gesztusokkal, kéz- és testtartással azonos időben a beszédet kísérhetik hezitációk, szünetek, megakadások, hiányos grammatika, amelyek által közvetve kö-

vetkeztethetünk arra, hogy hogyan befolyásolják az érzelmek a beszélő tudati működését.

Jelen tanulmányban az empirikus adatok közlése a legfőbb cél. Az adatok részletes elemzése részben a kötött terjedelem miatt nem lehetséges, részben pedig azért, mert az adatok feltárása és összesítése olyan további kérdéseket vetnek föl, amelynek megválaszolásához újabb mérési eredmények szükségesek. Eredményeinknek leginkább a kinyert információk által felmerülő kérdéseket tekintjük, amelyek irányt adnak a további mérésekhez.

Jelen tanulmány két részre oszlik. Először bemutatom azokat a mondatokat, amelyekben érzelmi váltás következik be, majd megvizsgálom a befejezetlen mondatok és az érzelmek viszonyát.

2. Az érzelmek és a kogníció kapcsolata

Megfigyeléseinkben és kísérleteinkben ahhoz a nyelvi produkciót vizsgáló kutatási hagyományhoz csatlakoztunk, amely a hibák felől közelíti meg a nyelvi folyamatokat, a kognitív és érzelmi feldolgozás nyelvi eredményeit [5].

Az érzelmek és a nyelv kapcsolatát vizsgálva felmerül az a kérdés, hogy „vajon az érzelem a kognitív-reprezentációs rendszer integrált részének tekinthető-e, vagy pedig önálló és bizonyos tekintetben mindentől független primér válaszrendszernek" [6]. A kutatók egy része amellett érvel, hogy az érzelmi rendszer autonóm és independens egységet alkot, és az érzelmi reakciók sok esetben megelőzik a kognitív folyamatokat. A kutatók másik része az egymással kölcsönhatásban lévő rendszerekként vizsgálja az érzelmi és kognitív rendszert. A kétféle nézetet leginkább a kogníció fogalmának értelmezési tartomány határozza meg [7]. Kísérleteinkben azokhoz a kutatókhoz csatlakozunk, akik az érzelmi-kognitív modellt egységes, integrált és egymással kölcsönhatásban lévő reprezentációs rendszernek fogják fel. Az érzelmek és a beszélt nyelv vizsgálata során arra törekszünk, hogy olyan kognitív-érzelmi modellt állítsunk fel, amelyben központi szerepet kap az a gondolat, hogy az egyes érzelmi állapotok aktiválódása hozzájárul bizonyos kognitív tartalmak létrejöttéhez.

Beszédprodukciós modellek létrehozói szerint a nyelvi szintek egyik alkotóeleme a mondat szintje. Például Dell (1986) beszédprodukció-elmélete szerint a beszéd tervezése négy szinten, négy szakaszon keresztül történik: szemantikai, mondattani, morfológiai, fonológiai. Dell is egyetért abban, hogy a gondolkodás egyik szegmense a grammatikai gondolkodás [8]. Ezt és a grammatikai gondolkodást elfogadó más elméleteket alapul véve számos kérdést tehetünk fel az érzelmek és a beszéd kapcsolatáról. Ezzel megerősítünk klasszikus eredményeket, másrészt pedig új következtetéseket is levonhatunk.

„A mai napig nincs pontos ismeretünk arról, hogy az érzelmi és a kognitív folyamatok hogyan befolyásolják a beszédtervezés és -kivitelezés folyamatát" [9]. Erre a hiányzó információra szeretnénk a figyelmet felhívni, és olyan kutatási lehetőségeket felvázolni, ami hozzájárulhat ilyen irányú ismereteink bővítéséhez.

A HuComTech Korpusz adatainak megvizsgálása után talán lehetővé válik annak a sejtésnek a közelebbi vizsgálata, amely szerint az érzelmi szint a beszédtervezés egyik

szintje. Mindezt annak alapján feltételezhetjük, hogy „a beszédtervezés gyakorlatilag bármely szintje hatással lehet a kivitelezésre, és módosíthatja a létrehozott közlést" [10]. Ha ez így van, és az érzelmi szint is hatással van legalább a szintaktikai gondolkodásra, akkor egy szempontnak már megfelelne ez a sejtés.

3. Az érzelem fogalma

Ekman úgy véli, hogy egy kialakult érzelem az első ezredmásodpercekben uralkodik a cselekedeteinken, gondolatainkon, beszédünkön [11]. Ennek alapján azt feltételezzük, hogy az érzelmek első ezredmásodpercei meghatározzák a további másodpercek kognitív döntéseit, és arra keressük a választ, hogy van-e kapcsolat az érzelmek és a mondatok (mondatszerkezetek) között. Fontos megjegyezni, hogy mi nem a kialakulásuk felől figyeljük az emocionális folyamatokat, hanem az érzelmi folyamatok kialakulásának és létrejöttének azt az időbeli szakaszát vizsgáljuk, amely egy másik egyén számára külső formális jegyek alapján jól észlelhető (esetünkben: jól látható). Egy érzelem kialakulása és láthatóvá válása közötti időszakaszt figyelmen kívül hagyjuk, és arra az időintervallumra koncentrálunk, amikor az adott érzelem már kivehető az arcon, illetve a testrészeken: kézen, testtartáson, fejmozgáson. Az érzelem fogalmának tisztázásához nagyban hozzájárulnak a HuComTech Korpusz adottságai, hiszen a formális és informális társalgás érzelmeit az annotátorok képkockánként (0,4 másodperc) értelmezték. Olyan érzelmeket (vagy hangulatokat) vizsgálunk tehát, amelyek percepcionálisan felfogható, minimum 0,4 másodpercig tartó információk. Az érzelmeknek nem a jellegére koncentrálunk (érzés, érzelem, hangulat és attitűd nem különül el), hanem az arc és a testrészek adott időtartam alatt jól kivehető és értelmezhető változásaira.

Ekman alapvető érzelmeit (szomorúság, fájdalom, düh, undor, megvetés, kellemes érzések) alapul véve [12] a következő érzelmi címkéket használtuk: *happy, sad, tense, recall, surprise, natural, other*. Mindezek fokozatokkal is elláthatók voltak az annotátorok számára: *strong, moderate, reduced*.

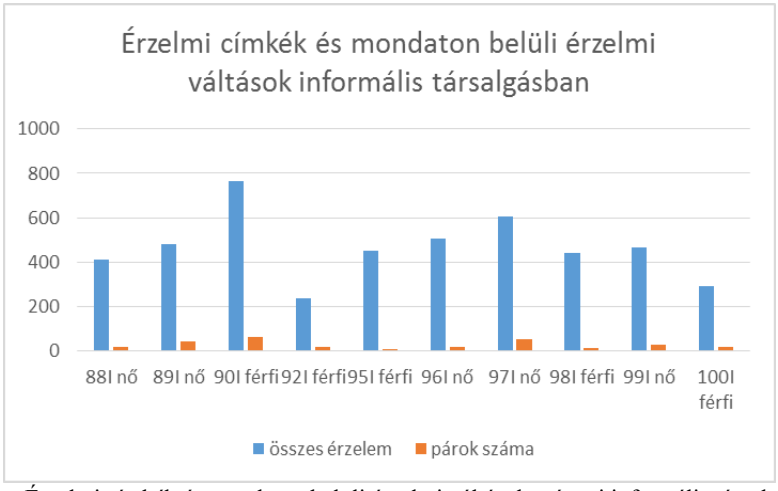
4. Az érzelmek dinamikája a mondatban

Az érzelmi állapotok változatosságának megragadása a mondatokban azért fontos, mert az érzelmi állapotok kifejezésének vizsgálata a kommunikáció funkcióinak árnyaltabb meghatározására és az érzelmek társalágásban betöltött szerepének részletesebb megfigyelésére ad lehetőséget.

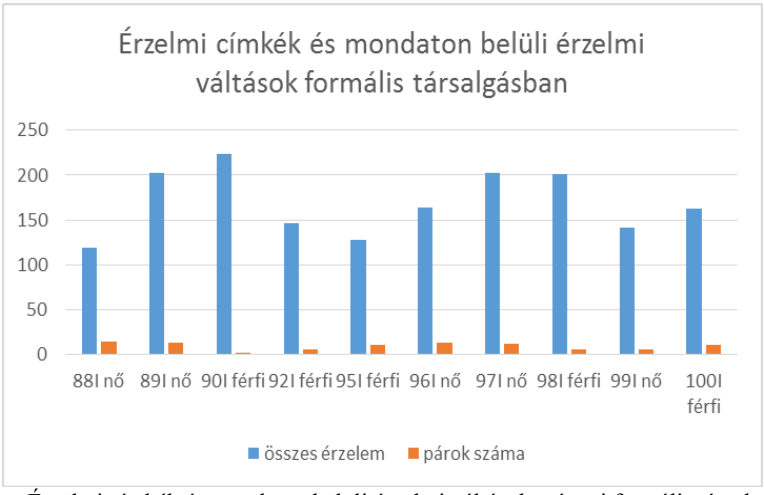
Az érzelmek dinamikájának feltárása céljából azt az adathalmazt vizsgáltuk meg, amely azt mutatja meg, hogy egy mondatban (ami temészetesen több, akár 15-20 tagmondatból is állhat), hányszor következik be érzelmi váltás. Megfigyeléseinkben 10 darab formális és 10 darab informális beszélgetés video anyagát használtuk fel.

Azokat az eredményeket listáztuk, amikor két érzelem váltotta egymást egy mondatban. A három, négy, öt, hat érzelem váltakozása nincs jelölve a táblázatokban.

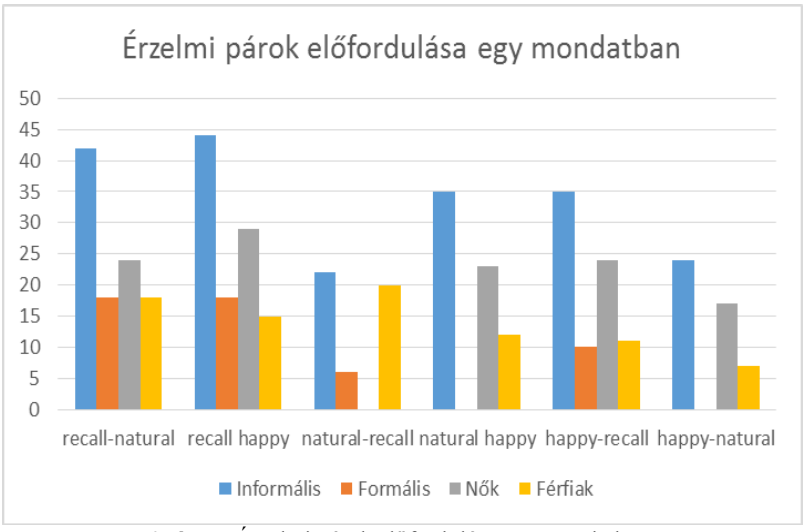
Megfigyelhető, hogy leggyakoribb érzelmek variálódnak és váltakoznak a mondatokban, ám nem a gyakoriságuk sorrendjében. A természetes intuíció szerint talán a leggyakoribb érzelmi váltás az, amikor a leggyakoribb érzelem vált át a második leggyakoribb érzelembe, második leggyakoribb váltás az, amikor a leggyakoribb érzelem vált át a harmadik leggyakoribb érzelembe, a harmadik leggyakoribb váltás pedig az, amikor a második leggyakoribb érzelem vált át a leggyakoribb érzelembe, és így tovább.



1. ábra: Érzelmi címkék és mondaton belüli érzelmi váltások arányai informális társalgásban.



2. ábra: Érzelmi címkék és mondaton belüli érzelmi váltások arányai formális társalgásban.



3. ábra: Érzelmi párok előfordulása egy mondatban.

Az 1. és a 2. ábrán látható, hogy azok a mondatok, amelyeken csupán egyetlen érzelem halad át, értelmezhetően jelentősebb számban vannak jelen, mint azok a mondatok, amelyekben érzelmi váltás történik. Az összes érzelmi címkéhez képest értelmezhetően kevés tehát azoknak a mondatoknak a száma, amelyben két (vagy több) érzelem található egymás után. Mégis ezeket a mondatokat találtuk érdekesnek ahhoz, hogy megállapításokat tehessünk a mondatok és az érzelmek viszonyáról, mert ezekben a mondatokban a grammatikai folyamat és az érzelmi váltakozások folyamata egyazon időszakaszon belül történik.

A 3. ábrán azt láthatjuk, hogy a kötetlen beszélgetésekben van értelmezhetően a legtöbb érzelmi váltás egy mondaton belül. Az informális beszélgetésekben többféle váltás szerepel és több előfordulásban is, mint a formális beszélgetésekben. A nők esetében pedig értelmezhetően több érzelmi váltás figyelhető meg, mint a férfiaknál. Ezek az adatok a természetes intuíciót támasztják alá: az érzelmek nem mechanikusan követik egymást, tehát nem gyakoriság alapján szekvenciálódnak az érzelmi váltások a mondatokban. Ennek oka lehet a szemantika, a kontextus, vagy a kommunikatív helyzet, amit tükröz az adott mondat. A mondat tehát a kommunikációs helyzetben való viselkedés és a kommunikációs helyzethez való viszonyulás tükröződése, nem pedig egy előre kiszámítható folyamat eredménye.

A beszéd és az érzelmek kapcsolatát keresve kijelenthetjük, hogy a mondatokon belüli érzelmi váltások különbséget mutatnak formális és informális, valamint férfiak és nők esetében. Arra nézve azonban nincs statisztikai szignifikancia, hogy az egyes beszélőknél mennyire gyakoriak az adott érzelmek. Arra nézve sincs statisztikai szignifikancia, hogy beszélőnként milyen gyakoriak és mik az érzelmi váltások. Az érzelemváltást létrehozó tényezők nem az érzelmi előfordulások gyakoriságának megfelelően változnak, azaz az érzelmi váltások nem automatikusan generálódnak az érzelmek számához igazodva. Ennek egyik oka az lehet, hogy nem várható a beszélőktől, hogy elvárt módon reagáljanak a feltett kérdésekre, mert egyéni, hogy milyen részletességgel, milyen átéléssel beszél a szubjektum az élete eseményeiről.

5. Az érzelmek és a befejezetlen mondatok kapcsolata

Azt is megvizsgáltuk, hogy egy befejezetlen tagmondat vagy mondat milyen kapcsolatban áll azzal az érzelemmel, amelyet a befejezetlen tagmondat vagy mondat kimondásának időintervallumában jelöltek az annotátorok.

Gósy Mária szerint „nemegyszer kideríthetetlen az az ok vagy oksorozat (pl. fáradtság, figyelmetlenség is), amelynek megakadás az eredménye" [13]. Az ilyen kideríthetetlen okok érdekelnek minket, habár Gósy a fáradtság és figyelmetlenség mellett idézett tanulmányában nem emeli az érzelmi okokat, ami fontos lehet a megakadások hátterének vizsgálata során.

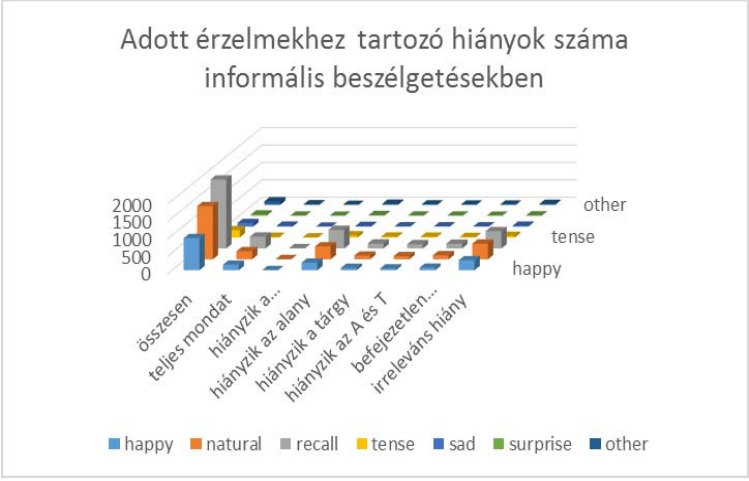
Befejezetlen mondatnak tekintjük azokat a tagmondatokat és mondatokat, amelyek elkezdődnek és valamilyen okból nem fejeződnek be. A megakadásjelenségek önkontrolljának négy lehetősége van: korrigálta a hibát vagy nem korrigálta, valamint észrevette vagy nem vette észre azt [14]. Önkontroll szerint csoportosítva a mondatok befejezetlenségét a következő kritériumok alapján lehet befejezetlen tagmondatnak tekintni az adott tagmondatot a HuComTech Korpuszban: ha a beszélő úgy fejez be egy mondatot vagy tagmondatot, hogy nem történik korrekció, vagy ha úgy korrigálja beszélő a hibát, hogy morfológiai változtatás történik. Ebben az esetben a korrekció előtt befejezetlen mondatként jelölve lezárul a mondat.

A befejezetlen tagmondatok, illetve befejezetlen mondatok érzelmi hátterének vizsgálatához a következő kérdéseket vetjük fel. Az intonáció vagy az annotátorok által mondatvégi írásjellel lezárt, teljes vagy (grammatikai sérülést nem okozó) hiányt tartalmazó tagmondatok és mondatok (*teljes mondat, hiányzik a főmondat, hiányzik az alany, hiányzik a tárgy, együtt hiányzik az alany és a tárgy, a hiány keresése irreleváns, tehát nem állapítható meg hiány a tagmondatban*), valamint a befejezetlen tagmondatok vagy mondatok esetén különbséget vártam az érzelmekkel való kapcsolatuk tekintetében. A befejezetlen mondat elvileg tartalmazhat több (vagy intenzívebb) érzelmi címkét, mint a lezárt (tag)mondatok. (Az érzelmek intenzitásának változását jelen tanulmányban nem vizsgálom.) Befejezetlen (tag)mondat érzelmi blokkja a címkék alapján lehet zavar, zavartság (*tense*), meglepettség (*surprise*), váratlan, hirtelen öröm (*happy*), váratlan, hirtelen szomorúság, levertség *(sad*).

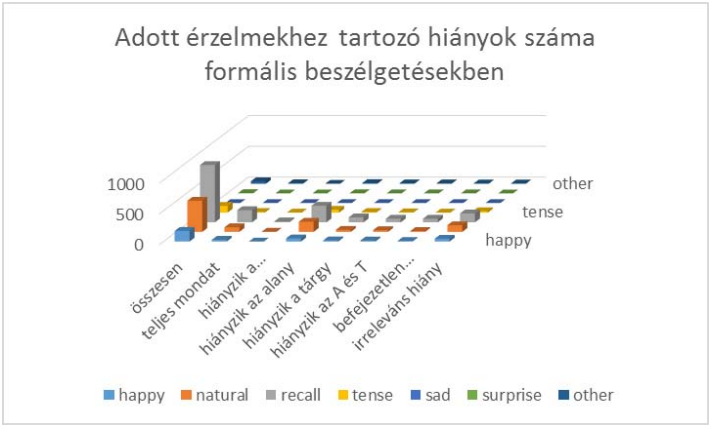
A hiányokat és az érzelmeket a formális és informális, valamint a férfi és nő közötti különbségek tekintetében vizsgáltuk meg. A 4. és 5. ábrán az érzelmek és a hiányok viszonya látható az informális és formális beszéletésekben. Mindkét ábrán jól látható, hogy a hatfajta érzelem és az other kategória eloszlása a mondatokban nem egyenletes. Számuk típusonként nem azonos, hanem három annotált érzelmi típus értelmezhetően magas számú. A legmagasabb a *recall*, azt követi a *natural*, a harmadik helyen pedig a *happy* kategória áll. Ennek egyik oka lehet az, hogy a riport során a beszélőknek folyamatosan figyelniük kellett a kérdezőre, hiszen a riporter aktívan beleszólt a beszélgetésbe, folyamatos kérdésfeltevésel irányította azt, valamint ő is sok történetet elmesélt. Ez a *recall* címkék számát jelentősen megnövelhette. A *natural* címke gyakorisága azt mutatja meg, hogy az annotátorok számára sok esetben nem volt leolvasható érzelem az arcokon. A mosoly a társalgás szándékára és fenntartásának jelölésére az egyik legfontosabb eszköz. Nem fordulhatott elő, hogy az annotátorok *happy* kategóriát állapítottak meg olyan esetekben is, amikor valaki zavartan mosolygott, vagy a mosoly mögött valójában nem vidámság húzódott meg. (Fontos megjegyezni, hogy a

jó és a rossz hangulatot feltételezően kiváltó kérdések aránya azonos volt.) Ekman szerint „a mosoly, ha csak finoman is, de félreérthetetlenül elárulja nekünk, hogy öröm váltja-e ki” [15]. Ha bízunk az annotátorok intuíciójában, akkor az öröm kategóriája nem keveredett össze más érzelmi állapotokkal.

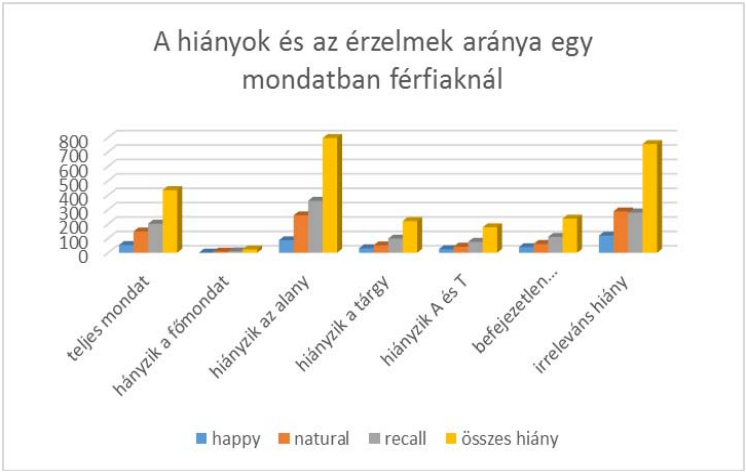
A többi érzelem a *natural, happy, recall* címkéknél értelmezhetően alacsonyabb számban jelenik meg. A riportok készítésének egyik fő célja volt a meglepetés okozása, zavarba hozás, a szomorúság és az öröm kiváltása. Az adatok alapján ez a cél nem volt teljesen sikeres. A legtöbb érzelem az alanyhiányos mondatokra és az irreleváns hiánykategóriába eső tagmondatokra és mondatokra esik, mert ebből a mondattípusból van a legtöbb.



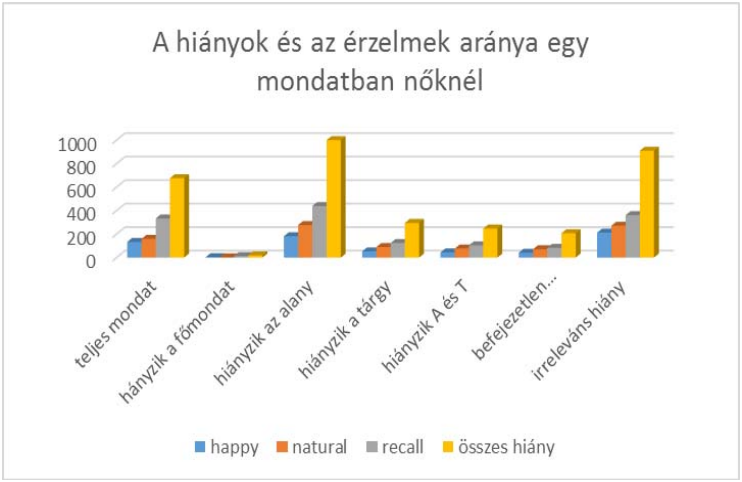
4. ábra: Adott érzelmekhez tartozó hiányok száma informális beszélgetésekben.



5. ábra: Adott érzelmekhez tartozó hiányok száma formális beszélgetésekben.



6. ábra: A hiányok és az érzelmek aránya egy mondatban férfiaknál.



7. ábra: A hiányok és az érzelmek aránya egy mondatban nőknél.

Az érzelmek és a hiányok viszonya a nők és férfiak között is különbségeket mutat (6., 7. ábra). Nők esetében nagyobb számban jelennek meg a hiányok és az érzelmek is. Szignifikáns különbség az adott 10 egyénre vonatkozóan ismételten nem mutatható ki azzal kapcsolatban, hogy valamelyik nem érzelmei és (tag)mondatbeli hiányai között jelentősebb összefüggés lenne.

A mondat tehát szakaszolja valamilyen módon az érzelmeket és kommunikációhoz való viszonyulást, de jelen tanulmányban még nincsenek olyan adatok, amelyekből arra lehetne következtetni, hogy beszélés és hallgatás (figyelés) közben milyen különbségek vannak az érzelmekben. Ha erre vonatkozóan lenne adatunk, akkor láthatnánk, hogy milyen módon szakaszolja a beszéd az érzelmeket.

Az eddigi eredmények újabb lehetőségeket vetnek fel az érzelmek és a beszéd kapcsolatának kutatásában, hiszen általuk a befejezetlen mondatot kiváltó érzelmek lokalizálásának három lehetőségére kereshetünk választ a jövőben:

Az adatok a mondat időintervallumára vonatkoznak, ami a mondat befejezése előtti közvetlen szakaszt mutatja meg. Tehát arra kaptunk választ az adatok megfigyelésével, hogy milyen érzelmek előzik meg közvetlenül a befejezetlen mondatot. Az eredményekből láthatjuk, hogy a közvetlen befejezést nem előzi meg tendenciaszerűen egy (annotátor által érzékelt) érzelem a mondaton belül. Érdemes lesz megvizsgálni ezért a befejezetlen mondat előtti és az azt követő mondatot, hogy azokban a szakaszokban tetten érhető-e nagy számban valamilyen érzelem.

Ha értelmezhetően jelentős számban lesz jelen érzelem a mondat előtti szakaszban, akkor feltételezhetjük, hogy a befejezést megelőző érzelemsor az, ami kiváltja a mondat mint beszédakuts befejezését, de az is lehet, hogy a kognitív tervezést kíséri az érzelem. Ha értelmezhetően jelentős számban lesz jelen érzelem a mondat befejezése utáni szakaszban, akkor feltételezhetjük, hogy a befejezést követő érzelem mutat rá a befejezés okára. A mondat befejezése utáni másodpercekben talán könnyebben érzékelhető az érzelem az arcon, mint a mondat befejezésének pillanatában. Ha ezt a két folyamatot a jövőben megvizsgáljuk, akkor megkísérelhetünk választ adni a befejezetlen mondatok érzelmi hátterére

6. Összegzés és kitekintés

Az érzelmeknek jelentős hatása van a gondolkodásra, az információ megszerzésére, feldolgozására és a gondolat kifejezésére. Általánosságban tehát azt vizsgáltuk, hogy a gondolkodásra milyen hatással vannak az érzelmek, és ez a hatás hogyan nyilvánul meg mondattanilag a beszédprodukcióban, valamint arra is választ kerestünk, hogy milyen érzelmi dinamikát vehetünk észre a mondatokban az érzelmi váltásokon kerresztül. Mindezt az ELAN programok segítségével tettük meg, valamint segítségünkre volt a PRAAT és a HuComTech csoport saját fejlesztésű szoftvere, a QUANNOT [16]. Részben ezeknek a szoftvereknek a lehetőségei alakították az eredmények elméleti és módszertani alapját és részben a terminológiai bázisát is. Kutatásunktól azt vártuk, hogy a multimodális HuComTech Korpusz szintaktikailag annotált dialógusainak adatait felhasználva összefüggéseket találjunk a szintaktikai elemek és a multimodális jegyek között.

Jelen tanulmányban tehát empirikus adatok közlése volt a cél. Egy adott mondaton belüli érzelemváltások dinamikájának megfigyelésére, és a befejezetlen mondatok létrejöttének érzelmi hátteré kerestük a válaszokat. Jelenlegi adatokból nem következtethetünk szignifikáns megállapításokra, de arra vonatkozóan elég információt kaptunk, hogy milyen irányokba indulhatunk el, hogy megláthassuk, hogy a beszéd és az érzelmek között milyen kapcsolat lehetséges.

Az adatbázisban háromféle érzelmet annotáló séma jött létre. Az első video és audio anyagot tartalmaz egyszerre, a második csak audio, a harmadik pedig csak video annotációkból áll. A háromféle modulban létrejött annotációk a befogadás szempontjából készültek el, hiszen az annotátorok intuícióin alapulnak. Ennélfogva azokat az

érzelmi reakciókat vettük alapul, amelyek valamennyi modulban mindenki által jól érzékelhetően jelen vannak, és nem pedig azokat az érzelmeket, amelyekről a beszélő beszámolhat ugyan, de mint érzelmi folyamatok a ripotalanyban zajlanak, és láthatatlanok maradnak a befogadás szemszögéből. A háromféle változat azt is lehetővé teszi majd, hogy a különböző modulokkal történő vizsgálatok eredményeit összevessük egymással és további következtetéseket vonjunk le [17].

Hivatkozások

1. Ekman, P.: Leleplezett hazugságok. Kelly Kiadó, Budapest (2007) 20
2. Eysenck, M. W., Keane M.T.: Kognitív pszichológia. Nemzeti Tankönyvkiadó, Budapest (2003) 353–364
3. Forgács, J.: Érzelem és gondolkodás. Kairosz Kiadó (2000) 11–33
4. Hunyadi, L., Kiss, H., Szekrényes, I.: Intelligent Decision Technology Support in Practice. Springer International Pulishing Switzerland (2016) , 231–257
5. Gósy, M.: Pszicholingvisztika. Osiris Kiadó, Budapest (2005) 80–88
6. Forgács, J.: Érzelem és gondolkodás. Kairosz Kiadó (2000) 14–15
7. Forgács, J.: Érzelem és gondolkodás. Kairosz Kiadó (2000) 15
8. Dell, G.S.: A spreading-activation theory of retrieval in sentence production. Psychological Review 93 (1986) 283–321
9. Tisljár-Szabó, E.: Érzelmek és beszéd. Szerk.: Pléh, Cs., Lukács, Á.: Pszicholingvisztika 2., Akadémiai Kiadó, Budapest (2014) 957
10. Gósy, M.: im.: .: Pszicholingvisztika. Osiris Kiadó, Budapest (2005) 89
11. Ekman, P.: Leleplezett hazugságok. Kelly Kiadó, Budapest (2007) 48
12. Ekman, P.: Leleplezett hazugságok. Kelly Kiadó, Budapest (2007) 20
13. Gósy, M.: Szerk., A spontán magyar beszéd megakadásainak hallás alapú gyűjteménye. Beszédkutatás 2004 (2004) 7
14. Gósy, M.: Szerk., A spontán magyar beszéd megakadásainak hallás alapú gyűjteménye. Beszédkutatás 2004 (2004) 16
15. Ekman, P.: Leleplezett hazugságok. Kelly Kiadó, Budapest (2007) 262
16. Pápay, K.: Rekurzió a nyelvben I. A beszélő személy akusztikai és vizuális gesztuskészlethasználtának vizsgálata multimodális korpusz alapján – beágyazások, beékelések és adaptáció. Szerk.: Hunyadi, L.,Tinta Könyvkiadó (2011) 38
17. Hunyadi, L.: On Multimodality in the Perception of Emotions from Materials of the HuComTech Corpus. CogInfoCom 2015 (2015) megjelenés alatt

Rádióműsorok elemzése a WordNetAffect érzelmi szótár segítségével

Lukács Gergely[1], Martos Tamás[2], Jani Mátyás[1], Takács György[1]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
[2] Semmelweis Egyetem, Mentálhigiéné Intézet
1089 Budapest, Nagyvárad tér 4.
{lukacs, jani.matyas, takacs.gyorgy}@itk.ppke.hu
martos.tamas@public.semmelweis-univ.hu

Kivonat: A hang alapú tartalom személyre szabásához az elmúlt évek technikai fejlődése, elsősorban az okostelefonok és a mobilinternet elterjedése, megteremtette a technikai hátteret. Ennek megfelelően az lejátszási lista készítés (angol: „playlist generation") fontos kutatási területté lépett elő.
Jelen munka célja a kevert beszéd-zene lejátszási listák készítésének nyelvtechnológiai vizsgálata. Előzetes kutatásunk alapján a hanganyagok szöveges leiratából elsősorban a hangulatnak van jelentősége a lejátszási lista készítésnél. A kérdés vizsgálatához rádióadók mintegy 2500 órányi műsorát vizsgáltuk meg. A felvételekben automatikus beszédfelismerővel a WordNetAffect érzelmi szótár szavait ismertük fel, majd az így kapott adatbázist elemeztük. Jellegzetes mintákat találtunk az érzelmi kategóriák együttes előfordulására és az érzelmek időbeli – heti, napi és óránkénti – változására vonatkozóan is.

1. Bevezetés

A hang alapú tartalom személyre szabásához az elmúlt évek technikai fejlődése, elsősorban az okostelefonok és a mobilinternet elterjedése, megteremtette a technikai hátteret. A közelmúltban számos személyre szabott zenei streaming szolgáltatás indult el, példaképpen a Spotify, az Apple Music, a Google Play Music vagy a Pandora. Az audió tartalmak személyre szabott kiválasztása és lineáris műsorfolyammá szerkesztése, az ún. lejátszási lista készítés (angol: „playlist generation") ezzel együtt fontos kutatási területté vált[1]. A lejátszási lista készítés történhet a hanganyagok akusztikai tulajdonságai, a szöveges tartalma (dalszöveg illetve beszédanyag leirata) és a felhasználói interakciók alapján. A szöveges tartalom figyelembevételénél elsősorban a kapcsolódó hangulatnak van jelentősége.

A beszéd-zene lejátszási listák kutatását a megfelelő adatbázisok hiánya nehezíti. Szerkesztett (pozitív és negatív mintákat is tartalmazó) beszéd-zene műsorfolyam készítése óriási költséggel járna, ugyanakkor a professzionálisan szerkesztett rádióadók felvételei jó támpontot adhatnak. Ezeknél nehézséget jelent ugyan, hogy csak pozitív mintákat tartalmaznak, és jelen vannak torzító tényezők is, mint a csatornák tematikussága, vagy az óra által meghatározott műsorstruktúra, pl. a hírek gyakori

ismétlése. Előnyük viszont, hogy nagy mennyiségű adat készíthető automatizáltan, és az adatok elemzése jó kiindulási pontként szolgál további megfontolásokhoz.

Jelen munka keretében professzionálisan szerkesztett rádióadók felvételeiből készítettünk egy adatbázist, részben automatikus beszédfelismerő segítségével, majd az elemeztük. A munka felépítése a következő. A 2. fejezet az irodalomkutatás eredményeit, a 3. fejezet a hangulatfelismerés lehetőségeit mutatja be. A 4. fejezet az adatbázis felépítésével és elemzésével foglalkozik, az. 5. fejezet az adatelemzés eredményeit mutatja be. Az összefoglalás és kitekintés az 6. fejezetben található.

2. Kapcsolódó munkák

Az első kutatások elsősorban a zenei lejátszási listákkal foglalkoztak, és főként akusztikai oldalról közelítették a feladatot. Újabb tanulmányok és vizsgálatok szerint ugyanakkor a nyelvtechnológiának is szerepe van, leginkább a számok közti hangulati harmónia figyelembevételénél [2].

A beszéd-zene lejátszási listák abban különböznek a pusztán zenei lejátszási listáktól, hogy a zeneszámok mellett beszédfelvételeket (pl. interjúk, aktuális hírek stb.) is tartalmaznak. Ezeknek az automatikus készítésére is mutatkozik igény. Amellett, hogy a felhasználó a saját ízlésének megfelelő zenéket hallgatja, össze van kötve a külvilággal és az aktuális eseményekkel [3]. Az erre vonatkozó eddigi kutatások a beszédzene átmenetek akusztikai tulajdonságait [4], a szöveges tartalom szerepét [5], valamint a tartalomfüggetlen, kizárólag a felhasználói visszajelzéseken alapuló ú.n. felhasználói szűrés (collaborative filtering) lehetőségeit vizsgálták [6].

Közösségi média hangulati elemzésével több kutatás is foglalkozik, példaképpen a Twitter elemzését írja le [7].

3. Hangulatfelismerés, a WordNetAffect hangulati szótár

Természetes nyelvi szövegek hangulatának felismerésére, mérésére több szótár, eszköz illetve módszer is elérhető, példaképpen a WordNetAffect (WNA) [8], a Linguistic Inquiry and Word Count (LIWC) [9], az Affective Norms for English Words (ANEW) [10] és a SentiWordNet [11].

Jelen munkához egy, a leggyakrabban használt kétértékűnél (pozitív/negatív) részletesebb kategóriarendszerre volt szükségünk. Választásunk ezért a WordNetAffect érzelmi szótárra esett, mely az Ekman-féle kategóriarendszert [12] használja, hat alapérzelmet különböztetve meg. Az érzelmek felismerése, mérése az WNA adott érzelmi kategóriájába tartozó szavak számának meghatározásával történik, további súlyozás nélkül. Vizsgálatunk során így az Ekman-féle kategóriákat különböztetjük meg, ezek az öröm (Ö), a meglepettség (M), a félelem (F), a düh (D), a szomorúság (Sz) és az undor (U).

4. Adatbázis

4.1. Eszközök, elkészítés

Eszközök, az elkészítés folyamata. Online elérhető rádióadókról készítettünk felvételeket, a hanganyagokat órás blokkokban rögzítettük. A felvételeken beszéd-zene szétválasztót és szünet detektort futtattunk. A beszéd szakaszokat beszédfelismerő segítségével leiratoztuk, valamint kulcsszókereső algoritmusok segítségével a WordNetAffect érzelmi szótár szavait kerestük.

Beszéd-zene szétválasztó. Beszéd-zene szétválasztóként a Xiph.Org[1] által készített Opus hangkodek egyszerű neuronhálóval és ennek a kimenetét simító két állapotú Markov-modellel készített beszédfelismerő algoritmusát vettük alapul és egészítettük ki további utófeldolgozással. A kiegészített algoritmus a beszéd és zeneszakaszok mellett a kevert szakaszokat is felismeri, melyekben lehalkított zene és beszéd egyszerre hallhatóak.

Az automatikus szegmentálás minőségének ellenőrzéséhez csatornánként 5-5 óra felvételen kézzel kijavítottuk a hibákat. Ezután kiszámoltuk a Cohen-féle kappát és alapértelmezett módon majd súlyozva is. Az automatikus felismerő három osztályt tud felismerni (beszéd, zene, beszéd-zene), az annotátorunk viszont időnként jelölte az egyéb (pl. taps, motorzaj) kategóriát is. A súlyozott kappa számításánál a mindkettő-beszéd és a mindkettő-zene (illetve ezek fordítottai) fele olyan súllyal szerepeltek, mint a többi tévesztés. Az 1. táblázat foglalja össze a kappa értékeket. Általánosságban 0,7-nél magasabb értékeket jónak szokták tartani. Egyedül az R4-es csatorna kappája alacsonyabb ennél. Érdekes jelenség, hogy az R3 komolyzenei adón kimagaslóan jól egyezik az automata a kézi annotálással. Ez valószínűleg részben annak köszönhető volt, hogy ritkábban szólt egyszerre beszéd és zene is, nem nagyon volt "átúsztatás", tisztábbak voltak a váltások.

1. táblázat: A beszéd-zene felismerés minősége

Rádióadó	Cohen's kappa	Súlyozott Kappa
R1	0,713	0,814
R2	0,730	0,822
R3	0,930	0,936
R4	0,601	0,601

Beszédfelismerő. Az automatikus leiratot és a kulcsszókeresést a Kaldi [13] nyílt forráskódú beszédfelismerő rendszer segítségével végeztük. Az alapértelmezetten beépített TEDLIUM adatbázisra építő receptet használtuk a leiratokhoz. Csak a legutolsó GMM-HMM modellt használtuk, mély-neuron hálókat nem alkalmaztunk. Ezt leszámítva a tanítást a recept változtatása nélkül végeztük. Kulcsszókereséshez a Kaldiban található, hipotézis-gráf indexelésén alapuló kulcsszókeresési eljárást használtuk. A kulcsszavaknak a WordNetAffect érzelmi szótár szavait választottuk, célunk az érzelmi kategóriák előfordulásának vizsgálata volt. Az automatikus leiratot csak a szószámláláshoz használtuk. A beszédfelismerés hibájára közvetlen mérésünk nincs. A

[1] http://www.xiph.org

beszédfelismerő korábbi változatánál, más adathalmazon a szófelismerési hiba (WER) 47,2% volt [5], ugyanakkora hibák kiegyenlítették egymást és nem befolyásolták számottevően az érzelemfelismerés pontosságát.

4.2. Adatok leírása

Az adatbázis a BBC rádióadó négy csatornájának adásait tartalmazza. A Radio 1 és a Radio 2 könnyűzenei adók, az előbbi fiatalok számára, az utóbbi felnőtteknek. A Radio 3 komolyzenei adó, a Radio 4 pedig irodalmi és aktuális eseményekkel kapcsolatos műsorokat ad.

A felvételek 2014. december 22. és 2015. január 20. között készültek, összesen 2 444 óra terjedelemben. Technikai okok miatt a felvételek nem fedik le a teljes időtartományt, szünetek előfordulnak. A beszéd, kevert beszéd-zene és beszéd arányát az egyes adóknál a 2. táblázat mutatja. A hangadatok áttekintése az 1. ábrán látható.

2. táblázat: Zene, beszéd/zene és beszéd arány a vizsgált rádióadóknál

	Zene	Kevert beszéd-zene	Beszéd
R1	75%	10%	15%
R2	65%	6%	29%
R3	83%	2%	16
R4	7%	9%	83%

A teljes adatbázis 10 572 398 szót tartalmaz. A WNA szavak száma 233 760, arányuk így 2,2%. A WNA érzelmi kategóriái között a WNA szavak megoszlása a következő: öröm: 45%, düh: 14%, félelem: 12%, szomorúság: 15%, meglepetés: 12%, undor: 3%. Az örömhöz tartozó, kiemelkedően magas érték egybecsenghet [14] megfigyelésével. A többi érzelemhez tartozó érték hasonló egymáshoz, kivéve az undorra vonatkozó alacsony számot.

Az adatbázis a jelen munkában bemutatott általános vizsgálatokhoz elegendő nagy. Speciális kérdésfelvetések esetén (pl. undor vizsgálata hétvégi éjszaka) ugyanakkor a jelen adatok alapján végzett megfigyelések bizonytalansága magas.

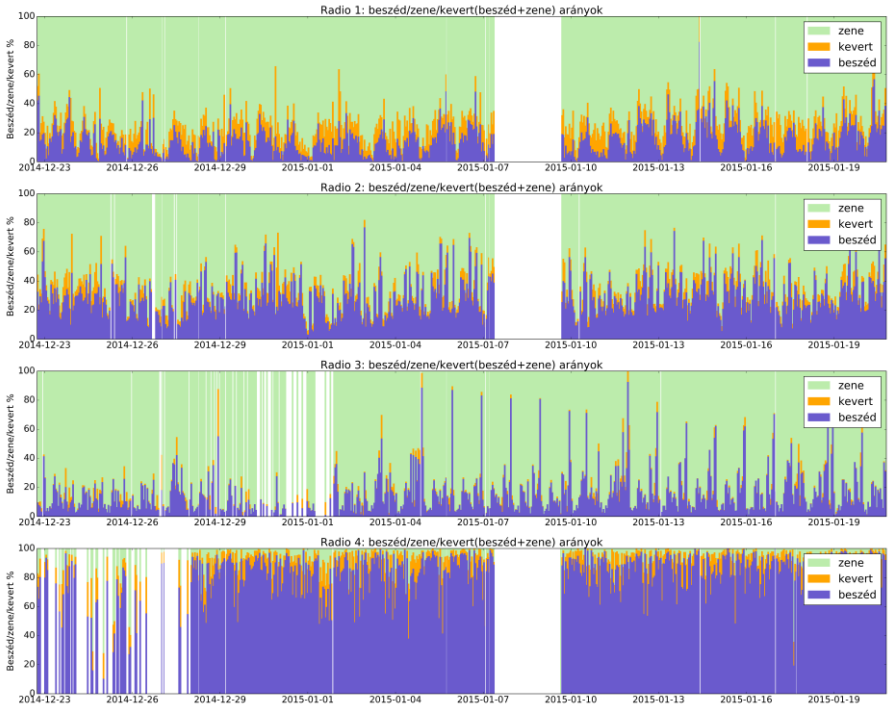
5. Vizsgálatok

5.1. Kiértékelési mérték

A vizsgálatoknál a következő mértéket vettük figyelembe. Kiindulási pontként a WNA szavak és az összes szó aránya szolgált, példaképpen egy adott vizsgálatnál (adónál, időtartományban stb.) a örömhöz tartozó WNA szavak aránya az összes elhangzott szó 1,1%-a.

Mivel az egyes érzelmi kategóriákhoz tartozó szavak előfordulási gyakorisága nagyban különbözik, ezért célszerű volt a WNA szavak arányának változását nem abszolút értékben mérni, hanem az adott esetre vonatkozó átlagos arányhoz viszonyít-

va. A WNA szavak (átlagos arányához mért) relatív arányán a trendek világosabban értékelhetőek, és az egyes érzelmi kategóriákra vonatkozóan is összehasonlíthatóak.



1. ábra: Rádiófelvételek áttekintése, beszéd-zene arány az idő függvényében.

5.2. Érzelmi kategóriák aggregálása

Az adatokon ellenőriztük, hogy az érzelmek együttes előfordulása mutat-e valamilyen mintázatot. Konkrétabban arra voltunk kíváncsiak, hogy az azonos, illetve hasonló érzelmi jelentésű szavak nagyobb valószínűséggel fordulnak-e elő egymáshoz közeli időpontban. Ha igen, akkor ez két előzetes feltevést igazol. Egyrészt azt jelzi, hogy az egyes érzelmi jelentésű szavak nem véletlenszerűen fordulnak elő egymáshoz közel, hanem az adott szöveg adott szövegkörnyezetére általánosabban jellemző érzelmi-hangulati állapotot tükröznek önmagukban is és az együttes előfordulásukból sejthető-en összességében is. A második igazolt feltevés az elsőből következik: amennyiben az egyes szavak érzelmi jelentése egy általánosabb érzelmi mintázatot jelez, akkor az elkülönült előfordulások összegzése is érvényes eljárás arra, hogy nagyobb, egybefüg-gő hangzó szöveg érzelmi jellegének leírására használjuk az így számított aggregált értékeket. (Az együttesen összegzendő szakaszok kijelölése természetesen szintén fontos gyakorlati kérdés, de ezt a problémát jelen érvelés most nem érinti.)

A következő ellenőrzést végeztük el: az adatbázisban megkerestük azokat a kódolt érzelmi szavakat, melyek elhangzása között legfeljebb 3 mp volt a különbség – azaz feltételezhető volt, hogy amennyiben létezik a szövegnek valamilyen tartósabb hangulati állapota, akkor a két érzelmi jelentésű szó körülbelül azonos érzelmi-hangulati állapotú beszélőtől származik.

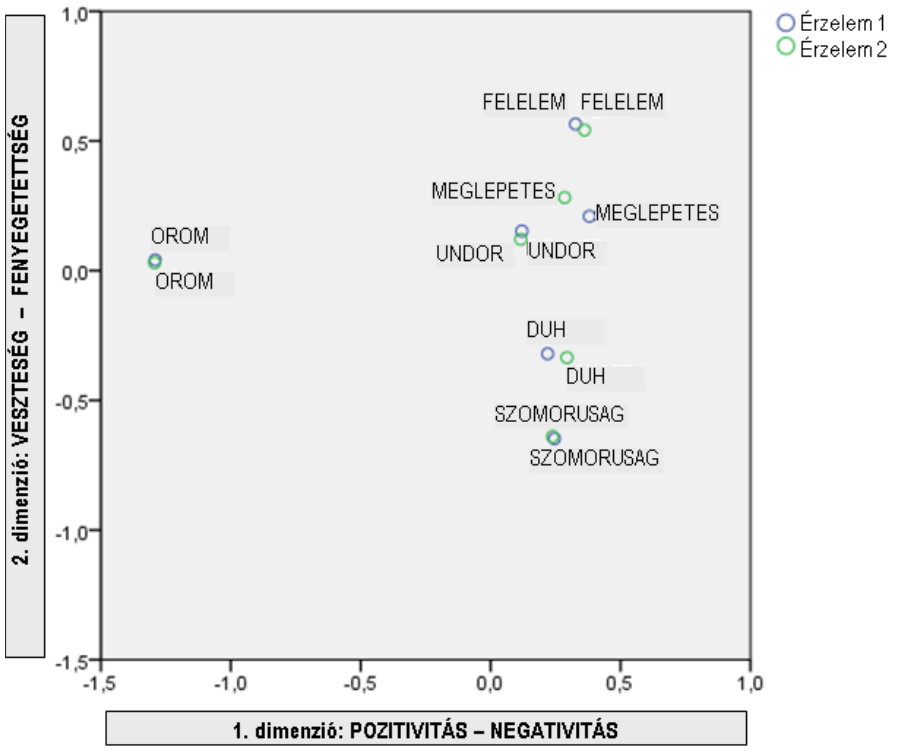
A kódolt anyagban összesen 91 255 ilyen szópár volt. Az együttes előfordulások mintázatának feltárására korrespondencia-elemzést (correspondence analysis, CA) végeztünk az SPSS programcsomag felhasználásával. A CA kategoriális adatok együttes előfordulásainak mintázatát jeleníti meg egy többdimenziós térben. Az egyes lehetséges párok együttes előfordulásainak valószínűségét a párok tagjainak egymástól való távolsága reprezentálja. (Azaz a közeli párok nagy valószínűséggel fordulhatnak elő 3 mp-en belül, míg a távoli párok együttes előfordulásának valószínűsége csekély.)

Eredmények: A CA alapján két dimenzió írja le az adatok varianciájának összesen 92,4 %-át. Ebből az 1. dimenzió magyarázott 77,1 %-ot, míg a második 15,3 %-ot. Az egyes érzelmi kifejezés-párok együttes előfordulásának valószínűségét a 2. ábrán mutatjuk be. Az ábra alapján két fontos megállapítást tehetünk.

1. A legnagyobb valószínűséggel ugyanazon érzelmi kifejezések jelennek meg egymástól legfeljebb 3 mp távolságban. (Azaz az időben első és a második előfordulást reprezentáló kétféle színű pont gyakorlatilag átfedésben van az ugyanazon érzelmeket reprezentáló szavak esetében.) Ez az eredmény arra utal, hogy a kódolt szövegek – viszonylag rövid időhatárokon belül – valóban egységes hangulati-érzelmi állapotokkal jellemezhetők. Másfelől pedig jelzi a kódolási rendszer érvényességét is.

2. A két dimenziót is megkísérelhetjük értelmezni. Az ábrán vízszintes dimenzió egyértelműen a pozitív (öröm) és a negatív érzelmeket jelölő szavak elkülönülését jelzi. A másik (az ábrán függőleges) dimenzió értelmezése kevésbé egyértelmű, de alább megkíséreljük. Itt a pozitív oldalon természetszerűleg nem volt változatosság. A negatív oldalon az egyik póluson a félelem, a másikon a szomorúság jelzi az egymástól elkülönülő érzelmeket. A félelemhez közelebb helyezkedik el a meglepetés és az undor, míg a szomorúsághoz a düh. Ez a dimenzió értelmezésünk szerint annak az érzelemfelosztásnak felel meg, ami a negatív érzelmeket aszerint különbözteti meg, hogy milyen tapasztalatra adott választ jeleznek. A félelem egy fenyegetésre, averzív inger közeledésére adott reakció, míg a szomorúság valamilyen pozitív élmény elvesztésére, távolodására adott válasz. A köztes érzelmek is besorolhatók ebbe a modellbe, hiszen a meglepetés és az undor inkább valamilyen közelítő averzív ingerhez kapcsolódik, míg a düh a szomorúsághoz hasonlóan a frusztrációra, azaz veszteségre adott reakció. A második tengelyt ezért veszteség vs. fenyegetés tengelynek neveztük el.

Összességében mindkét szintű mintázat (azonos érzelmek közelsége és a kétdimenziós érzelmi tér értelmezhetősége) arra utal, hogy a kódolási rendszer érvényes adatokat szolgáltat és az egyes érzelmek összesített gyakorisága is érvényes eljárás – beleértve akár a pozitív és negatív érzések elkülönült aggregálását is.



2. ábra: A 3 mp-en belül együttesen előforduló érzelmi jelentésű szavak korrespondancia-elemzésének eredménye (91255 szópár alapján). Az egyes dimenziók elnevezése a szerzők javaslata.

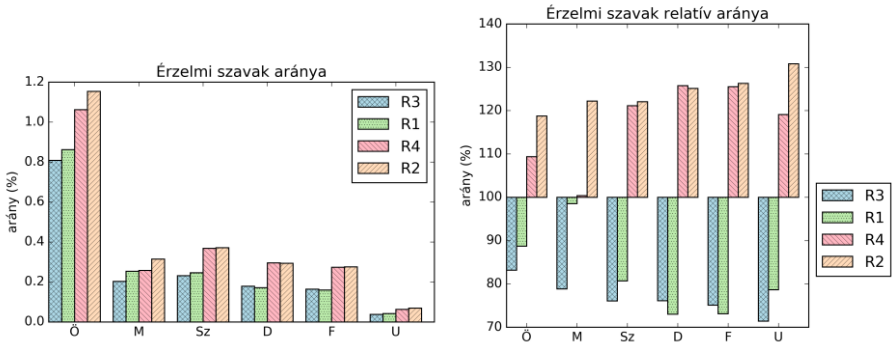
5.3. Érzelmi kategóriák időben aggregált aránya

Elsőként időben aggregálva vizsgáltuk a hangulati szavak előfordulását az egyes rádióadóknál, két mértéket használva: (1) a hangulati szavak aránya az összes elhangzott szóhoz viszonyítva, (2) és a hangulati szavak relatív aránya a négy csatorna átlagértékéhez viszonyítva (3. ábra).

Az R2 és az R4 esetében jelentősen gyakoribbak a hangulati szavak. Egy lehetséges magyarázat, hogy ezek több, mélyebb tartalmat akarnak közvetíteni, amire nagyobb beszédarányuk is utalhat. (A nagyobb beszéd arány, vagy akár a beszédtempó csatornánkénti különbözősége közvetlenül nem befolyásolja a vizsgált értékeket, hiszen a hangulati szavak és az elhangzott szavak arányát, illetve ennek az aránynak a változását vizsgáljuk.).

Az R2 és R4 örömre vonatkozó jellemzői ugyanakkor kevésbé kiugróak, mint a szomorúság, düh, félelem és undor esetében. Egy lehetséges magyarázat, hogy az

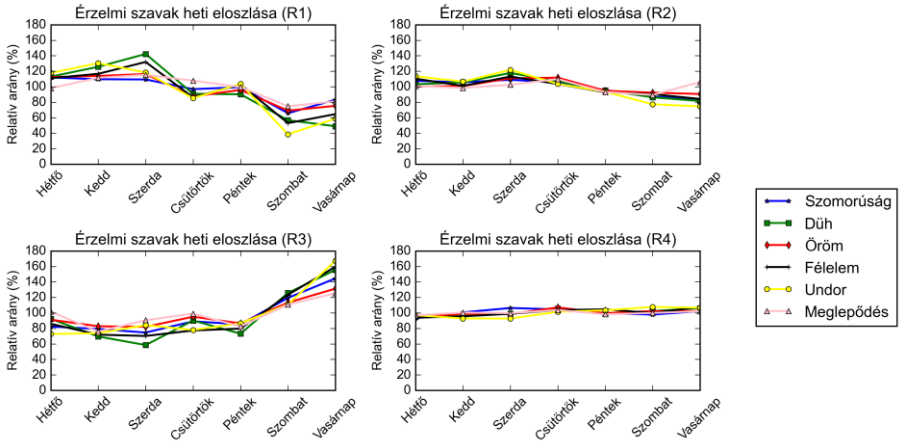
örömöt viszonylag könnyebb kommunikálni, és a több tartalmat szolgáltató adók (az R2 és az R4) az amúgy inkább háttérbe szorított negatív érzelmeknek is nagyobb teret ad. A meglepetésre vonatkozóan az R1 és az R4 esetén nem áll a fenti minta, az okok felderítéséhez további vizsgálatokra van szükség.



3. ábra: WNA érzelmi kategóriák aránya és relatív aránya a vizsgált rádiófelvételekben.

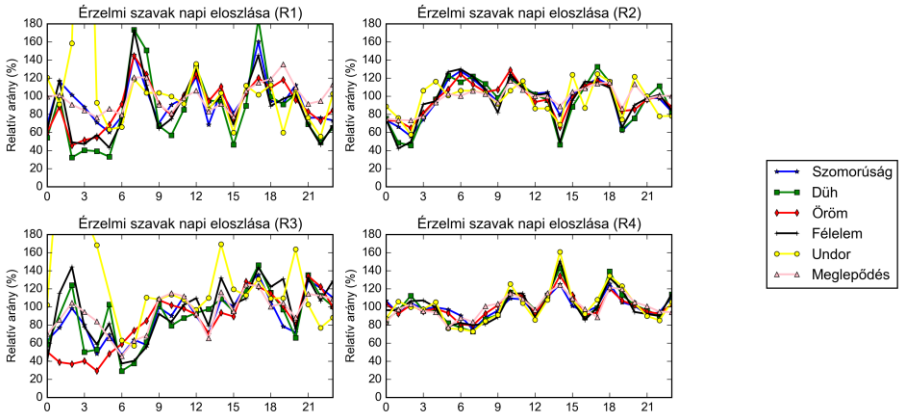
5.4. Időbeli minták

Heti minta. A WNA szavak arányának heti ritmusú változását a 4. ábra mutatja. Az érzelmi szavak az érzelmi kategóriától függetlenül együtt változnak. Az R1 esetén hétköznap magasabb, hét végén alacsonyabb a WNA szavak aránya. Az R3 esetében ellenkező minta figyelhető meg, a hétvégéken jelentősen magasabbak az értékek. Az R2 esetén jóval mérsékeltebb a heti ritmusú változás, ennek iránya pedig az R1-hez hasonló: hét közben magasabb értékek. Az R4 esetében alig észrevehető a heti minta, tendenciáját tekintve az R3-hoz hasonlóan hétvégén emelkednek, bár csak kis mértékben az értékek.



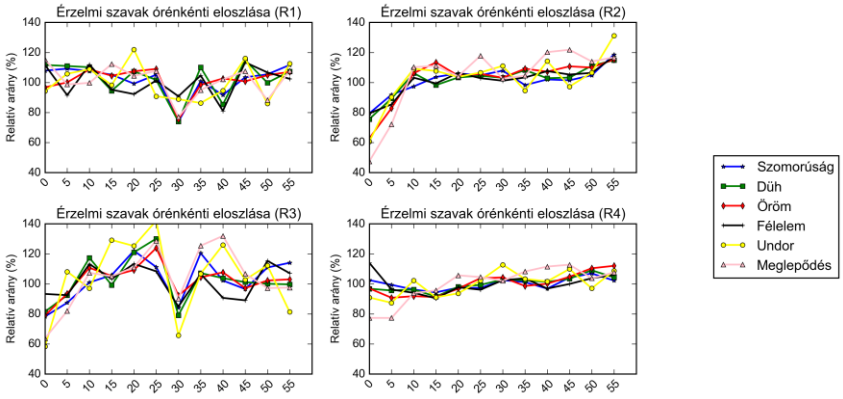
4. ábra: WNA érzelmi kategóriák heti ritmusa adónként.

Napi minta. A napi ritmusú változást az 5. ábra mutatja. Az érzelmek itt is nagyjából együtt mozognak. Általában jellemző a hajnali nyugodtabb és a nappali magasabb szint (az R1 és az R3 esetén az undorra vonatkozó hajnali kiugró értékek nagyon kis számú mintából történtek kiszámolásra). Az R1 adóra jellemzőek a reggeli és az esti kiugró értékek. Az R2 esetében kisebbek a változások, de a délutáni és éjszakai alacsonyabb értékek itt is megfigyelhetőek. Az R3 esetében egy délutánra, estére emelkedő tendencia látható. Ez esetleg összefügghet a heti ritmusban megfigyelt, hétvégére emelkedő értékekkel is. Az R4 napi ritmusán csak kisebb változások figyelhetőek meg, a délutáni csúcsok ennek ellenére felismerhetőek.



5. ábra: WNA érzelmi kategóriák napi ritmusa adónként.

Órai minta. Célszerűnek tűnik az óránkénti mintát is megvizsgálni, hiszen a rádióadók műsorszerkesztése ehhez igazodik. Az órák első perceiben, amikor általában a hírek hallhatóak, a WNA szavak aránya alacsony. Ugyanakkor ez még szembetűnőbb az öröm és a meglepetés esetében, ezekhez képest a félelem, düh, szomorúság, undor aránya általában viszonylag magasabb. Elsősorban a nagyobb beszédaránnyal rendelkező R2 és R4 adókra jellemző, hogy az órák folyamán enyhén növekszik a WNA szavak aránya.



6. ábra: WNA érzelmi kategóriák óránkénti ritmusa adónként.

6. Összegzés, kitekintés

Érdekes megfigyelés, hogy a hat vizsgált érzelem általában együtt mozgott, együtt mutattak fel alacsony vagy magas értékeket. Ez alapján a műsorok egyik legalapvetőbb (rejtett, implicit) tulajdonsága az lehet, hogy mennyire „telítettek" érzelemmel. Az (implicit) érzelmek milyensége, aránya csak második lépésben árnyalhatja az érzelmi összhatást.

Az időbeli változást tekintve jellemző heti, napi és óránként visszatérő mintákat találtunk. Az adók között ugyanakkor jelentős különbségek is vannak. Több szempont szerint hasonlóképpen viselkedik az R1 könnyűzenei és az R3 komolyzenei adó egyfelől, és a magasabb beszédaránnyal rendelkező R2 és R4 másfelől. Más szempont alapján pedig a két könnyűzenei adó, a fiatalokat célzó R1 és az idősebbeknek szánt R2 mutat hasonlóságot.

A Twitter adatok vizsgálata a WNA segítségével [7] alapján jellegzetes napi mintázatot mutatott, reggel a pozitív, este a negatív érzelmek mutattak viszonylag magasabb értéket. A rádióadások napi mintája ezt részben követi. A rádióadások ugyanakkor rendelkeznek egy jellegzetes óránként visszatérő mintával is, a műsorszerkesztés sajátosságai miatt. A Twitter adatainak vizsgálatakor a téli-nyári adatok összevetése is megtörtént [7], erre a rádióadások elemzéséhez nem áll rendelkezésre adatunk. A rádióadásokban ugyanakkor egy jellegzetes heti minta is felismerhető, a Twitter adatainak vizsgálata ilyen szempontból érdekes összehasonlításokra adhatna lehetőséget.

Érdekes lenne tovább validálni emberi észlelőkkel végzett kísérletben, hogy a szavak gyakoriságával kimutatott hangulati jellemzők mennyire korrelálnak a hallgatók által észlelt hangulattal. További kutatási irányt jelent a hang egyéb, nem nyelvi jellemzőnek (pl. hangerő, beszédtempó) a vizsgálata, illetve az egyes tényezők egymással való összefüggése.

Köszönetnyilvánítás

A munka a KAP15-059-1.1-ITK pályázat keretében valósulhatott meg, melyért a szerzők ezúton is köszönetüket fejezik ki.

Hivatkozások

1. Fields, B.: Contextualize Your Listening: The Playlist as Recommendation Engine. PhD thesis, Goldsmiths, University of London, London (2011)
2. Hu, Xiao, Downie, J. Stephen: When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In Proceedings of the 10th International Conference on Music Information Retrieval , Utrecht, The Netherlands (2010) 619–624
3. Lukacs, G., Pethesné, D.B., Madocsai, B.: Impact of Personalized Audio Social Media on Social Networks. In XXXIII. Sunbelt Social Networks Conference of the International Network for Social Network Analysis Abstract Proceedings , Hamburg, Germany (2013) 210

4. Jani, M., Lukács, G., Takács, Gy.: Experimental Investigation of Transitions for Mixed Speech and Music Playlist Generation. In Proceedings of ACM International Conference on Multimedia Retrieval , Glasgow, United Kingdom (2014) 392–398
5. Benyeda, I., Jani, M., Lukács, G.: Beszéd-zene lejátszási listák nyelvtechnológiai vonatkozása. In Proc. of XI. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2015 , Szeged, Hungary (2015) 257–268
6. Jani, M.: Fast Content Independent Playlist Generation for Streaming Media. In 12th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2015 , Marrakech, Morocco (2015) To appear
7. Strapparava, C., Valitutti, A.: WordNet-Affect: an Affective Extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation , Lisbon, Portugal (2004) 1083–1086
8. Pennebaker, J.W., Martha, E.F., Booth, R.J.: Linguistic Inquiry and Word Count. , Mahwah, NJ (2001)
9. Bradley, M.M., Lang, P.J.: Affective Norms for English Words (ANEW): Instruction manual and affective ratings, http://www.uvm.edu/~pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf, (1999)
10. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06 (2006) 417–422
11. Ekman, Paul: Facial expression of emotion. Am. Psychol. 48, (1993) 384–392
12. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding , Hilton Waikoloa Village, Big Island, Hawaii, US (2011)
13. Dodds, P.S., Clark, E.M., Desu, S., Frank, M.R., Reagan, A.J., Williams, J.R., Mitchell, L., Harris, K.D., Kloumann, I.M., Bagrow, J.P., Megerdoomian, K., McMahon, M.T., Tivnan, B.F., Danforth, C.M.: Human language reveals a universal positivity bias. Proc. Natl. Acad. Sci. 112, (2015) 2389–2394
14. Lampos, V., Lansdall-Welfare, T., Araya, R., Cristianini, N.: Analysing Mood Patterns in the United Kingdom through Twitter Content. CoRR. abs/1304.5507, (2013)

V. Szaknyelv, speciális nyelvhasználat

A magyar jelnyelvi korpusz létrehozásának és annotálásának kihívásai

Bartha Csilla[1], Varjasi Szabolcs[1], Holecz Margit[1]

[1] Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Többnyelvűségi Kutatóközpont, 1068 Budapest, Benczúr u. 33.

Kivonat: A 2015. október 31-én zárult JelEsély Projekt keretében egy hozzávetőlegesen 1750 órányi jelnyelvi korpusz jött létre. Országos terepmunka során 147 szociolingvisztikai interjú készült 5 régióban és 9 helyszínen, 27 grammatikai teszt során pedig 54 adatközlővel készültek felvételek (interjúnként 2 adatközlővel). Ahhoz, hogy a létrejött videoalapú korpusz kereshető, kutatható és felhasználható legyen, szükség van egyrészt a korpusz annotálására, amely folyamat során különféle információkat kapcsolunk a felvételekhez, másrészt a jelnyelvi felvételek fordítására. Írásunkban a jelnyelvi korpuszépítés és annotáció egyedi kihívásait ismertetjük, melyek többsége két okra vezethető vissza, melyek összefüggenek a jelnyelvek sztenderdizációjának kérdéseivel is. Egyrészt a jelnyelveknek nincs kidolgozott és elfogadott írásrendszerük, másrészt a jelnyelvekre – a sztenderdizálatlan hangzó nyelvekhez hasonlóan –jellemző a nagyfokú változatosság. A kereshető, immár géppel is olvasható korpuszok számos további kutatási lehetőséget biztosítanak, az alapvető statisztikai vizsgálatokon túlmenően is. A szociolingvisztikai kutatások mellett lehetővé válik korpuszalapú szótár létrehozása, valamint egy valós nyelvhasználaton alapuló grammatika megalkotása is. Vizsgálhatóak továbbá diskurzusjelenségek, pragmatikai sajátosságok és a siketes jelek is. A korpusz ezen kívül oktatási célokat is szolgálhat, például tan-, és segédanyagok létrehozásával.

1. Bevezetés

A siket közösség Magyarország harmadik legnagyobb nyelvi kisebbsége, annak ellenére, hogy „a veleszületett vagy szerzett halláskárosodás folytán a siket közösségek nem etnikai alapon szerveződnek, nyelvi kisebbségek abban az értelemben is, hogy sajátjukként *bármely más (hangzó) nyelvvel egyenértékű teljes, autonóm természetese nyelvet, jelnyelvet használnak*" [1: 85 – kiemelés az eredetiben]. Munkálataink során, a Többnyelvűségi Kutatóközpontban a siketségnek a jelnyelvet forrásként kezelő, nyelvi-szociokulturális megközelítésére alapozunk, szemben a fogyatékosságparadigma deficit-alapú megközelítésével: „[…] a kulturális, antropológiai értelmezés a siketséget egy olyan embercsoport létállapotának, adottságának tekinti, amely tagjainak közös vonása, hogy a világot elsődlegesen vizuálisan érzékelik, akiket közös kultúra, hasonló tapasztalatok, viselkedési szokások jellemeznek, s legfőképpen, közös nyelvet, a jelnyelvet használják, amely elsődleges kommunikációs és megismerő szerepe mellett – más nyelvi közösségekhez hasonlóan – önazonosságuk szimbóluma

is [2: 79]". Ebből adódóan a siket közösségek tagjai tehát nemcsak siketek és nagyothallók lehetnek, de hallók is (pl. siket szülők halló gyermekei, siket gyermekek családtagjai és a közösséghez csatlakozó, annak értékeivel, nézeteivel azonosuló hallók) [vö. 4, 5].

A 2009. évi CXXV. törvény a magyar jelnyelvről és a magyar jelnyelv használatáról mérföldkő volt a siket közösség életében. Nemcsak azért, mert a magyar jelnyelvet önálló, természetes nyelvként ismeri el, hanem azért is, mert biztosítja a jogi keretet a bilingvális oktatás 2017-től való bevezetésére. A bilingvális oktatás kidolgozásához azonban szükség van a magyar jelnyelv oktatási célú sztenderdizációjára. Ez a folyamat csak a siket közösség tagjainak bevonásával valósulhat meg, a megalapozásához pedig szociolingvisztikai alapon megtervezett, korpuszalapú empirikus nyelvészeti kutatásra van szükség. Ezt a célt tűzte ki a TÁMOP 5.4.6/B-13/1-2013-0001 *A magyar jelnyelv sztenderdizációjának elméleti és gyakorlati lépései* (Jel*E*sély) elnevezésű projekt.

A következőkben a projekt során létrehozott, folyamatos fejlesztés alatt álló korpuszt mutatjuk be.

2. A korpusz bemutatása

2.1. A korpusz mint kutatási bázis

Leech már a 90-es évek elején megfogalmazta, hogy a korpusznyelvészet valójában egy módszertani bázis, így könnyen alkalmazható a nyelvészet különféle területein, például a fonetikában vagy a szociolingvisztikában [32]. Rundell pedig a következő jövőképet vázolja 1996-ban: „Mindazok számára, akik a nyelvtanulás, nyelvi leírás, illetve nyelvoktatás bármely területén dolgoznak, a korpusz használata olyan természetessé és nélkülözhetetlenné fog válni, amilyen a lexikográfusok számára jelenleg [38]."

A korpuszok osztályozása többféle módon történhet, ily módon a szakirodalomban is különféle korpusztípusokkal találkozhatunk. A *referenciakorpusz* célja, hogy átfogó információt adjon egy nyelvről, annak minden fontos változatát és a szókincs jellegzetességeit is reprezentálja, ezáltal megbízható nyelvtanok, szótárak, tezauruszok és egyéb nyelvi referenciaanyagok alapjául szolgálhatnak [40]. Az anyagok kiválasztása során meghatározásra kerülnek azok a paraméterek, amelyek alapján adott szövegek a korpusz részévé válhatnak. Ez magába foglalja a lehető legtöbb szociolingvisztikai változó figyelembe vételét, valamint az egyes szövegtípusok arányának meghatározását. A *monitorkorpuszok* lehetővé teszik, hogy a nyelv időbeli változását is nyomon követhessük, a *párhuzamos korpuszok* esetében pedig a szövegek mellett megjelennek azok különböző nyelvű fordításai is. *Összehasonlítható korpuszról* akkor beszélhetünk, ha több mint egy nyelv vagy nyelvváltozat hasonló szövegei jelennek meg, a hasonlósági kritérium azonban nincs pontosan definiálva. A korpuszok többféle jellemzővel írhatóak le, ahol minden jellemzőnek van egy „alapértelmezett"értéke. Ha bármely jellemző eltér ettől, akkor már *speciális korpusz*ról beszélünk. Az alapértel-

mezett értékek: mennyiség=nagy, minőség=autentikus, egyszerűség=egyszerű szöveg, dokumentált=igen.

Jelnyelvek esetében a korpuszok nemcsak dokumentálják, megőrzik az egyes jelnyelveket, de ezzel együtt autentikus szövegekhez is hozzáférést biztosítanak. (Jel)Nyelvek reprezentatív mintáit szolgáltatják, miközben grammatikák vagy szótárak alapjait is képezhetik, hosszú távon pedig nyomon követhető a nyelv változása is.

A jelnyelvészeti korpusznyelvészet kialakulása a technológiai fejlődés függvényeként is értelmezhető. A hangzó és írott nyelvi korpuszok a 20. második felétől kezdve váltak egyre elterjedtebbé (Rundell 1996-ban kifejti, hogy az angol mellett egyéb nyelveken is megindultak a korpuszmunkálatok, a 90-es években pl. már több mint 12 nyelven voltak különböző korpuszok Európában. [38:7]. Fontos azonban kiemelni, hogy már a Chomsky előtti időszakban is voltak korpuszalapú kutatások, melyet korai korpusznyelvészetnek nevezünk [34]. Ez elsősorban helyesírási konvenciók meghatározására használt gyűjtemények, illetve a nyelvelsajátítás nyomon követésére vezetett naplók formájában valósult meg.

McEnery és munkatársai a kisebbségi nyelvi tervezés problematikájával kapcsolatban emeli ki, hogy széles körű kutatásokat és szoftveres erőforrásokat nem lehet hatékonyan létrehozni korpuszos források hiányában, emellett egynyelvű és párhuzamos korpuszokból származó adatokra is szükség van [34]. Az ind nyelvekkel kapcsolatban hangsúlyozzák, hogy a magas fokon sztenderdizálatlan szövegek esetében a szöveges kódolás kulcsfontosságú kihívást jelent.

Ez a kihívás a jelnyelvek esetében még hangsúlyosabban jelenik meg, ahol problémát jelent az eltérő modalitás és írásbeliség hiánya is. A jelnyelvek esetén -- kezdetleges formában ugyan – a notációs rendszerek kialakulásával (ld. lentebb) indulhattak meg a gyűjtések. A 90-es években több fontos előrelépést érdemes kiemelni, egyrészt a tárolóeszközök közül a digitális CD, majd a nagy teljesítményű háttértárak váltották föl az élő nyelvi szövegek kazettáit; másrészt megjelentek a beszélt nyelvi korpuszok is. Közülük külön kiemelendő a Wellington-korpusz, hazai tekintetben pedig a Budapesti Szociolingvisztikai Interjú (BUSZI) [30], majd a kétezres évekből a Kárpát-korpusz [27] és a BEA adatbázisa [7]. Az írott korpuszok közül a Magyar Nemzeti Szövegtár a legfontosabb. A magyar nyelvterületen létrehozott különböző adatbázisok elérhetőek a Nyelv- és beszédtechnológiai platform honlapján [35].

Bartha rávilágít arra a visszásságra, hogy habár az emberek jelentős része a mindennapi tevékenységei során a legutóbbi időkig (a számítógép és egyéb eszközök megjelenéséig, melyek új nyelvhasználati lehetőségeket hoztak magukkal) a beszédet részesítette előnyben (az írott nyelvvel szemben), ám a hozzáférhető korpuszok fordított arányokat mutatnak [3]. Ez részben azzal magyarázható, hogy a beszélt nyelvi diskurzusminták gyűjtése és átírása lényegesen nagyobb nehézséget jelent, mint az írott nyelvieké.

A nagymennyiségű adatok tárolása, a nyelvi adatok dokumentálása és megfelelő rendszerezése a legtöbb empirikus adatokkal dolgozó kutató számára fontos kérdéssé vált, pl. a szociolingvisztikában is [vö. 28]. A nagy mennyiségű szövegek tárolása az ezredfordulóra már adott volt, azonban ahhoz, hogy az egyszerű, általában CD-n tárolt jelnyelvi archívumokból valódi korpuszok jöhessenek létre, további fejlődésre volt szükség, így a jelnyelvi korpusznyelvészet kialakulásában fáziskéséssel kell számolnunk. Jelenleg több területen hiányzik még ezen tudományág kiforrott módszertana,

amely lehetőséget teremt egyrészt a fejlődésre, másrészt az írott nyelvi korpuszok tanulságainak implementálására.

A jelnyelvi korpuszok többnyire jelenleg is fejlesztés alatt állnak [13]. Habár már 1910 és 1920 között is készült korpusznak tekinthető gyűjtemény [29], de ezt követően hosszú idő telt el, míg a modernnek tekinthető korpuszok létrehozását célzó projektek elindultak a kétezres évek elején. Ezek közül legjelentősebbek a 2006-2008 között futó holland projekt a nijmegeni Radboud Egyetem koordinálásában [15], a veszélyeztetett nyelvi státusszal rendelkező ausztrál jelnyelv (Auslan) nyelvtanát és diskurzusstratégiáit dokumentáló 2004-től 2007-ig zajló korpuszprojekt [17], a brit jelnyelv (BSL) korpuszát létrehozó három és fél éves (2008 januárja és 2011 júniusa között futó) projekt [8], valamint az a jelenleg is tartó 15 éves projekt, amely a német jelnyelv korpuszának létrehozását tűzte ki célul [16]. A 2-3 év alatt összeállított nyersanyagok feldolgozásán, közzétételén, és felhasználásán (szótárak, oktatási anyagok, grammatikai vizsgálatok stb.) folyamatosan dolgoznak.

2.2. A magyar jelnyelvi korpusz létrehozása és feldolgozása

2.2.1. A korpusz felépítése

A korpusz fő alkotóelemei szociolingvisztikai és grammatikai tesztek felvételei, melyek több hónapon át zajló, országos terepmunka során készültek el. 7 mintavételi pontról (Budapest, Szeged/Hódmezővásárhely, Békéscsaba, Debrecen, Kaposvár, Sopron/Győr, Vác) 16 siket terepmunkás részvételével összesen 147 szociolingvisztikai interjú készült el, melyek közül 67 budapesti és 80 vidéki. Az interjúk 345 kérdésből álltak, a felvételek három kamerával való rögzítése pedig átlagosan 3-4 órát vett igénybe.

A grammatikai tesztek során 27 terepmunka alatt összesen 54 adatközlővel készültek felvételek (interjúnként 2 adatközlővel, melyek közül 11 teszt vidéki, 16 pedig budapesti adatközlővel készült). A grammatikai tesztek (a magyar jelnyelv alapgrammatikájának megírásához szükséges elicitációs tesztsorok) felvétele 5 kamerával zajlott, és átlagosan két órásak voltak.

A nyers videofelvételek feldolgozása többlépcsős munkafolyamatban zajlott. Az anyagokat először archiváltuk és vízjeleztük, ezt követően konvertáltuk. A korpusz nyersanyaga hozzávetőlegesen 1750 órányi, ami 6,5 terabájtnyi adatot jelent.

2.2.2. A korpusz feldolgozása

Ahhoz, hogy a videoalapú korpusz kereshető, kutatható és felhasználható legyen, szükség van a korpusz annotálására, mely folyamat során különféle információkat kapcsolunk a felvételekhez (pl. a felvételek közben megjelenő kézformák, a használt jelek magyar megfelelője stb.). Az annotációs részfolyamatokban nemcsak annotátorok, de fordítók és ellenőrzők is dolgoztak.

A kutatási céloknak megfelelően más-más protokollt alkalmaztunk a szociolingvisztikai és a grammatikai korpusz annotálásakor. A külföldi jelnyelvi korpuszprojektek áttekintését követően a szociolingvisztikai anyagoknál a jelnyelvalapú fordítás volt az elsődleges, amely azt jelentette, hogy a fordítók jelről-jelre haladtak folyamatosan, és nem csupán tartalmi összefoglalót készítettek. Ezáltal biztosítható a jelelni nem

tudó kutatók számára a korpusz anyagához való hozzáférés magyar nyelven, hiszen a jelnyelveknek, köztük a magyar jelnyelvnek nincsen általánosan elfogadott és széleskörűen használt írásrendszere. Habár több kezdeményezés is született a jelek írásbeli rögzítésére, mint például a HamNoSys [18] vagy a SignWriting [39], de ezek egyrészt jól jelelők számára is sokszor nehezen olvashatóak, másrészt jelelni nem tudó kutatók számára nem hozzáférhetőek[1]. A jelnyelvi videók magyar nyelvű fordítására tehát annak ellenére is szükség volt, hogy bizonyos esetekben elfedik a jelnyelv változatosságát (vö. [26]), illetve azt sugallja, hogy a jelnyelv és a magyar hangzó nyelv elemei között lehetséges az egyértelmű megfeleltetés, de ez természetesen nem igaz [vö. 5, 42, 43]. A jelnyelvek a hangzó nyelvekhez hasonlóan természetes nyelvek [41] melyek ugyanakkor a magyartól és más hangzó nyelvektől nagymértékben eltérő struktúrával és nyelvi eszközkészlettel rendelkeznek. Mindezek ellenére szükséges a halló kutatók számára is hozzáférhetővé tenni a korpuszt. Folyamatosan készülnek a magyar fordítások, mely menetét és irányelveit a későbbiekben fogjuk ismertetni.

A számítógéppel feldolgozható jelnyelvi korpusz létrehozásakor az egyik legnagyobb kihívás a magyar jelnyelv (manuális és/vagy nonmanuális komponensekből álló) elemeinek következetes azonosítása a korpusz egészében. Ennek a kérdésnek a megoldására a nemzetközi gyakorlatban kétféle megoldást találunk [vö. 10, 11, 18]. Egyik út az ún. notációs rendszerek használata, amelyek célja, hogy olyan pontos fonológiai leírást adjanak a jelekről, hogy azok kivitelezése lemásolható legyen. Ilyen notációs rendszerek kialakítása elsősorban a jelnyelvkutatás korábbi időszakára jellemző. A legismertebb közülük a HamNoSys rendszer, amelyet a hamburgi egyetem munkatársai fejlesztettek ki. A jelnyelvi lexikográfiában és korpuszelemzésben használt egyik szoftveres megoldás az iLex rendszer, melynek központi részét képezi a HamNoSys-ben történő átírás.

A másik megoldás a jelnyelvi írásrendszer hiányának kiküszöbölésére egy következetes jelölés alkalmazása, amely minden jelformát egyedileg azonosít. A jelnyelvi jelek egyedi formai azonosítóját ID-Glosszoknak nevezzük [22, 25, 12]. Mivel több ezer jelről van szó, ezért a gyakorlatban nem alkalmazhatunk tetszőleges kódrendszert (például számokat) – ez megnehezítené a gyakorlati felhasználást, ellehetetlenítené a keresést. Fontos megemlíteni, hogy habár az ID-Glosszok elnevezése utalhat az adott jel központi jelentésére, ez a megfeleltetés nem szükségszerű, de megkönnyítheti az adott kódhoz tartozó forma felidézését. Nem utal továbbá az ID-Glossz hangzó nyelvi szófaja az adott jel szófajára, annál is inkább, mivel a szófajiság megítélése különbözik a jelnyelvekben és a hangzó nyelvekben [37]. Jelnyelvek esetében kevésbé élesek a szófaji határok, a szófaji felosztásról pedig még nem született konszenzus a nemzetközi szakirodalomban.

2.2.3. Az ELAN szoftver alkalmazása

A projekt során áttekintett és mintául szolgáló jelnyelvi korpuszok (a holland [15], a brit [8], az ausztrál [22]) a hamburgi és a lengyel kivételével a Max Planck Institute által fejlesztett ELAN szoftvert használják, amely lehetővé teszi multimédiás anyagok

[1] Természetesen ezek mellett is számos alternatív lejegyzési módszer használatos, melyeket a gyakorlati igény hívta életre, gyakran találkozhatunk velük a jelnyelv mint idegen nyelv képzések során akár a diákok, akár az oktatók esetén.

annotálását. Alkalmas egyszerre több videó párhuzamos lejátszására, ez különösen fontos a jelnyelvi annotáció szempontjából. Maximálisan négy kamerakép egyidejű megtekintését biztosítja, valamint lehetőség van a felvételek utólagos összeszinkronizálásra abban az esetben, ha a felvételeket nem egyszerre indították. A program hátránya, hogy (az iLex-el szemben) nem kapcsolódik közvetlenül lexikai adatbázishoz, azonban – köszönhetően annak, hogy az ELAN szabad forráskódú – a készülő szótár és a szótár mögött álló lexikai adatbázis közötti kommunikációt sikerült megoldanunk.[2]

Az egyes elemzési szempontok külön szinteken, úgy nevezett tierekben jelennek meg, pl. kézforma vagy mozgás. A különböző adatokat tartalmazó tierek száma végtelen lehet.

Sem az ELAN-nak, sem az iLexnek nem volt magyar nyelvű változata a JelEsély projekt kezdetén, annak ellenére, hogy számos más nyelven elérhetőek. Az akadálymentesítés biztosításának érdekében elkészült a magyar fordítás, amely jelenleg még a mindenki által használt funkciókra tér ki, a bonyolultabb keresési és néhány egyéb, ritkán használt funkció fordítása még nem történt meg.

Az ELANban bizonyos elemzési szinteken az annotátorok egy legördülő listából kiválaszthatják az annotációs értékeket, ezeket a listákat kontrollált szótáraknak (controlled vocabularies, a továbbiakban CV) nevezzük. A CV-k nagy segítséget jelentenek a következetes annotálás elősegítésére, valamint elkerülhetőek az elütések is általuk. Használatuk azonban megköveteli, hogy az annotáció kezdete előtt meghatározzuk az adott kategória lehetséges elemeit. Az ELAN eredetileg nem teszi lehetővé a kontrollált szótárak értékeinek módosításait a munka kezdetét követően, azonban más projektek saját fejlesztésű scriptjei ezt a problémát már megoldották.

A jelnyelvi korpuszok létrehozásánál megkerülhetetlen az elemzési szempontok előzetes összeállítása. A JelEsély projekt grammatikai és szociolingvisztikai munkacsoportjaival együttműködve jött létre három sablon, melyek tartalmazzák azoknak az elemzési szinteknek a listáját, melyeket a magyar jelnyelv és (jel)nyelvhasználat vizsgálatakor előzetesen fontosnak tartottunk. A jövőbeni kutatásokhoz összesen 140 különféle elemzési szempontot határoztunk meg résztvevőnként (a szociolingvisztikai-grammatikai, célzott grammatikai és szótári annotáció során). Ezek egymással részben kompatibilisek, és van lehetőség a későbbi egyesítésre.

2.2.4. Az annotáció kihívásai

Annak ellenére, hogy a projekt során külön kezeljük a szociolingvisztikai és a grammatikai korpuszt, továbbá, hogy ezek feldolgozása más-más módon és céllal kezdődött el, hosszú távon mindkettő feldolgozásakor ugyanazokkal a kihívásokkal szembesülünk. A következő szakaszokban ezeket a kihívásokat foglaljuk össze, a jelenlegi állapotot bemutatva, függetlenül attól, hogy az eddigi munkánk során melyik részkorpusszal kapcsolatban merültek fel.

[2] Az ELAN-ban az ID-Glosszok listája a szótári adatbázisból frissíthető. Ez jelenleg csak egyirányú szinkronizációt jelent, az optimális ugyanakkor az lenne, ha az ELAN-ban megadott, új ID-Glosszok is bekerülnének a szótári adatbázisba, amely megfelelő ellenőrzési protokoll után megjelenhetne a szótári felületen is.

Az annotátorok és fordítók kiválasztásakor is fontos volt a siket közösség tagjainak lehető legnagyobb mértékű bevonása. A terepmunka és a további kutatási feladatok tervezéséhez hasonlóan itt is fontos volt, hogy az egyéni kompetenciákra és preferenciákra építve (a magas fokú magyar jelnyelvi kompetencia mellett a magyar nyelvtudás, illetve megfelelő számítógépes ismeretek voltak szükségesek) osszuk szét a feladatokat az annotátorok között. Külön nehézség volt a szociolingvisztikai annotáció során a potenciális CODA (Child of d/Deaf Adult, siket szülő halló gyermeke) munkatársak felkutatása. Később nagyothallók és a közösség által elismert tolmácsok bevonása jelentett megoldást. A szociolingvisztikai anyagok lejegyzése során próbáltunk alkalmazkodni a lejegyzők igényeihez (voltak, akik számára a fordítás azonnali gépelése volt a gyorsabb, míg mások a diktafonba fordítást preferálták). Hasonló elvek alapján kerültek kiválasztásra a grammatikai annotációt végző munkatársak is.

Kiemelten fontos volt az annotátorok oktatása annak érdekében, hogy megismerjék, és készségszinten tudják kezelni az annotációhoz használt szoftvert; valamint, hogy megértsék a feladatot, biztosítandó az annotáció következetességének megőrzését. A formális oktatás mellett folyamatosak voltak az informális megbeszélések, továbbá több feladatspecifikus leírás is készült számukra.

A legtöbb annotátor nem a Többnyelvűségi Kutatóközpontban végezte a munkáját, hanem otthonról. Jelenleg még nem épült ki nagymennyiségű videófájlok kezelésére és mozgatására alkalmas hálózat, ennek megvalósítását a későbbiekben tervezzük, mivel ennek hiányában az annotáció (főként kiadott fájlok és feladatok) dokumentálása, folyamatkövetése nagy adminisztratív terhet jelent.

A jelnyelvi videók hangzó nyelvre való fordítása során több elméleti és módszertani problémával szembesültünk, melyek közül néhányat már érintettünk. Annak ellenére, hogy a fordítói protokoll készítésekor törekedtünk a feladat pontos leírására, a jelnyelvi fordítás – hasonlóan a hangzó nyelvihez – nem törekedhet arra, hogy egyszerre adja vissza a jelnyelvekre jellemző sajátos mondatszerkezetet és jelentésalkotási stratégiát; valamint a mondat jelentésének megértéséhez szükséges magyar nyelvtani rendszert követő fordítást. Ez az elméleti probléma a gyakorlatban azt jelentette – annak ellenére, hogy CODA (siket szülő halló gyermekeként felnőtt, esetünkben mindkét nyelven magas kompetenciájú személy), vagy a közösség által elfogadott tolmács végezte a fordítási munkákat –, hogy több munkatárs nem vállalta a feladatot, vagy első elvállalás után nem folytatták a munkát. Ez elsősorban azzal magyarázható, hogy a jelnyelvi sajátosságokat visszaadó, jelről jelre haladó magyar fordítást kértünk a fordítóktól, nem pusztán tartalmi fordítást. Ez pedig olyan feladat, amellyel a legritkább esetben találkoznak mindennapos nyelvi környezetükben a tolmácsok és a CODÁ-k is. Az annotátorok és a fordítók egyéni kompetenciáihoz nagymértékben kellett alkalmazkodni a fordítás során, bizonyos esetekben még akkor is, ha ez módszertani problémákat is felvetett. A kutatás során a hosszú távú cél, hogy az annotációhoz használt szoftver felületén megjelenve a fordítások időben összekapcsolódjanak a releváns beszédeseménnyel (jelelési eseménnyel). Fontos volt továbbá szem előtt tartani, hogy a projekt szűk időkerete megkövetelte a gyors munkavégzést. Emiatt döntöttünk később úgy, hogy a számítógépet nem jól kezelő annotátorok diktafonba fordítsák a jelnyelvi videók anyagát, ami pedig később kerüljön begépelésre. Ez ugyan nem alkalmas a videókkal való azonnali összekapcsolásra, ugyanakkor nagymértékben meggyorsította a munkát. A projekt szellemiségével összhangban a gépelők között

látássérült munkatársak bevonására is sor került, emellett a számítógépet készségszinten használó, és nagy sebességgel gépelő munkatársak ELAN oktatása is folyamatban van. Kidolgozásra került továbbá az az eljárási mód, ahogyan a különböző szövegformátumú (de a videókkal nem összekapcsolt) fordítások ELAN-ba importálhatóak, ahol már a felvételekkel összekapcsolva, időkódokkal jelennek meg. Mivel szövegszerkesztőkben nem jeleníthető meg párhuzamosan a jelnyelvi változat és a fordítás, ezért a fordítások ellenőrzése problémát jelentett. Az ELAN-ba való későbbi importálás során a fordítások újraellenőrzésére és megfelelő szegmentálására sort kell keríteni.

Az általánosan elfogadott jelnyelvi írásrendszer hiánya mellett számos további problémával szembesültünk, amely a jelnyelvek sajátosságaiból adódnak. Ilyen alapvető kérdéskör a jel kezdetének és a jel végének a meghatározása, amely a videóanyagok tokenizálása során jelentkezett. Annak ellenére, hogy nincs egységes álláspont a nemzetközi szakirodalomban ezzel kapcsolatban sem, szükséges volt meghatározni, hogy az annotátorok milyen kritériumok alapján járjanak el a szegmentáció során. A későbbiekben tervezzük ennek a felülvizsgálatát, ellenőrzését is. A jelszegmentáció alapvető kérdése, hogy a jelelést folyamatos jelfolyamnak (ahol egy jelhez nemcsak az ún. tiszta fázis, hanem az átvezető mozgások is hozzátartoznak), vagy *jel→átmeneti mozgás→jel* folyamnak tekintjük. Számos oka van annak, hogy végül az első lehetőség mellett döntöttünk. A legfontosabb, hogy ne egy előre meghatározott konstrukcióval közelítsünk az egyik legfontosabb jelnyelvi elem felé, ne egy adott elméleti elgondolás mentén tekintsünk egy jelenséget jelnek, míg egy másik jelenséget átmeneti mozgásnak, hanem valóban alulról-felfelé építkezve, az adatokból elindulva határozzuk meg a jel fogalmát.

Ezek alapján „tág" szegmentumokat hoztunk létre, tehát a jel akkor kezdődik, amikor a kéz vagy kezek irányváltást kezdenek, miután az előző jel kivitelezéséhez szükséges összes mozgást befejezték ÉS/VAGY amikor a kéz vagy kezek elkezdik megváltoztatni a kézformát, ha az nem része az előző jel artikulációjának. A jelnek vége van: (1) Még mielőtt a kéz vagy kezek elkezdenének irányt változtatni, miután befejezték az aktuális jel kivitelezésének összes releváns mozgását ÉS/VAGY (2) még mielőtt a kéz vagy kezek elkezdenék megváltoztatni a kézformát, ha az nem része az előző jel artikulációjának. Továbbá (3) amikor a kéz vagy kezek elkezdenének visszatérni a pihenési pozícióba (pl. keresztbe tett karok, kezek a csípőn, vagy karfán, vagy a test mellett.). A kéz vagy kezek kivitelezési helyen való megállítása és pihentetése (a kézforma megtartásával) a jel részét képezi. A szakasz addig tart, amíg a „pihenés" véget ér, és a kéz vissza nem tér a nyugalmi helyzetbe vagy el nem mozdul egy következő jel kivitelezése felé. A félbehagyott jeleket, és minden kezekkel kapcsolatos jelenséget szegmentálni kell (ez alól kivétel a nyelvileg nem értelmezhető cselekvés). Hezitálásokat, szókereséseket és egyéb (feltehetően) megakadás-jelenségeket is szegmentálni kellett.

További alapvető problémát jelent a magyar jelnyelv kézforma-állományának a kérdése. A magyar jelnyelv szublexikális szintjeinek leírására korábban született monográfia természetesen foglalkozik a kézformák kérdéskörével is: [42], [43], de a probléma tisztázását célzó további vizsgálatok még folyamatban vannak. A magyar jelnyelvben használt, fonémának tekinthető kézformák meghatározása nélkül nem lehetséges a jelenségek következetes jelölése, ráadásul ennek a kérdésnek nagy jelen-

tősége van a sztenderdizációs folyamat egészét és a hallásállapottól független módon értelmezett jelnyelv-tanulói közösséget nézve is.

A jelnyelv fonológiai[3] komponensei, tehát a kézkonfiguráció (kézforma és kézformaváltás, orientáció, érintkezés testrésszel vagy másik kézzel, egy- vagy kétkezes) mellett a mozgás, a kivitelezési hely, a nonmanuális elemek, valamint orális elemek (szájkép) vesznek részt a jelnyelvi produkcióban [42]. A nyelvleírásnak csakúgy, mint a korpuszépítésnek alapvető feladata meghatározni a fenti kategóriák lehetséges értékeit (például a lehetséges mozgástípusokat). A külföldi jelnyelvi korpuszmunkálatok és grammatikai leírások, valamint egyéb nem nyelvészeti, de releváns kutatások alapján (pl. emócióelemzés és gesztuskutatás) meghatározott elemek, illetve a hazai siket közösség képviselőinek meglátása alapján dolgoztuk ki ezeknek a kategóriáknak a rendszerét, melyek az annotáció jelen szakaszában tesztfázisban vannak.

A jelnyelvi korpuszok létrehozásának és annotálásának számos hasonlóan új területe van, amelyekre jellemző, hogy több esetben empirikusan nem igazolt állítások, csoportosítások és hipotézisek várnak tesztelésre. Annak érdekében, hogy az annotálást végző munkatársak egy következetes segédlethez hozzáférjenek, létrehoztunk egy ún. annotációs vitaanyagot, amely tartalmazza egyrészt a munkafolyamat protokollját, másrészt eligazítást ad a jelnyelvi annotáció néhány kérdésében (a lexikális és fél-lexikális jelek és a nonmanuális komponensek, ismétlések és az artikuláció annotálása, stb.) Másik célja, hogy az annotációt tervező munkatársak közös referenciaanyagot hozzanak létre, amelyben az egyes nyelvi elemek annotációját megvitathatják. Ahogy a neve is sugallja, ez a dokumentum nem tekinthető véglegesnek. Az annotációs vitaanyag több hasonló külföldi anyag mintájára készült el [26, 9, 14, 43], elsősorban Trevor Johnston korábban hivatkozott anyagán alapul, amelyet évről-évre frissítve elérhetővé tesz, és az ausztrál jelnyelvi korpusz annotációja során használják.

A korpuszannotáció ciklikus volta lehetővé és szükségessé is teszi a projekt indulásakor meglévő tudásunk újraértelmezését. Az új ismeretek, új kihívások lehetővé teszik az annotációhoz kidolgozott rendszer folyamatos fejlesztését, fejlődését.

2.3 A korpusz felhasználási lehetőségei

A korpusz széleskörű felhasználási lehetőségeit röviden már érintettük a 2.1. fejezetben. A kereshető, immár géppel is olvasható korpuszok számos további kutatási lehetőséget biztosítanak, az alapvető statisztikai vizsgálatokon túlmenően is. A szociolingvisztikai kutatások (pl. területi és társadalmi változatosság) mellett lehetővé válik korpuszalapú szótár létrehozása, melynek során kiemelkedően fontos irányelv a jelnyelv-központúság; valamint egy valós nyelvhasználaton alapuló grammatika megalkotása is. Vizsgálhatóak továbbá diskurzusjelenségek, pragmatikai sajátosságok és a siketes jelek is. A korpusz ezen kívül oktatási célokat is szolgálhat, például tan-, és segédanyagok létrehozásával.

[3] Stokoe a fonológia, fonéma és allofón mintájára bevezeti a kerológia, keréma, alloker fogalmakat [41], de ezt a megkülönböztetést később ő maga sem látja szükségesnek. Egyrészt a közös fogalomrendszer rávilágít a hangzó nyelvek és jelnyelvek közös vonásaira, valamint ezek a fogalmak ugyanolyan adekvátak és megfelelőek a jelnyelvek leírásakor is, mint hangzó nyelvek esetében [6].

A jól annotált, számítógéppel feldolgozható korpusz a szótárkészítés alapja lehet, a felvételekből származtatott jeladatbázis a nemzetközi kutatási normáknak megfelelve jelnyelv-központú szótár létrehozását teszi lehetővé. A korpusz szociolingvisztikai vizsgálatokra, területi- és társadalmi, illetve rejtett változó mentén való vizsgálatokra is alkalmas, amennyiben mind a jelnyelvi szöveg, mint az interjú metaadatai rendelkezésre állnak a vizsgálathoz. Az előállt adatbázis korpuszalapú, valódi nyelvhasználatból származtatott grammatika készítését teszi lehetővé, hiszen alapot jelenthet a jövőben az összes nyelvi szint vizsgálatára a fonológiától a pragmatikáig.

A korpusz alapvető funkciója a magyar jelnyelv archiválása, mivel egyedülálló értékként bír a kortárs magyar jelnyelvhasználatot tekintve, a jövőben pedig történeti anyagként szolgál, így a későbbiekben lehetséges lesz a magyar jelnyelv különböző szintjein történő változások vizsgálata. Az ELAN-ban annotált korpuszban futtathatóak az egyes jelformákra való önálló keresések, mivel az annotáció során létrehozott címkét összekapcsolja a videó megfelelő szegmensével.

A korpusz kiváló tanítási anyagként is felhasználható, pl. a siketek oktatásakor, számtalan formában (segéd-, és példaanyagok, siket kultúra tantárgy, nyelvtan, stb.). A hallók jelnyelvoktatásának fejlesztésében is kulcsszerepe van a jelnyelvhez való hozzáférés kérdésének, a jelnyelvet hallóként, idegen nyelvként tanulók – a kezdő szinttől a tolmácsszintig, az egyszeri érdeklődőtől a siket gyermekek halló szüleiig – nagy hasznát vehetik a korpusz anyagának.

A korpusz kiindulási alapja lehet az automatikus jelfelismerő rendszereknek, illetve a számítógépes jelnyelvi modellezésnek, mivel természetes változatossággal rendelkező nyelvi anyagról van szó.

Felsőoktatásban a nyelvészeti terepmunka és a korpusznyelvészet órákon különösen, de antropológiai és minden egyéb, terepfelvételeket használó tudományterületnek kiváló példát szolgáltat a Jel*E*sély projekt jelnyelvi korpusza az adatkezelésre, adatfeldolgozásra. A korpusz nyelvi adatait az ELAN szoftverben megtekintve lehetővé válik, hogy egy időben 4 kamerakép vizsgálatával a legaprólékosabban megfigyeléseket tegyük a magyar jelnyelv jelenségeivel kapcsolatban. Kiváló konvertálási adottságok jellemzik a korpuszt (példamondatok exportálhatóak a jelnyelv tanításhoz, valamint a nagy sebességű videó feliratozás is lehetségessé vált). A korpuszból emellett számos alapvető statisztika kinyerhető.

3. Résztvevők

A projekt során számos terület szakemberei (szociolingvisták, elméleti nyelvészek, pszichológusok, szociológus, jogász stb.) dolgoztak együtt. A terepmunkák során kizárólag siket terepmunkásokkal dolgoztunk, az annotálás/fordítás/lejegyzés folyamataiban pedig siket, nagyothalló, CODA és halló munkatársak dolgoztak együtt, összesen 35-en. A részfeladatok összehangolásához precíz és részletes dokumentálásra volt szükség, valamint a jelnyelvi tolmácsokkal való állandó együttműködésre.

4. További tervek és feladatok

A korpuszhoz kötődő munkálatok során a jövőben is további számos kihívással kell szembenéznünk. Ilyen például a széleskörű annotálás és az ID-Glossz adatbázis kidolgozása, valamint a nemzetközi jelnyelvi korpuszokkal való átjárhatóság megteremtése.

A következő fontos lépés az ID-Glossz adatbázis kidolgozása, amely nemcsak biztosítja a könnyű keresést, de a későbbiekben a szótárépítéshez is nélkülözhetetlen. Ennek a folyamatnak a szótárkészítési vonatkozása a lemmatizáció, amelynek a jelnyelvekre alkalmazható nemzetközi standardjai még nem adottak. Álláspontunk szerint a lemmatizáció elveinek kidolgozása is csak a siket közösség bevonásával lehetséges.

Ahogy már említettük, fontos további feladat a szociolingvisztikai korpuszanyagok fordításainak ELAN-ba való átemelése és azoknak a videók megfelelő szegmenseivel való összekapcsolása.

A módszertanában nemzetközileg is úttörőnek számító *JelEsély* projekt a magyar jelnyelv átfogó, korpuszalapú grammatikai leírásával, korpuszával és szótárával e kutatások nélkülözhetetlen kiindulását jelentik a minőségi kétnyelvű oktatás elméleti, módszertani és gyakorlati feltételrendszere meghatározásának és az új oktatási program kimunkálásának.

Köszönetnyilvánítás

A tanulmányban leírtak nem valósulhattak volna meg a Jelesély Projekt (Támop 5.4.6/B-13/1-2013-0001) támogatása nélkül. Köszönetet mondunk a Jel*E*sély projekt megvalósítóinak, valamennyi siket és halló munkatársnak, különösen a technológiai előkészítésében és archiválásában résztvevő Tarr Zoltánnak és Gál Ferencnek, valamint a kontrollált szótárak elemeinek kidolgozásában nyújtott támogatásáért Szabó Mária Helgának.

Hivatkozások

1. Bartha, Cs: A kétnyelvűség alapkérdései. Nemzeti Tankönyvkiadó, Budapest (1999)
2. Bartha, Cs., Hattyár, H.: Szegregáció, diszkrimináció vagy társadalmi integráció? – A magyarországi siketek nyelvi jogai. In: Kontra, M., Hattyár, H. (eds.): Magyarok és nyelvtörvények. Teleki László Alapítvány, Budapest (2002) 73–123
3. Bartha, Cs.: A Kárpát-medencei kisebbségi magyar nyelvi korpusz. Korpuszépítési és kutatási lehetőségek. Kézirat. MTA Nyelvtudományi Intézet, Budapest (2002)
4. Bartha, Cs.: Siket közösség, kétnyelvűség és a siket gyermekek kétnyelvű oktatásának lehetőségei. In: Ladányi, M., Dér, Cs., Hattyár, H. (eds.): „…még onnét is eljutni túlra…". Nyelvészeti és irodalmi tanulmányok Horváth Katalin tiszteletére. Tinta Könyvkiadó, Budapest (2004) 313–332
5. Bartha, Cs., Hattyár, H., Szabó, M. H.: A magyarországi siketek közössége és a magyarországi jelnyelv. In: Kiefer, F. (ed.): Magyar Nyelv. Akadémiai Kiadó, Budapest (2006) 852–906

6. Battison, R.: Analysing Signs. In: Valli, C., Lucas, C.: (eds.) Linguistics of American Sign Language: An Introduction. Gallaudet University Press, Washington (2000) 199–218
7. Bea – Magyar Spontán Beszéd Adatbázis http://www.nytud.hu/adatb/bea/index.html (é.n.)
8. Cormier, K., Fenlon, J., Rentelis, R., Schembri, A.: British Sign Language Corpus Project: A corpus of digital video data of British Sign Language 2008–2011. University College London, London (2011)
9. Cormier, K., Fenlon, J., Gulamani, S., Smith, S.: BSL Corpus Annotation Conventions (2015) http://www.bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v2_-Feb2015.pdf
10. Crasborn, O., Sloetjes, H., Auer, E., Wittenburg, P.: Combining video and numeric data in the analysis of sign languages within the ELAN annotation software In: Vettori, C. (ed): LREC 2006, II. Workshop proceedings. Representation and processing of sign languages. ELRA, Paris (2006) 82–87
11. Crasborn, O., Sloetjes, H.: Enhanced ELAN Functionality for sign langauge corpora In. Crasborn, O., Hanke, T., Efthimiou, E., Thoutenhoofd, E. D., Zwitzserlood, I.: Proceedings of LREC 2008, Sixth International Conference on Languge Resources and Evaluation (2008) 39–43
12. Crasborn, O., de Meijer, A.: From corpus to lexicon: the creation of ID-glosses for the Corpus NGT In: Crasborn, Efthimiou, Fontinea, Hanke, Kristoffersen, Mesch (eds.): Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (2012) 13–17
13. Crasborn, O.: „Sign Language Corpora." Sign Language Corpora Wiki. Online: http://sign.let.ru.nl/groups/slcwikigroup/wiki/7f8aa/sign_language_corpora.html (2013) (2014. 03. 08)
14. Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., de Meijer, A., Sáfár, A.: Annotation Conventions for the Corpus NGT. (2015) http://www.bslcorpusproject.org/wp-content/uploads/CorpusNGT_AnnotationConventions_v3_Feb2015.pdf
15. Crasborn, O., Zwitserlood, I., Ros, J.: Corpus NGT. An Open Access Digital Corpus of Movies with Annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen. [Available at: http://www.ru.nl/corpusngt] (én) (2015.12.03)
16. DGS-Korpus. Online: http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dgs-korpus.html (é.n.) (2014. 03. 09)
17. ELP, Endangered Languages Project: Corpus of grammar and discourse strategies of deaf native users of Auslan (Australian Sign Language). http://www.hrelp.org/grants/projects/index.php?lang=9 (é.n.)
18. Hanke, T.: HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts In: Streiter, O., Vettori, C. (eds): LREC 2004, Workshop proceedings. Representation and processing of sign languages. ELRA, Paris (2004) 1–6
19. Hattyár, H.:. A siketoktatás elméleti és gyakorlati kérdései. Educatio 9. (2000) 776–790
20. Hattyár, H.: Jelnyelvek – Természetes emberi nyelvek eltérő modalitással. In: Ladányi, M., Dér, Cs., Hattyár, H. (eds.): „…még onnét is eljutni túlra…". Nyelvészeti és irodalmi tanulmányok Horváth Katalin tiszteletére. Tinta Könyvkiadó, Budapest (2004) 342–346
21. Hattyár, H.: A magyarországi siketek nyelvelsajátításának és nyelvhasználatának szociolingvisztikai vizsgálata. Doktori Disszertáció ELTE BTK, Budapest (2008)
22. Johnston, T.: The lexical database of Auslan (Australian Sign Language). Sign Language & Linguistics (2001) 145–169
23. Johnston, T., Schembri, A.: Australian Sign Language (Auslan): An introduction to sign language linguistics. Cambridge University Press, Cambridge (2007)

24. Johnston, T.: The Auslan Archive and Corpus. In D. Nathan (ed.): The Endangered Languages Archive—http://elar.soas.ac.uk/languages. Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London, London (2008)
25. Johnston, T.: From archive to corpus: transcription and annotation in the creation of signed language corpora. In: Roxas, R. (ed.): 22nd Pacific Asia Conference on Language, Information, and Computation. De La Salle University, Cebu, Philippines (2008) 16–29
26. Johnston, T.: Auslan Corpus Annotation Guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University, Sydney, Australia (2014)
27. Kárpád-medencei Magyar Nyelvi Korpusz: http://corpus.nytud.hu/mnszworkshop/index.html (2006)
28. Kendall, T. On the History and Future of Sociolinguistic Data. In: Language and Linguistics Compass (2008) 332–351
29. Konrad, R.: Sign Language Corpora Survey http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/SL-Corpora-Survey_update_2012.pdf (2012)
30. Kontra, M.: A Budapesti Szociolingvisztikai Interjú. MTA Nyelvtudományi Intézet, Élőnyelvi Kutatócsoport. Kézirat. Budapest http://buszi.nytud.hu/ (1987)
31. Lancz, E., Barbeco, S.: A magyar jelnyelv szótára. Siketek és Nagyothallók Országos Szövetsége, Budapest (1999)
32. Leech, G.: „Corpora and theories of linguistic performance." In: Svartvik, J. (ed.): Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter. (1992) 105–122
33. McEnery, T., Wilson, A.: Corpus Linguistics. Lancaster University, Lancaster (2001)
34. McEnery, T., Sebba, M., Burnard, L.: Minority Language Engineering (MILLE) – Summary Report (é.n.)
35. Nyelv- és beszédtechnológiai platform (sz.n.) http://www.hlt-platform.hu/online-adatbazisok.html
36. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of LREC 2014. http://clara.nytud.hu/mnsz2-dev/ (2014)
37. Pfau, R., Steinbach, M., Woll, B. (eds.), Sign language. An international handbook (HSK - Handbooks of linguistics and communication science). Mouton De Gruyter, Berlin (2012)
38. Rundell, M.: The corpus of the future, and the future of the corpus. Talk at 'New Trends in Reference Science' (1996)
39. SignWriting History. SignWriting® Site.: www.signwriting.org/library/history/history.html (é.n.) (2014.3.10.)
40. Sinclair, J.: EAGLES. Preliminary recommendations on Corpus Typology. (1996) http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html
41. Stokoe, W. Sign Language Structure: An Outline of Visual Communication Systems of the American Deaf. Studies in Linguistics: Occasional Paper No. 8. University of Buffalo. Buffalo, NY (1960)
42. Szabó, M. H.: A magyar jelnyelv szublexikális szintjének leírása. Akadémiai Kiadó, Budapest (2007)
43. Szabó, M. H., Mongyi, P.: A jelnyelv nyelvészeti megközelítései. Magyar Jelnyelvi Programiroda, Budapest (2005)
44. Wallin, L., Mesch, J., Nilsson., A-L.: Transcription guide lines for Swedish Sign Language discourse. https://www.diva-portal.org/smash/get/diva2:389066/FULLTEXT01.pdf (2010)

Jogszabályok hivatkozásainak automatikus felismerése és a belső hivatkozások struktúrája

Hamp Gábor[1], Syi[1], Markovich Réka[2, 3]

[1] BME Szociológia és Kommunikáció Tanszék 1111 Budapest, Egry József u. 1.
hampg@eik.bme.hu, i@syi.hu
[2] ELTE Filozófiatudományi Doktori Isk., Logika Tanszék 1088 Budapest, Múzeum krt. 4/I
[3] BME Üzleti Jogi Tanszék, 1111 Budapest, Magyar Tudósok körútja 2.
markovich@phil.elte.hu

Kivonat: A jogrendszert alkotó jogszabályok nem egymástól független létezők: hálózatot alkotnak – ebből és hierarchikus viszonyukból következően beszélhetünk rendszerről. Ebből következően a jogszabályokban sok hivatkozást találunk más jogszabályokra. De mivel a hivatkozás a másik szöveghely behívásával redundanciát kerül, nemcsak egy másik jogszabály szöveghelyére történő hivatkozásnak van értelme: az adott jogszabályon belül történő hivatkozás szintén gyakori. A jogszabályok automatikus logikai elemzésének aspektusait feltárni célzó kutatásunkban szükségessé vált a hivatkozások kezelése. Az automatikus hivatkozásfelismerés a logikai elemzés számára történő szövegelőkészítés mellett más, igen komoly gyakorlati jelentőséggel is bír.

A jogszabályok alkotta szövegkörnyezet folyamatosan változik: új jogszabályok születnek, mások részben vagy egészében hatályukat vesztik. A jogszabályi környezet változásai miatt a hatályban lévő jogszabályok hivatkozásai könnyen és gyakran elavulnak. Ebben a helyzetben elemi gyakorlati érdek, hogy az elavult hivatkozások feltárását, frissítését gépi eszközökkel lehessen támogatni. Mi a kutatásunkban a jogszabályok logikai elemzése automatizálásának aspektusait vizsgáljuk. Tekintettel arra, hogy a logikai elemzés a szöveg mondatszintű egységein kivitelezhető, a szövegelőkészítés részeként vált szükségessé – a gyakorlati jelentőséggel is bíró – hivatkozáskezelés.

E cél eléréséhez első lépésként a jogszabályi hivatkozások felismerését, valamint a teljes szövegen belüli pozícióinak gépi azonosítását kell elvégeznünk, majd a hivatkozásból ki kell nyerni a hivatkozott jogszabály vagy a hivatkozott jogszabályi szerkezeti egység azonosításához szükséges információt. A továbbiakban – ha az egyértelmű tárgyalásmód azt megkívánja – a hivatkozott jogszabályi egységet (a jogszabály egészét vagy annak egy részét) célnak (céljogszabálynak), a hivatkozás tartalmazó jogszabályi egységet forrásnak (forrásjogszabálynak) nevezzük. A forrásoldali tehát a hivatkozó, a céloldali a hivatkozott jogszabály (jogszabályi szerkezeti egység). Ha a forrás- és a céljogszabály egybeesik, belső hivatkozásról beszélünk.

Amennyiben nem esik egybe, azaz külső hivatkozással állunk szemben, a meghatározó lépések egyike a hivatkozott jogszabály beazonosítása. A jogszabályoknak – elméletileg – egyedi címük van, így egyértelműen lehet rájuk hivatkozni. A jogszabályszerkesztésről szóló rendelet pontosan előírja, hogyan kell a jogszabályok címét megadni, és hogyan kell rájuk hivatkozni [2]. Az elfogadott hivatkozási módok közül nem mindegyik alkalmas arra, hogy teljes biztonsággal fel lehessen ismerni valamely

jogszabály címét kizárólag szintaktikai mintázat alapján, ezért a jogszabály-hivatkozások felismeréséhez szükség van egy olyan külső címtárra, amely tartalmazza az összes jogszabály összes lehetséges címét. Ennek hiányában a jogszabályokra való hivatkozások felismerése nem garantált. Egy ilyen címtár webes források feldolgozásával előállítható [10], mi is létrehoztunk egy ilyen forrást, de ebben a tanulmányban a külső linkekkel érdemben nem, csak problémafelvetés szintjén foglalkozunk.

A jogszabályszerkesztésről szóló rendelet meghatározza a jogszabályok lehetséges szerkezeti egységeit és azok hierachiáját: alpont, pont, bekezdés, szakasz, alcím, fejezet, rész és könyv) [2]. A szerkezeti egységeket jelöléssel (címekkel, számokkal, egyéb jelekkel) látja el a jogalkotó – az egyértelmű azonosítás és ezzel az egyértelmű hivatkozási lehetőség érdekében. A hivatkozás egy konkrét jogszabály adott szerkezeti egységén belül helyezkedik el, és vagy egy jogszabályra teljes egészében, vagy egy (akár az azonos, akár egy másik) jogszabály adott szerkezeti egységére mutat. A hivatkozások gépi felismeréséhez ismernünk kell a hivatkozások lehetséges típusait, és egyértelmű eredményeket nyújtó azonosítási technikát kell használnunk a jogszabályokra és azok szerkezeti egységeire történő utalásokhoz.

A jogszabályszerkesztésről szóló rendelet III. Fejezete elkülöníti a *rugalmas*, valamint *merev hivatkozásokat*, azaz a tartalmi körülírással megadott, illetve az adott jogszabály azonosítóit tételesen megadó utalásokat, valamint megkülönbözteti a *belső*, illetve *külső hivatkozásokat*. Utóbbi felosztás szerint arra figyelnek, hogy az adott jogszabály saját magára, pontosabban saját magán belül egy másik szerkezeti egységre vagy egy másik jogszabály egészére vagy részére utal. A rendelet megadja ezeknek a hivatkozástípusoknak a használati szabályait és szintaxisát. A gépi elemzéshez azonban további tipizálásra is szükségünk van. El kell különíteni az *egyszerű* és a *halmozott hivatkozásokat*, valamint az *egyszeres* és *többszörös hivatkozásokat* egymástól. A halmozott hivatkozások több szerkezeti egységre utalnak a céloldalon (pl. *…a 37. § (2)-(5) bekezdésében…*). Többszörös hivatkozásokról akkor beszélünk, ha az adott szerkezeti egységben egynél több egyszerű vagy halmozott hivatkozás van (pl. *Az (1) bekezdés a) pontjában felsoroltak közösen, illetve az (1) bekezdés b)-d) pontjában meghatározottak közösen is gyakorolhatják a fenntartói jogokat.)*; ha egy szerkezeti egységen belül egyetlen hivatkozás van, akkor arra egyszeres hivatkozásként utalunk. Így háromféle hivatkozáshalmozással találkozhatunk. Létezik tehát

i) forrásoldali halmozott hivatkozás, amikor a forrásoldali szerkezeti egységben több, egymástól egyértelműen elkülöníthető hivatkozás szerepel (*a 80. § (2) bekezdésének a) pontja, valamint a 83. szakasz (1) bekezdésének f) pontja*),
ii) céloldali halmozott hivatkozás, mikor egy egyszerű forrásoldali hivatkozás több forrásoldali pontra mutat (*a 80. § (2) bekezdésének a)-c) és f) pontjai*),
iii) kétoldali vagy kétszeresen halmazott hivatkozás, amikor egyrészt a forrásoldalon több hivatkozás van egy szerkezeti egységen belül, másrészt a forrásoldali hivatkozások közül legalább egy többszörös céloldali hivatkozást tartalmaz (*a 80. § (2) bekezdésének a)-c) pontjai, valamint a 83. § (1) bekezdésének f) és g) pontjai*).

Az a tény, hogy a jogszabályok jelölt szerkezeti egységekből állnak, lehetővé teszi azt, hogy a jogszabályok bármelyik szerkezeti egységére egyértelműen tudjunk hivatkozni. Figyelni kell azonban arra, hogy a szerkezeti egységekhez rendelt jelölések nem feltétlenül egyediek, hiszen míg a szakasz számozása folyamatos, addig a pontok

számozása minden szakasz alatt, az alpontok számozása minden pont alatt újrakezdődik, így önmagukban nem alkalmasak az egyértelmű azonosításra. Tehát pl. 1. §-ból csak egy lehet egy törvényben, (1) bekezdésből több is. Az egyértelmű hivatkozást azáltal biztosíthatjuk, ha kihasználjuk a jogszabályok hierarchikus tagoltságát. A hierarchiára építve úgy hivatkozhatunk egyértelműen, hogy az adott szerkezeti egység jeléhez sorban hozzákapcsoljuk az őt magába foglaló összes szerkezeti egység jelét is, vagyis adott szerkezeti egységet azzal a *hierarchikus jelsorozattal* azonosítunk, amelyet az egymásba ágyazott szerkezeti egységekhez tartozó jelekből állítunk össze.

x. évi y-ról szóló z. törvény + 2. § + (1) bekezdés + a) pont + ab) alpont

A fenti kifejezéssel egyértelműen rámutathatunk a szóban forgó szerkezeti egységre annak ellenére, hogy a) pontból is, ab) alpontból is több lehet az adott jogszabályon belül.

Mivel a jogszabályok szerkezeti egységei meghatározott típusokba sorolhatóak, egy ilyen hierarchikus jelsorozat megadható azzal, hogy az egyes összetevői milyen típusba tartoznak és az adott típusnak milyen azonosítója van. A fenti példában szereplő jelsorozatot így formalizálhatjuk:

<j:123;s:1;b:2;p:a;a:ab>

ahol a ‘j’ a jogszabály, az ‘s’ a szakasz (§), a ‘b’ a bekezdés, a ‘p’ a pont, az ‘a’ az alpont mint szerkezetiegység-típusok jele, a kettőspont után következő szám- és betűkombinációk pedig a konkrét szerkezeti egységek – jogszabályban szereplő – jelei. Egy szerkezeti egységnek a fenti formalizmust követő azonosítója kinyerhető az adott szerkezeti egységben szereplő explicit, valamint a szerkezeti egység hierarchikus beágyazódásáról vonatkozó implicit információkból. Ha rendelkezésre áll egy jogszabálycímtár erőforrásként, akkor azt arra is használhatjuk, hogy a jogszabályok egyedi jeleként a címtárban található egyedi azonosítót használjuk ’j’ értékeként. Ha így teszünk, akkor az egymásba ágyazott szerkezeti egységek típusából és jeléből álló szimbólumkettősök sorozatával minden szerkezeti egységet egyedi módon azonosíthatunk. Ez a jelsorozat azért is fontos, mert ugyanezt a mintázatot előállíthatjuk az automatikusan megtalált és felcímkézett hivatkozásokból is. Amikor megtalálunk és elkülönítünk egy hivatkozást, azt összetevőkre bontva ugyanolyan módon felépíthetünk egy jelsorozatot, ahogy azt a szerkezeti egységek azonosításakor tesszük. Az elemzett jogszabályszövegekből kinyert hivatkozásokat jellemző jelsorozat értékeit pedig összehasonlíthatjuk a jogszabálykorpuszban található céloldali szerkezeti egységek jelsorozataival, és ha azok között megtaláljuk a keresett mintázatot, akkor összekapcsolhatjuk a forrást a céloldali szerkezeti egységgel. Ez az összekapcsolás természetesen csak akkor (és olyan mértékben) működik, amikor (és amilyen mértékben) megtalálhatóak a hivatkozott jogszabályok a korpuszunkban.

A gépi feltárás célja az, hogy a hivatkozásokat megtaláljuk a forrásoldalon, feltárjuk, hogy hány hivatkozás van az adott szerkezeti egységen belül, majd minden egyes hivatkozást felbontva beazonosítsuk azokat a jogszabályokat, illetve azon belüli szerkezeti egységeket a céloldalon, amelyekre a hivatkozások mutatnak. A feladat tehát az, hogy az elemzett jogszabályszövegen belül:

i) megtaláljuk a teljes hivatkozás kezdő- és zárópozícióját (többszörös hivatkozás esetén több értékpárt kell megtalálnunk),

ii) feltárjuk, hogy egyszerű vagy halmozott hivatkozásról van-e szó (utóbbi esetben elkülönítjük a hivatkozási célpontok közös, illetve egyedi összetevőit),
iii) azonosítjuk az egyes hivatkozásokban hivatkozott jogszabályt, és
iv) meghatározzuk a hivatkozott szerkezeti egység(ek) azonosító jelsorozatát.

A hivatkozás homogén abban az értelemben, hogy mind a forrásoldalon, mind a céloldalon szerkezeti egységeket kapcsol össze. Ez a homogenitás összhangban van azzal az tézissel, amely a jogszabályok legkisebb – hivatkozási – egységének a szerkezeti egységeket tekinti (és nem engedi meg az azokon belüli, például mondatokra irányuló utalásokat). A gépi hivatkozásfelismerés gyakorlatában azonban többféle inhomogenitással is találkozhatunk. Egy hivatkozás két végpontjának pontos meghatározása eltér a jogi korpuszon belül: a céloldalon csak a szerkezeti egységet adjuk meg (az őt azonosító jelsorozattal), a forrásoldalon ellenben meg kell adni a hivatkozás szerkezeti egységen belüli pozícióját is. Utóbbi esetben pontosabb lokalizálást kell elvégeznünk. Erre azért van szükség, mert egy szerkezeti egységen belül előfordulhat többszörös hivatkozás, amiket el kell tudnunk különíteni egymástól, és meg kell tudnunk mondani, hogy melyik hivatkozás hol helyezkedik el a szerkezeti egységen belül.

A hivatkozások inhomogének lehetnek abban az értelemben is, hogy az – egyetlen – forrásoldali szerkezeti egység több szerkezeti egységre mutat a céloldalon. Akkor fordulhat elő ilyen helyzet, amikor a hivatkozás olyan szerkezeti egységre mutat, amelynek vannak alárendelt szerkezeti egységei. A forrásoldalon pl. csak annyit adnak meg, hogy a céloldalon egy másik jogszabály valamely bekezdésére hivatkoznak (ebben az értelemben ez egy egyszerű hivatkozásnak számít), ám a hivatkozott jogszabály ismeretében kiderülhet, hogy a szintaktikai értelemben véve egyszerű utalás szemantikai értelemben valójában halmozott hivatkozásnak minősíthető, hiszen több szerkezeti egység – pont és alpont – tartozik a terjedelmébe. Ezeket nevezhetjük rejtett vagy látens halmozott hivatkozásnak. Ennek legnyilvánvalóbb esete az, amikor egy hivatkozás egy másik jogszabály teljes szövegére mutat.

A céloldali hivatkozások feloldásának további nehézségét jelentik azok az – igen gyakori – esetek, amikor egy szerkezeti egység egyszerre kétféle minőséggel is rendelkezik. A bekezdések esetében tipikus megoldás az, hogy ugyanaz a szerkezeti egység a bekezdés első szakasza is egyben. Az ilyen helyzetek kétféle problémát jelentenek. Egyrészt könnyen megtörténhet az, hogy az első szakaszra utaló jelölést nem teszik ki (például azért, mert a bekezdés csak egy szakaszból áll). A szerkezeti egység felismerő modulnak képesnek kell lennie arra, hogy kezelje az ilyen helyzeteket. Másrészt könnyen zavart okozhat az is, hogy a forrásoldalon egyszer úgy hivatkoznak az adott bekezdésre, hogy a bekezdés egészére akarnak utalni, máskor viszont csak a bekezdés első szakaszára (ami „fizikailag”, szintaktikailag ugyanaz a szerkezeti egység). Ezt a kétféle hivatkozási igényt csak úgy tudjuk kielégíteni, ha az ilyen bekezdésekhez (szerkezeti egységekhez) kétféle tipizálást rendelünk, és így kétféle hivatkozási lehetőséget biztosítunk.

A jogszabályok szerkezeti egységeire való hivatkozásnak szintaktikailag jól felismerhető szabályai vannak. Ezek a hivatkozások jobbról könnyen lezárhatóak: könnyen meg lehet találni a hivatkozási sor utolsó elemét. A hivatkozás kezdetének megtalálása a külső hivatkozások esetében jóval nehezebb feladat. Elvileg mindig a tartalmazza a jogszabály „nevét”. Korábban említettük, hogy a jogszabályokra (pontosabban a jogszabályok címére) való hivatkozás lehet merev vagy rugalmas, ami a

kontextusmentes vagy a kontextusfüggő hivatkozási technikának felel meg (mások ezt a dichotómiát direkt vagy relatív hivatkozásnak nevezik [8]). A kontextusmentes címhivatkozás felismerése elvileg viszonylag problémamentes kellene hogy legyen, hiszen a hivatkozás a jogszabály teljes címét tartalmazza, aminek szigorú – a jogszabályszerkesztésről szóló rendelet előírásait követő – mintázatát fel lehet ismerni, ám e felismerés parszolási techikákat igényel a jogszabály címének pontos beazonosításához. A rugalmas hivatkozás detektálása még nehezebb, hiszen a felismeréshez további információra, kontextusismeretre van szükség. A gyakorlatban minden belső hivatkozás kontextuális hivatkozásnak minősíthető, hiszen egy konkrét jogszabályon belül sosem fognak a szöveg egy másik pontjára a jogszabály teljes címével hivatkozni. Egyszerűbb eset, amikor az adott szövegben a jogszabály önmagára úgy hivatkozik, hogy „e törvényben", hiszen ekkor (és ebből) tudhatjuk, hogy a törvény itt saját magára hivatkozik (más kérdés, hogy vajon van-e értelme ezt valódi hivatkozásként kezelnünk). Hasonló módon ismerhetők fel azok a belső hivatkozások, amelyek alacsonyabb szintű szerkezeti egységre utalnak (pl *e bekezdés a) és b) pontja*). Bár a belső hivatkozások csak kontextuálisok lehetnek, azért ennek a kontextualitásnak lehetnek fokozatai. Vannak erős – kontextuális – hivatkozások, amikor adott hierarchikus szintre vonatkozó utalást egyértelműen adnak meg (*a 3. § (2) bekezdés a) pontja),* és vannak gyenge hivatkozások, amikor a hivatkozás valamely eleme relatív, kontextusfüggő (*e § (2) bekezdése)*.

A külső hivatkozások között vannak nem kontextuális hivatkozások *(a nemzeti felsőoktatásról szóló 2011. évi CCIV. törvény 3. § (2) bekezdése*), de előfordulhatnak köztük kontextuálisok is, amelyeken belül elkülöníthetünk további altípusukat. Gyakori megoldás, hogy a jogszabály elején bevezetnek a későbbiekben többször hivatkozni kívánt külső jogszabályra valamilyen rövidítést, és a továbbiakban ezt az alakot használják a linkek felépítésekor. Ezek a rövidítések gyakran olyan alakúak, amelyek más jogszabályokban is előfordulnak, az adott jogszabály egyedi azonosítójaként, rövid címeként szolgálnak. A kontextualitás erősebb fokát jelenti az a megoldás, amikor olyan rövidítést alkalmaznak, amely csak az éppen adott jogszabályban érvényes, és semmiképpen sem vihető át jogszabályról jogszabályra. Beszédes példája ennek a *Módtv.* rövidítés, ami mindig egy módosító törvény alkalmilag bevezetett – és ezáltal – alkalmi rövidített megnevezése, azonban módosító törvényből nagyon sok van, és ez a minősége egyrészről csak az adott kontextusban érdekes, másrészről a rövidítés jelölete ennek megfelelően változó.

Fentebb bemutattuk, hogy a hivatkozásokat és a szerkezeti egységeket ugyanazzal a logikával bonthatjuk fel komponensekre, ugyanazzal a technikával azonosíthatjuk egyedi módon. Az azonosításban kiemelt szerepe van az azonosító jelsorozat kezdő értékének, amely a szerkezeti egységet magába foglaló jogszabály egészére utal. Célszerű ezt a kitüntetett szerepet azzal jelezni, hogy ezt az első komponenst nevezzük a hivatkozás *fejének*. A hivatkozások tipizálásakor, a kontextusok kezelésekor a hivatkozás fejében található információt hasznosíthatjuk. Az, hogy külső vagy belső hivatkozásról van-e szó, csak akkor derül ki, amikor már megtaláltuk a hivatkozás fejét, vagyis már ismerjük a hivatkozásba foglalt jogszabály címét. Ha ez megegyezik az éppen elemzett jogszabály címével, akkor belső hivatkozásról van szó, ha eltér a két cím egymástól, akkor külső hivatkozásról beszélhetünk.

A halmozott hivatkozásoknak több típusa lehet. Előfordulhat olyan halmozás, amikor egymástól teljesen elkülöníthető, semmilyen átfedést, közös részt nem tartalmazó

egységeket sorolnak fel egymás után. Ilyen esetben mindegyik hivatkozást önmagában lehet feltárni. Azok a halmozott hivatkozások, amelyek ugyanazon jogszabály több, különböző helyére mutatnak, egyfajta hivatkozásfelsorolásnak is tekinthetők. Az ilyen felsorolásoknak több típusa is előfordulhat. Lehetnek „folytonos" hivatkozások, amikor szigorúan egymás után követő, azonos típusú szerkezeti egységek sorozatára utalnak egy kötőjeles rövidítéssel: *a 23-37. §.* Az egyértelmű céloldali lokalizálás érdekében az ilyen sorozatot fel kell bontani elemeire, hogy minden egyes szerkezeti egységre – külön-külön – rá tudjunk mutatni.

A felsorolás másik fajtája a diszkrét hivatkozási sor, amikor folytonos sorozatba nem rendezhető elemeket vesszővel elválasztva sorolnak fel: *a 16. § (2), (3), (5) bekezdése*. Ilyenkor könnyebb feladat felbontani a sorozatot elemi egységekre, de ekkor is meg kell tenni, hogy a célirányban egyértelmű utalásokhoz juthassunk.

Ez a jelenség nem keverendő össze azzal a – korábban leírt – helyzettel, amikor olyan látens halmozott hivatkozásról van szó, amely a forrásoldalon egyetlen, fel nem bontható hivatkozást találunk, ami viszont a céloldalon mégis több szerkezeti egységet fog össze, mert a hivatkozás magasabb hierarchikus szintű egységre mutat.

Az alkalmazott technika

A jogszabályok szerkezeti egységeinek tipizálására, a szerkezeti egységek közti hierarchikus alárendeltségi (szülő-gyerek) kapcsolatok feltárására, a szerkezeti egységeket egyértelműen azonosító hierarchikus jelsorozat képzésére, a Goody-listák megtalálására [5, 7], tipizálására, illetve a Goody-mondatok előállítására reguláris kifejezéseket használunk.

Speciális hivatkozások, kivételkezelés

A normál, szabályos hivatkozások mellett olykor előfordulnak speciális, atipikus utalások is a jogszabályokban. Ezek egy része – különböző okok miatt – nem is igazán tekinthető valódi hivatkozásnak. Vegyük sorba őket.

A hivatkozások között előfordulhatnak *egymásba ágyazott hivatkozások*. Ilyen esetekben egy adott jogszabály adott szekezeti egységére való hivatkozáson belül egy másik jogszabály valamely szerkezeti egységére mutató hivatkozást találhatunk. Íme, egy példa erre: *e törvénynek a 2012. évi CCVIII. törvénnyel megállapított 85/A. § (1) bekezdése*. Az idézetet úgy kell értelmeznünk, hogy egy belső hivatkozásban (*e törvénynek a ... 85/A. § (1) bekezdése)* szerepel egy külső *(2012. évi CCVIII. törvénnyel*) hivatkozás (amely épp itt a törvény egészére, nem pedig annak egy szerkezeti egységére mutat).

Vannak jogszabályok (tipikus módon törvények), amelyekben – a jogszabály által szabályozni kívánt területhez tartalmilag egyáltalán nem kapcsolódva – a jogszabály valamelyik szerkezeti egységében megváltoztatják egy másik jogszabály valamely szerkezeti egységének a szövegét. Ezek a betétek tartalmilag függetlenek a „befogadó" szöveghez képest, ezért vendégszövegeknek tekinthetőek, az ilyen hivatkozásokat ezért nevezhetjük *vendégszöveg-hivatkozásoknak*. Nézzük meg a következő példát:

> (2) A Magyarország helyi önkormányzatairól szóló 2011. évi CLXXXIX. törvény 109. § (2) bekezdése a következő szöveggel lép hatályba:
> „(2) A képviselő-testület kizárólag a nemzeti vagyonról szóló törvényben meghatározott személyekkel köthet vagyonkezelési szerződést." [3]

Ez a sajátosság azért említésre méltó a hivatkozások feldolgozása szempontjából, mert a vendégszövegekben szereplő belső hivatkozásokat a befoglaló jogszabály szempontjából külső hivatkozásoknak kell tekinteni, míg a beágyazott jogszabály, a vendégszöveg felől tekintve ezeket belső hivatkozásoknak lehetne minősíteni.

Nem túl gyakran, de előfordul olyan *általános hivatkozás*, amikor nem konkrét, egyedi (partikuláris) jogszabályra utal a jogszabály szövege, hanem valamilyen általános, nem egyedi hivatkozást tesz: *A szexuális szolgáltatás fogalmát és reklámozásának további korlátozásait külön törvény állapítja meg.* A 'külön törvény' típusú utalás annyit jelent, hogy a szóban forgó fogalom (itt a szexuális szolgáltatás fogalma) a jelen törvényben nincs az értelmező rendelkezések között, de a jogrendszer által ismert, kezelt fogalomról van szó. Ezek a hivatkozások formai jellemzőiket tekintve hasonlíthatnak az „igazi" hivatkozásokhoz, és csak azzal lehet elkülöníteni őket, hogy a típushivatkozások nem tartalmaznak minden szükséges partikuláris adatot, ami egy konkrét, egyedi jogszabály azonosításához szükséges (tipikus módon hiányzik az időre és a jogszabály konkrét jelére, „sorszámára" vonatkozó információ).

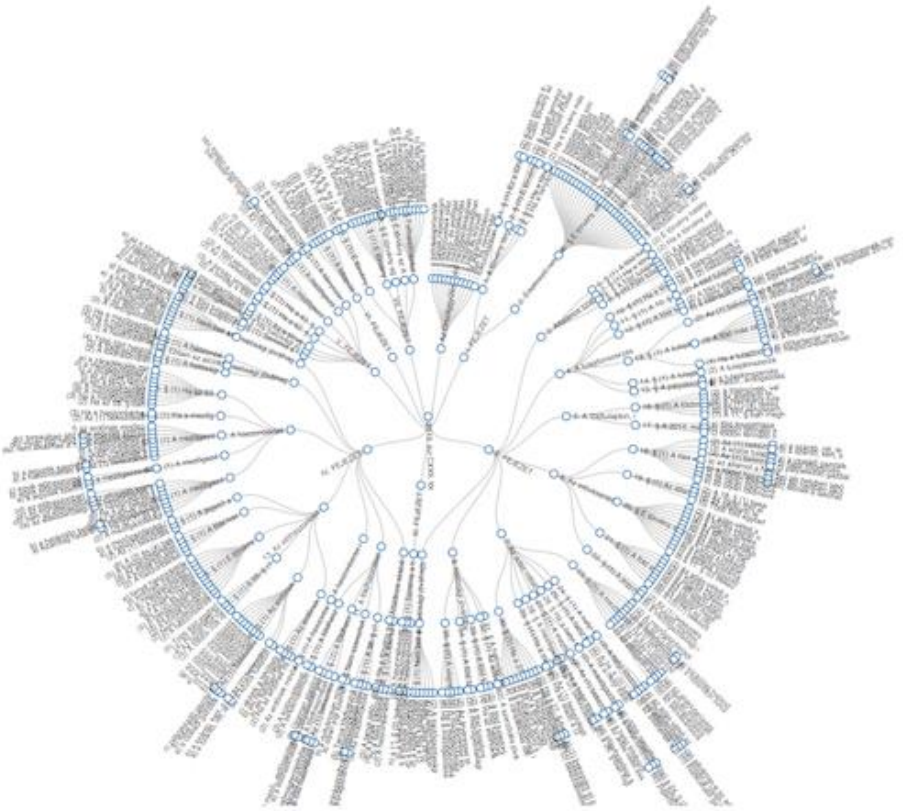
Az előző problémakörhöz hasonlóak azok a jogszabályok, amelyek egyértelműen az elemzett jogszabály megalkotásának időpontját követően létrehozott, létrehozandó jogszabályokra tesznek utalást. Az ilyen hivatkozások ugyanúgy nem teljesek, mint a fentebb bemutatott „általános jogszabályok", érdemes mégis külön említeni őket, és külön foglalkozni velük, mert ezek a *biankó jogszabályok* idővel jó eséllyel megszületnek, bekerülnek a teljes jogkorpuszba, amikortól fogva már ugyanolyan hivatkozható, partikuláris jogszabályként funkcionálnak, mint a többiek. Tipikus példaként említhetjük a törvények végrehajtási jogszabályait. A biankó jogszabályokat éppen azért érdemes külön kezelnünk, mert ha az elemzést időben később végezzük el, azután, hogy a biankójogszabály már hatályosították, akkor a biankójogszabály a többi jogszabállyal teljesen „egyenrangú" elemként szerepel a korpuszban, miközben a jelenben még eltérő a státusuk.

Goody-listák és hivatkozások

A jogszabályszövegek jelentős hányada jelölt listákból áll, amelyek elliptikus mondatokként is jellemezhetőek. Ezeket korábban Goody-listáknak neveztük el, és a sajátosságaikat több szempontból is elemeztük [5, 6, 7, 11]. A hivatkozások feltárásban a Goody-listák akkor jelenthetnek gondot, amikor a hivatkozások „átfolynak" a Goody-listák fejtételei és listatételei között. Mivel a Goody-listák elliptikus mondatainak kiegészítését, teljes mondatokká egészítését korábban már megoldottuk, így itt ezt a technikát alkalmazva a probléma könnyen eliminálható. A hivatkozásfeltárás feladatát azon a feldolgozott szövegen érdemes megkezdeni, amely már tartalmazza a Goody-listák hiányos mondatainak kiegészítését.

Vizuális ábrázolás

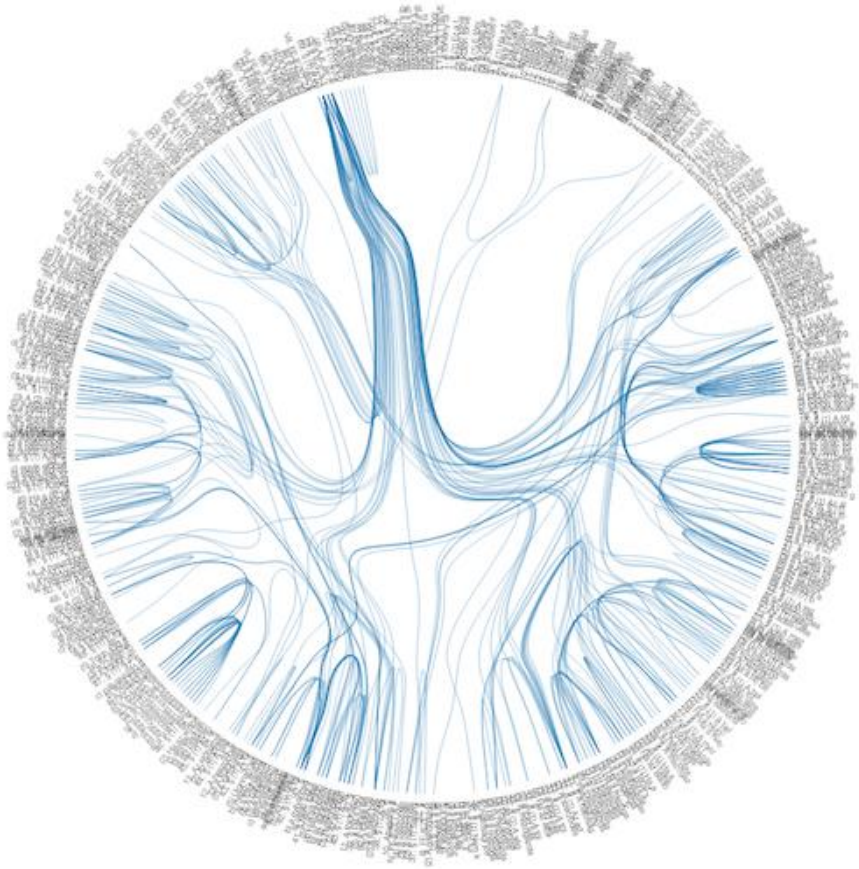
A jogszabályok szövege hierarchikus struktúrájú, ám a hierarchikus elrendezés megszokott szabályaira (pl. a hierarchikus szintek egyenletes eloszlására) a jogszabályok szövegének kialakításakor nem lehetnek figyelemmel. A jogalkotó részletezni/tételezni ott fog, ahol erre tartalmilag (szemantikailag/szintaktikailag) szükség van. Így előfordul, hogy nagyon egyenlőtlen eloszlású hierarchiák jönnek létre, amikor a jogszabályok egyik részében nagyon gazdag alstruktúra jelenik meg, míg máshol kevés szöveg, lapos struktúra a jellemző. A „szép” hierarchikus megjelenítést (és az egyértelmű hivatkozást) nehezíti a – már bemutatott – kettős tipizálás kényszere is (a bekezdések és azok első szakaszai esetében).



1. ábra: a 2013. évi CXXII. törvény szerkezete

A jogszabályok struktúrájának elemzéséhez, a struktúrák tipizálásához kerestünk adatvizualizációs technikákat [13]. A kiválasztott javascript-könyvtárakra [9] támaszkodva elkészítettük a 2013. évi CXXII. törvény [1] belső szerkezetét mutató ábrát (1. ábra). Munkánk következő szakaszában szeretnénk mérhetővé tenni a belső szerkezet jellegzetességeit, hogy ezek alapján a különböző struktúratípusok jól elkülöníthetőek legyenek.

Még érdekesebb az az ábra, amely a jogszabályok belső hivatkozásait mutatja (2. ábra). Első ránézésre is kitűnik, hogy a jogszabályon belül sok egymáshoz közeli helyekre mutató, *szomszédos hivatkozás* van (általában a szakaszokon belül vannak így összekötve a bekezdések, pontok, alpontok egymással). Ez a jellegzetesség vélhetőleg minden jogszabály esetében megfigyelhető lesz. Már nem annyira kiterjedten, de még mindig elég sok jogszabályt jellemezhet az az itt megfigyelhető másik vonás, hogy a jogszabály zárórendelkezései között sok visszamutató, belső linket találhatunk. A magyarázat egyszerű: gyakran előfordul, hogy az egyes részek különböző időpontban lépnek hatályba, mely időpontok belső hivatkozással történő felsorolását találjuk meg a jogszabályszöveg lezárásaként. Számos jogszabály belső hivatkozásainak hasonló módon történő ábrázolása összevethetővé teszi a belső linkek strukturáját, mely – munkahipotézisünk szerint – rámutathat bizonyos (pl. jogágankénti vagy jogszabálytípusonkénti) jellegzetességekre, eltérésekre. A belső linkek szerkezetének teljes tipizálásához, a típusok értelmezéséhez még nem áll elég tudás a rendelkezésünkre. Első menetben arra van szükség, hogy sok jogszabály vizualizációját elemezve a tipizáláshoz szükséges szempontokat megtaláljuk, hogy aztán mérhetővé tegyük a típusalkotáshoz szükséges dimenziókat.



2. ábra: a 2013. évi CXXII. törvény belső linkjei

Köszönetnyilvánítás

A tanulmány létrejöttében segített a K112172 sz. OTKA kutatás támogatása. Köszönjük Király Péternek az adatvizualizációhoz nyújtott segítségét.

Hivatkozások

1. A mező- és erdőgazdasági földek forgalmáról szóló 2013. évi CXXII. törvény
2. A jogszabály szerkesztésről szóló 61/2009. (XII. 14.) IRM rendelet
3. A nemzeti vagyonról szóló 2011. évi CXCVI. törvény
4. Friedl, J.E.F.: Mastering Regular Expressions, O'Reilly Media (2006)
5. Hamp, G., Syi, Markovich, R.: Elliptikus listák jogszabályszövegekben. In: Tanács A., Varga V., Vincze V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2015) 273–281.
6. Markovich R., Hamp, G., Syi.: Jogszabályszövegek gépi elemzésének tanulságai. Jogelméleti Szemle, 2. (2015) 64–73.
7. Markovich, R., Syi, Hamp, G.: Elliptical lists in legislative texts. In: ICAIL '15 Proceedings of the 15th Int. Conf. on Artificial Intelligence and Law (2015) 192–195.
8. Martínez, M., de la Fuente, P., Vicente, D-J.: Dealing with the automatic extraction of references from legislative digital libraries. Veille Strategique, Scientifique et Technologique 2004 (VSST 2004) 281–288
9. mbostock's blocks, http://bl.ocks.org/mbostock
10. Sartor, G.. Legislative Information and the Web. In Sartor, G., Palmirani, M., Francesconi, E., Biasiotti, M.A. (eds.) Legislative XML for the Semantic Web, Springer (2011) 11–20.
11. Syi, Hamp, G., Markovich, R.: Goody-listák a jogszabályszövegekben. Három tételben. JEL-KÉP 3. (2015) 13–24.
12. Palmirani, M., Brighi, R., Massini, M.: Automated extraction of normative references in legal texts. In: ICAIL '03, Proceedings of the 9th Int. Conf. on Artificial Intelligence and Law (2003) 105–106.
13. Tufte, Edward R.: Visual Explanations: Images and Quantities, Evidence and Narrative, Cheshire, CT: Graphics Press (1997)

Digitális Konzílium – egy szemészeti klinikai keresőrendszer

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat A klinikai dokumentumok feldolgozása a nyelvtechnológia egyik kiemelkedő és igen hasznos alkalmazási területe. A klinikai körülmények között létrejövő beteglapok igen sok hasznos információt tartalmaznak a beteg mellett az orvosok számára is. Ezek tárolási módja azonban nem teszi lehetővé ezeknek az információknak az elérését.
Cikkünkben egy magyar nyelvű szemészeti dokumentumokat feldolgozó láncot és a feldolgozott dokumentumokra épülő összetett keresőrendszer első változatát mutatjuk be. Az alkalmazott módszerek a klinikai dokumentumok sajátosságait veszik figyelembe az előfeldolgozás első lépéseitől kezdve a keresőfelület kialakításáig.

1. Bevezetés

A klinikai dokumentumok, beteglapok feldolgozása a nyelvtechnológia egyik fontos alkalmazása, amely egyre jobban elválik az orvosbiológiai szövegekre épülő alkalmazásoktól [4,10,3]. A kórházi körülmények között létrejövő dokumentumok sok hasznos információt tartalmazhatnak más orvosok vagy szakértők számára. Sokszor azonban a dokumentumok formátuma és tárolási módja, illetve a létrehozásukat „támogató” eszközök ellehetetlenítik a tényleges tartalomhoz való hozzáférést, elzárva ezzel az orvosok elől ezt a gazdag információforrást.

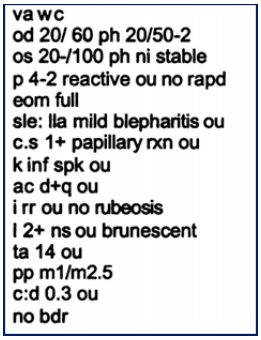
A klinikai dokumentumok létrehozásának módszere két fő csoportba sorolható. Az egyik esetben egy EHR (Electronic Health Records) rendszeren keresztül kerülnek rögzítésre. Ekkor az orvosok vagy az asszisztensek egy, a rendszer által felajánlott sablont töltenek ki, melynek eredménye valamilyen strukturált dokumentum. A dokumentum szerkezetének részletessége függ a dokumentációs rendszertől és a felhasználói szokásoktól is. A klinikai dokumentumok létrehozásának másik módja a hagyományos, kézzel írt dokumentációs szokásokhoz hasonlítható. Ekkor, bár a dokumentumok létrehozása és tárolása is számítógépen történik, ez tulajdonképpen csak mint egy írógép használható. A létrejött dokumentum pedig csupán egy egyszerű szövegfile, amiben a dokumentum tartalmi szerkezetének nyomai csak a szintén kézzel rögzített formázási jegyekben fedezhetők fel.

Cikkünkben egy olyan rendszer prototípusát mutatjuk be, ami magyar nyelvű szemészeti dokumentumok egy gyűjteményét alakítja át kereshető, strukturált

formába, és az így feldolgozott leletanyaghoz keresőfelületet is biztosít. A magyar kórházakban az EHR rendszerek használata hagy némi kívánnivalót maga után. Bár a legtöbb esetben használnak ilyen rendszereket, az ezeket használó orvosok és asszisztensek a rendszer rugalmatlansága és bonyolultsága miatt inkább saját dokumentációs szokásaikat tartják meg, a rögzített információt a felkínált sablon egyetlen mezőjében rögzítve.

Különösen nehéz terület a szemészeti dokumentáció, ahol az EHR rendszerek alkalmazása más országokban sem jár igazán sikerrel [1,9,2]. A sokféle mérési eredmény rögzítésének módja (amik közül néhány táblázatos formát igényel, a többit egyetlen szám vagy rövid szöveg, vagy akár ábra ír le) olyan rendszert igényelne, aminek a létrehozása és használata igen nehéz.

A szemészeti dokumentáció egy másik sajátos tulajdonsága, hogy a feljegyzések sietve, a vizsgálat közben jönnek létre. Ezért igen jellemzőek a gyakran ad hoc rövidítések, az elírások, az angol, a latin és a magyar nyelv vegyes használata, illetve a szöveges leírásokban is gyakran hiányos nyelvtani szerkezetek jönnek létre. Ezért a legtöbb, általános magyar szövegekre jól alkalmazható elemzőrendszer nem használható közvetlenül ezeknek a szövegeknek az elemzésére. Ez a feldolgozási lánc olyan alapvető elemeire is igaz, mint a tokenizálás, mondathatárfelismerés, vagy szófaji egyértelműsítés. Az 1. ábra egy angol, míg a 2. ábra egy magyar nyelvű szemészeti leletet mutat be, amiken jól látszanak a szemészeti dokumentáció során használt nyelvezet sajátosságai. A leletek feldolgozása tehát nem csak a magyar nyelvből fakadó összetettség miatt nehéz, hanem a szemészeti domén jellemzője bármely nyelv esetén.


va wc
od 20/ 60 ph 20/50-2
os 20-/100 ph ni stable
p 4-2 reactive ou no rapd
eom full
sle: lla mild blepharitis ou
c.s 1+ papillary rxn ou
k inf spk ou
ac d+q ou
i rr ou no rubeosis
l 2+ ns ou brunescent
ta 14 ou
pp m1/m2.5
c:d 0.3 ou
no bdr

1. ábra. Egy angol nyelvű szemészeti lelet részlete

A bemutatott rendszer célja, hogy mindezeket a nehézségeket kezelni tudja. A feldolgozás során a nyers szöveges dokumentumokból indul ki, amikben először a helyesírási hibák automatikus javítása történik, majd a rövidítések felismerése és feloldása, és a kórtörténet, illetve a dokumentumstruktúra automatikus meghatározása, melynek során minden dokumentum minden sora egy adott tartalmi

kategóriába sorol be a rendszer. Végül az így feldolgozott és strukturált dokumentumokat indexeljük, és egy speciális keresőfelületen keresztül kereshetőek.

2. A magyar szemészeti korpusz

Vizsgálataink során szemészeti osztályon keletkezett anonimizált magyar nyelvű dokumentumokat használtunk. A dokumentációs rendszer hiánya, vagy annak elnagyolt használata miatt a dokumentumok struktúrájára az esetek nagy részében csak az esetleges (a korpuszban nem is egységesen használt) formázási jegyek, illetve a dokumentumok tartalmának értelmezése utal. Továbbá, a feljegyzések jelentős részének tárolása redundáns módon történt, a teljes vagy részleges kórtörténet többszöri, kézzel történő másolása miatt. A dokumentumok szöveges része nagyon sok helyesírási hibát, rövidítést, idegen nyelvű terminológiát tartalmaz, amiknek a lejegyzése nem követ semmilyen helyesírási normát. A 2. ábrán egy példadokumentum látható az eredeti formájában.

AMBULÁNS KEZELŐLAP

Státusz
2010.10.19 12:28

Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
V:0,7+0,75Dsph=1,0
1,0 +0,5 Dsph élesebb

+2.0 Dsph mko Cs IV

St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla rekciók rendben, lencse tiszta, jó vvf.
Átfecskendezés mko sikerült.

Olvasó szemüveg javasolt: +2.0 Dsph mko.
Éjszakánként műkönnygél ha szükséges.
Kontroll: panasz esetén

Diagnózis

DIAGNÓZISOK megnevezése	Kód	Dátum	Év	K	V	T
Látászavar, k.m.n.	H5390	2010.10.19				3

Beavatkozások

Kód	Megnevezés	Menny.	Pont
11041	vizsgálat	1	750

2010.11.16

2. ábra. Egy szemészeti lelet annak eredeti formájában

A szemészeti dokumentumok egyedi sajátossága a szöveges részek közé illesztett táblázatos, vagy általázatos adat. Ezek a szövegfeldolgozás során zajt képeznek az adott környezetben. Ilyen nem szöveges információk a laboreredmények, a szemvizsgálat során mért látásélességi mérések eredményei, illetve speciális elválasztó karakterek, vagy más speciális karakterek sorozatai. A dokumentumok ezen részeire is jellemző a szabványos forma hiánya, így dokumentumról dokumentumra (vagy orvosról orvosra) változó lehet a leírás módja.

A dokumentumok szöveges részeit tekintve is számos jelentős különbség fedezhető fel az általános magyar nyelvű szövegek és a szemészeti domén között,

ami magyarázatot ad arra, hogy az általános szövegeken akár bonyolultabb feladatok esetén is jól teljesítő eszközök miért nem alkalmazhatóak a klinikai dokumentumokra. A különbség a két szövegtípus között nem csak azok tartalmában nyilvánul meg, hanem már a nyelvtani szerkezetekben és a szövegekben előforduló szóalakokban is. A két domén részletes összehasonlítása megtalálható [12]-ben és [14]-ben.

3. Előfeldolgozási lépések

A fent ismertetett sajátos, igen zajos és a bármilyen helyesírási és szerkesztési normához való igazodást nélkülöző nyelvezettel lejegyzett klinikai dokumentumok automatikus feldolgozása igen nehéz, ezért az előfeldolgozási lánc létrehozása és adaptálása minden további lépés előfeltétele és komoly munkát igényel. Az összes előfeldolgozási lépés részletes ismertetése meghaladja a jelen cikk korlátait, ezért csak röviden ismertetjük őket.

A **mondatok szegmentálása és a tokenizálás** az első alapvető lépés a szöveges részek feldolgozása során. Ennek megvalósítása két lépésben történik, először a tokenizálás, aztán pedig a mondatvéget jelző indikátorok log-likelihood alapú osztályozásával a mondathatárok felismerése. A módszer eredményessége 96,0%-os F-mértékkel jellemezhető a tokenizálásra és 91,89%-os F-mértékkel a mondathatár-felismerésre. [6]

A **helyesírási hibák automatikus javítása** során célunk egy kvázisztenderd forma elérése volt a szöveges részekben. Ehhez egy statisztikai gépi fordítón alapuló módszert alkalmaztunk, de a rendszer párhuzamos korpuszból való tanítása helyett a fordítási modellt egy javaslatgeneráló rendszerrel helyettesítettük, ami minden szóalakhoz lehetséges javításjelölteket generált. A dekóder feladata ezek közül a megfelelő alak kiválasztása, a nyelvmodell pedig a szövegkörnyezetet képviselve súlyozza a lehetséges alakok valószínűségét. A módszer 87,23%-os pontossággal működik.[14]

Már a tokenizálás és szegmentálás feladatát is jelentősen megnehezítette a rengeteg rövidítés, illetve azok igen sokféle megjelenési formája. Egy kezelhető reprezentáció elérése érdekében pedig különösen fontos lépés a jelentéssel bíró egységek azonosítása, így a **rövidítések felismerése és feloldása**. [15] A szövegek normalizálása során történő azonosítás és feloldás mellett a felderített rövidítéseket és feloldásaikat később a keresőrendszerbe is integráltuk, szinonimahalmazokként megadva őket, ezzel lehetővé téve a rövidített és kifejtett előfordulások megfeleltetését a keresés során.

Az előfeldolgozási lánc utolsó eleme a dokumentum szöveges részeinek **szófaji egyértelműsítése**, melyre a PurePos adaptált verzióját használtuk [7]. Az egyértelműsítő teljesítménye ugyan elmarad az általános szövegeken mérhetőtől, de a klinikai szövegeken is 90% fölötti pontosságot ért el.

A fenti lépések alkalmazása után létrejött egy kvázi-normalizált, illetve lemmatizált és morfoszintaktikai címkékkel ellátott reprezentáció a dokumentumok szöveges részeire. Ezután a rendszerhez **szemantikai információkat** adtunk

hozzá. Mivel magyar nyelvre nem nagyon létezik használható a szemészeti domént megfelelően reprezentáló lexikai szemantikai erőforrás, ezért egy olyan módszert alkalmaztunk, amivel a korpuszból létrehozott disztribúciós szemantikai modell alapján definiáltunk egy szemantikailag releváns fogalmi rendszert [11].

4. Dokumentumszerkezet azonosítása

A dokumentum szerkezetének azonosítását két lépésben valósítottuk meg. Először egy a formázási jegyeket azonosító algoritmust alkalmaztunk, majd a dokumentumban szereplő minden sort egy-egy tartalmi egységet jelölő címkével azonosított osztályba soroltunk be [13]. Az 1. táblázat az osztályozáshoz definiált tartalmi egységeket foglalja össze.

1. táblázat. A dokumentum állításainak kategorizálásához használt kategóriacímkék és jelentésük

címke	**jelentés**	**leírás**
`Tens`	Szemnyomás	A szemnyomás mérésének eredménye
`V/Refr`	Refrakció	Refrakció adatok
`Ana`	Anamnézis	A beteg panaszai, korábbi betegségek, családi kórtörténet, stb.
`Dg`	Diagnózis	Az aktuális diagnózis
`Beav`	Beavatkozás	Alkalmazott kezelés, kivéve ha műtét vagy gyógyszeres terápia
`Vél`	Vélemény	Az orvos által megfogalmazott vélemény, kivéve ha diagnózis vagy kezelés
`St`	Státusz	A beteg aktuális állapota
`Ther`	Terápia	Felírt vagy alkalmazott gyógyszeres kezelés
`BNO`	BNO	Betegségekhez vagy kezelésekhez tartozó BNO kódok
`T`	Teszt	Alkalmazott vizsgálatok, kivéve a `Rl` kategóriába tartozók
`V`	Visus	Látásélesség vizsgálatának eredményei
`Rl`	Réslámpa	Réslámpával végzett vizsgálatok
`Kontr`	Kontroll	Kontrollvizsgálatra vonatkozó információk
`Műtét`	Műtét	Előírt vagy végrehajtott műtétek
`XXX`	-	A fenti kategóriák egyikébe sem sorolható állítások

4.1. Kórtörténet azonosítása

Igen gyakran előfordul, hogy az aktuális beteglap egyes részei a betegre vonatkozó korábbi dokumentumokból származó részleges vagy teljes másolatok, esetenként kisebb változtatásokkal. Így bár ezek a részletek csupán redundáns információt tartalmaznak (a teljes betegtörténetre vonatkozóan), felismerésükre nem volt elegendő a teljes egyezés vizsgálata. A másik nehézséget ezek azonosítása során az jelentette, hogy sok olyan feljegyzésrészlet is van, ami közel minden

vizsgálatnak része, és gyakran tartalmilag is hasonlóak, így egy egyszerű illeszkedésvizsgálat során tévesen azonosíthatnánk másolatként ezeket. A tényleges másolatok és változataik felismerésére ezért egy olyan módszert alkalmaztunk, melynek alapja a dokumentumok adott részeinek `md5` kódolt változatának összehasonlítása [13]. A módszer alkalmazása után az egyes betegekhez tartozó kórtörténet visszakövethetővé vált, az eredetileg egyetlen dokumentumban összefűzve tárolt információ időrendi sorrendben kinyerhető és egységekre bontható.

4.2. Feljegyzett sorok osztályozása

A formázási jegyek alapján azonosított nagyobb szerkezeti egységeken belül is többféle tartalmi egységbe sorolható információ került feljegyzésre a dokumentumokban. Ezek azonosítását két lépésben oldottuk meg.

Először a szövegek előfeldolgozott változata alapján definiáltunk a szófajcímkék és a leggyakoribb kifejezésekhez rendelt szemantikai kategóriacímkék alapján olyan mintázatokat, amik az adott állítást valamilyen tartalmi egységbe sorolták. Például az igék alapvetően ritka használata miatt, ha egy múlt idejű igealak környezetében bizonyos szemantikai csoportba tartozó szavak kerültek, akkor az adott mondat nagy valószínűséggel az anamnézis kategóriába sorolható.

A második lépésben a többi sort osztályoztuk. Ehhez először az egyes tartalmi egységekhez indikátorszavakat gyűjtöttünk a korpuszból (néhány példa látható a 2. táblázatban). Az adott kategóriához tartozó indikátorszót tartalmazó sorokat megcímkéztük a kategóriacímkével, majd az így azonosított sorokból minden kategóriához egy szózsák modellt hoztunk létre tf-idf súlyozással. A címkézetlenül maradt sorokat pedig ezekhez a modellekhez hasonlítva kategorizáltuk [13].

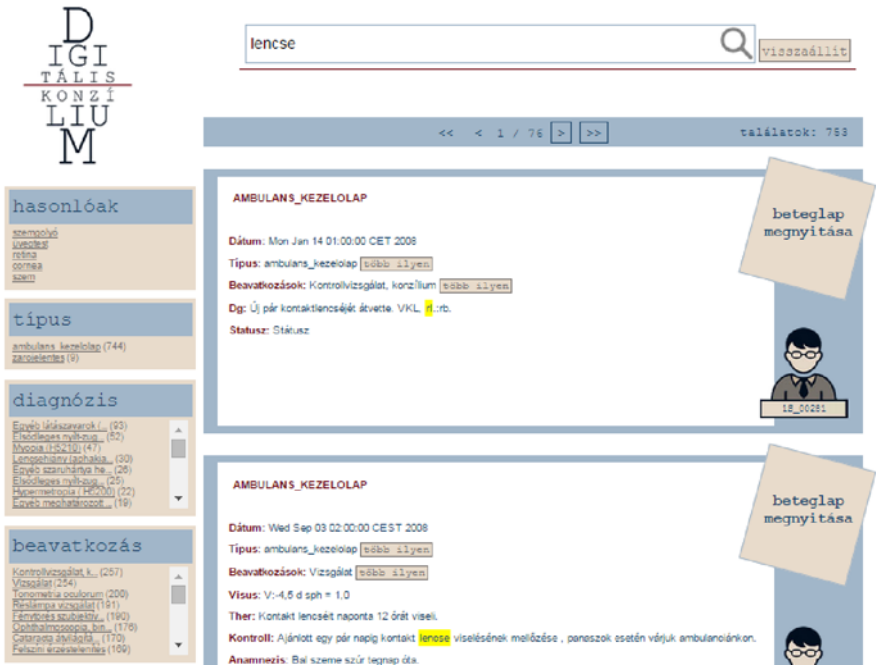
2. táblázat. Néhány példa a címkékre és a hozzájuk tartozó indikátorkifejezésekre

`Ana`	`T`	`RL`	`Ther`
egyéb betegség	eredmény	réslámpa	th
család	ultrahang	macula	szemcsepp
korábbi	Topo	fundus	terápia
hypertonia	Schirmer	rl	rendelés
anamnézis		lencse	javasolt
panasz			

5. A keresőrendszer

Célunk egy olyan keresőrendszer megvalósítása volt, amiben a dokumentumok strukturált és feldolgozott állapotban kerülnek eltárolásra és válnak kereshetővé oly módon, hogy akár tartalmi egységre, akár más szempont alapján való megszorítások, illetve összetett lekérdezések is megfogalmazhatóak legyenek. Ezért

a tartalmi címkékkel megjelölt dokumentumokat egy olyan XML-sémának megfelelően alakítottuk át, ami az egyes tartalmi egységeket elkülöníti, de a dokumentumok eredeti formáját is megőrzi. Az összes dokumentum így átalakított formáját egy `Solr` alapú keresőrendszerben indexeltük és kereshetővé tettük. A tárolt szerkezethez, illetve a dokumentumok sajátosságaihoz és a keresőrendszer eredeti célkitűzéseihez illeszkedő keresőfelület a 3. ábrán látható.



3. ábra. A keresőrendszerhez készült keresőfelület prototípusa

5.1. Csoportosítás

Az összetett, ugyanakkor rugalmas és könnyen kezelhető keresőfelület megvalósítása érdekében a rendszer különböző szempontok szerinti dinamikus csoportosítást valósít meg az aktuális keresés során kapott eredménylista alapján. Így a találatok halmaza tovább szűkíthető az azokban található diagnózisok, vizsgálatok és alkalmazott kezelések mentén. Továbbá, szintén a keresés lefuttatása után a felületen állítható a találati eredmények időbeli korlátozása is.

A találati lista szűrése mellett pedig kapcsolódó kifejezéseket is hozzáadhatunk az eredeti lekérdezéshez. Ezekre a rendszer tesz ajánlatot, miután a felhasználó beírta az általa keresett kifejezést. Az ajánlatok listája a szemészeti korpuszból korábban létrehozott fogalmi rendszer alapján jön létre, melyek akár

többszavas kifejezéseket vagy rövidített alakokat is tartalmazhat. A kategorizáció a fogalmak disztribúciós modellje alapján hierarchikus klaszterezéssel jött létre, melynek során a teljes hierarchiából koherens, tömör csoportok kivágásával jöttek létre az egymáshoz kapcsolódó kifejezések halmazai [16]. Így tehát a keresőkifejezés beírása után az azzal egy csoportban szereplő kifejezéseket ajánlja fel a rendszer, melyek nem csupán szinonimák, hanem lazábban kapcsolódó kifejezések is lehetnek. Például a *lencse, szemgolyó, üvegtesti tér, retina, kornea, szem* egy csoportba került kifejezések, a keresés során pedig jól finomíthatják vagy kiegészíthetik egymást, az automatikus felajánlásuk pedig megkönnyíti a felhasználó dolgát.

5.2. Szinonimák és rövidítések

A szinonimák azonosítása mind az indexelés, mind a lekérdezés során megtörténik. A rendszerben tárolt szinonimák meghatározása szintén a korpusz alapján történt. Szinonimaként tároljuk továbbá a rövidítések különböző változatait és azok feloldásait is, melyeket az előfeldolgozás során határoztunk meg. Például a *mindkét szem* kifejezés lekérdezése során a találati listában ennek a kifejezésnek az összes változatára illeszkedő dokumentumokat találunk. Így a magyar rövidített alaknak megfelelő *mksz* (és változatai), vagy a latin alaknak megfelelő rövidített változatok *ou, o.u., o.utr. OU* (oculi utriusque 'mindkét szem') is illeszkedő találatok lesznek.

A rövidítések szinonimaként való kezelése azonban zajt is eredményezhet a rendszerben, a nagyon sokféle és nagyon gyakori rövidítések miatt. Ezért nem egyetlen általános szinonimalistát tárolunk, hanem a tartalmi egységeknek megfelelően a rövidítéseket is osztályoztuk. Így ha egy rövidítés az egyik tartalmi egységben gyakori, míg a többiben nem használatos, akkor azt csak az ahhoz az egységhez tárolt és használt szinonimalistában tároljuk.

5.3. Helyesírási változatok

Mivel a rendszert magyar nyelvű dokumentumok keresésére alkalmazzuk, a helyesírási változatok kezelése nem egyszerű, de igen fontos feladat. Bár a dokumentumok előfeldolgozása során alkalmaztunk egy modult a helyesírási hibák javítására, a lekérdezés ellenőrzése csak online történhet. Ennek megoldására a Humor [8,5] morfológiai elemző orvosi terminológiával kiegészített változatát [7] integráltuk a rendszerbe.

5.4. Kórtörténet és beteglapok megjelenítése

A találati lista megjelenítésekor a dokumentumok feldolgozott és strukturált változata jelenik meg, amiben minden sor valamilyen tartalmi egységbe került besorolásra. A keresőkifejezésre való illeszkedést pedig kiemeléssel jelöljük. A felhasználónak azonban lehetősége van az eredeti dokumentum megjelenítésére is, annak eredeti alakjában. Mivel a rendszer megcélzott felhasználói köre az orvosok, ezért különösen fontos, hogy a számukra ismert és használt megjelenítést

is biztosítsunk. Amikor a felületen erre a megjelenítési módra váltunk, akkor az aktuális dokumentum egy léptethető időskálán jelenik meg, melyen így a kórtörténet átláthatóvá és visszakövethetővé válik közvetlenül a felületről. Az időskála bármely pontjáról visszatérhetünk az eredeti találati listához.

5.5. A *visus* mező

Mivel a keresőrendszer célzottan a szemészeti kezelőlapokhoz készült, ezek egyik jellemző sajátossága a beteg szemvizsgálata során kapott optikai jellemzők táblázatszerű feljegyzése (visus). Ennek során a beteg látását javító dioptria-, látásélesség-, szemnyomás- stb. értékeket az orvosok saját szokásaiknak megfelelő, általában semmilyen szabványhoz nem igazodó formában írják le. Sok esetben azonban ezek azok a részek az egyes dokumentumokban, amik a legfontosabb információkat tartalmazzák, a diagnózis is sok esetben nagyrészt ezekből határozható meg. Mivel ezek az adatok különböző, általában diszkrét skálákhoz igazodnak, ezért a keresés során a konkrét értékek mellett azok tartománybeli elhelyezkedése is fontos lehet. Mintaillesztési módszerek alkalmazásával megvalósítottunk egy olyan funkciót, amelyben akár szövegesen, akár konkrét értékekkel megadott keresés illeszthető a dokumentumokban különböző formátumban tárolt numerikus visusértékekre. Például a *közepes fokú rövidlátás* keresőkifejezés találatai között megtaláljuk azokat a dokumentumokat, ahol a betegnél mért szferikus dioptriaérték negatív előjelű, és a megfelelő tartományba esik. Ezzel egy időben a keresőrendszer többi funkciói is aktívak, így pl. az előbbi kifejezés valamely rövidített latin megfelelőjét (pl. *myop.med.grad.*) tartalmazó dokumentumok is találatok maradnak.

6. Konklúzió

Cikkünkben egy olyan keresőrendszer első verzióját mutattuk be, amely szemészeti dokumentumokhoz biztosít intelligens keresési felületet. A szemészeti kórlapok a klinikai dokumentumokon belül is speciálisan nehezen kezelhető részhalmazt jelentenek, ezért több előfeldolgozási lépés kifejlesztésére és alkalmazására volt szükség. Az előfeldolgozott dokumentumokat és az azokhoz tartozó hozzáadott információkat (rövidítések, kapcsolódó kifejezések, írásváltozatok stb.) egy keresőrendszer számára indexelhetővé tettük. Bemutattuk továbbá azt a keresőfelületet, ahol a dokumentumokban való keresésre összetett lekérdezéseket adhatunk meg, illetve a találati listát szűkíthetjük, bővíthetjük, a dokumentumok pedig több nézetben is megjeleníthetők.

Jelen cikkünkben a keresőrendszer első változatát mutattuk be, amely még természetesen további fejlesztésre szorul, de már ebben az állapotában is hiánypótló eszköz lehet a szemészettel foglalkozó orvosok számára, amit gyakorló szemészorvosok is megerősítettek.

Hivatkozások

1. Chiang, M.F., Read-Brown, S., Tu, D.C., Choi, D., Sanders, D.S., Hwang, T.S., Bailey, S., Karr, D.J., Cottle, E., Morrison, J.C., Wilson, D.J., Yackel, T.R.: Evaluation of electronic health record implementation in ophthalmology at an academic medical center (an american ophthalmological society thesis). Trans Am Ophthalmol Soc 111, 70–92 (Sep 2013)
2. Elliott, A., Davidson, A., Lum, F., Chiang, M., Saaddine, J.B., Zhang, X., Crews, J.E., Chou, C.F.: Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions. Am J Ophthalmol 154(6 0), S63–S70 (Dec 2012), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4037529/
3. Friedman, C., Johnson, S., Forman, B., Starren, J.: Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc Annu Symp Comput Appl Med Care pp. 347–51 (1995)
4. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 35, 128–44 (2008)
5. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
6. Orosz, Gy., Novák, A., Prószéky, G.: Hybrid text segmentation for Hungarian clinical records, Lecture Notes in Artificial Intelligence, vol. 8265. Springer-Verlag, Heidelberg (2013)
7. Orosz, Gy., Novák, A., Prószéky, G.: Lessons learned from tagging clinical Hungarian. International Journal of Computational Linguistics and Applications 5(1), 159–176 (2014), http://www.gelbukh.com/ijcla/2014-1/
8. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, College Park, Maryland (1999)
9. Redd, T.K., Read-Brown, S., Choi, D., Yackel, T.R., Tu, D.C., Chiang, M.F.: Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. Journal of AAPOS 18(6), 584–589 (2014)
10. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association 1(2) (Mar/Apr 1994)
11. Siklósi, B., Novák, A.: Identifying and clustering relevant terms in clinical records using unsupervised methods. In: Besacier, L., Dediu, A.H., Martín-Vide, C. (eds.) Statistical Language and Speech Processing, pp. 233–243. Lecture Notes in Computer Science, Springer International Publishing (2014)
12. Siklósi, B., Novák, A.: A magyar beteg. X. Magyar Számítógépes Nyelvészeti Konferencia pp. 188–198 (2014)
13. Siklósi, B., Novák, A.: Restoring the intended structure of hungarian ophthalmology documents. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics (2015)
14. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. Computer Speech and Language in press(0), – (2014), http://www.sciencedirect.com/science/article/pii/S0885230814000795

15. Siklósi, B., Novák, A., Prószéky, G.: Resolving abbreviations in clinical texts without pre-existing structured resources. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. ELRA (2014)
16. Siklósi, B.: Clustering relevant terms and identifying types of statements in clinical records. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, vol. 9042, pp. 619–630. Springer International Publishing (2015)

VI. Szintaxis

Egyszer „van”, hol nem „van”: A létige kezelése függőségi nyelvtanokban

Simkó Katalin Ilona[1], Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
kata.simko@gmail.com
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Cikkünkben három függőségi nyelvtani elemzést hasonlítunk össze a *van* létige kezelésének szempontjából. Az elméletek előnyeinek és hátrányainak tárgyalása után bemutatjuk, milyen eredményeket ér el egy szintaktikai elemző az egyes elemzésekben. Az ULA és LAS eredmények mellett részletes kézi hibaelemzést is végeztünk az adott szerkezet hibatípusaira koncentrálva. A cikk célja megtalálni a magyar *van* létige különböző típusainak számítógépes elemzésére leginkább alkalmas elméletet, valamint hangsúlyozni a feladatnak leginkább megfelelő elméleti keret megtalálásának és a kézi hibaelemzésnek a fontosságát.

Kulcsszavak: szintaxis, létige, kopula, hibaelemzés

1. Bevezetés

A nyelvi jelenségek nagy részére nem létezik egyetlen elfogadott nyelvészeti leírás, nemcsak az egyes keretek biztosítanak különböző megoldásokat a problémára, hanem az azokon belüli elméletek között is komoly eltérések lehetnek. Jelen cikk célja egy ilyen, a szakirodalomban már sokféleképpen leírt jelenség, a *van* ige lehetséges szintaktikai kezeléseinek megvizsgálása a számítógépes nyelvészetben, dependencia-nyelvtani keretben. A *van* ige kezelésére nem csak az elméleti nyelvészet ad számos lehetőséget, a számítógépes nyelvészet is több kísérletet tett erre [1]. Az ezek közötti választás nem egyértelmű, mindegyik megközelítésnek megvan az előnye és a hátránya is.

Cikkünk célja ezeknek az elemzéseknek a több szempontú összehasonlítása. A különböző elméletek szerint annotált korpuszon tanított elemzők eredményeinek szokásos, ULA és LAS százalékokban kifejezett összehasonlítása mellett kézi hibaelemzést végeztünk a relevánsnak tartott mondatrészek figyelembevételével.

2. A magyar létige függőségi nyelvtanokban

Cikkünkben a magyar *van* ige dependencia (függőségi) nyelvtanbeli lehetséges kezeléseit vizsgáljuk. Ennek oka, hogy ebben a keretben rendelkezésre áll három különböző elmélet szerinti teljes annotációval is ugyanaz a korpusz, a Szeged Korpusz Népszava alkorpusza [2]. A három elmélet így azonos feltételek mellett volt összehasonlítható.

2.1. Létige és kopula

A magyar *van* létigének – és a létigének számos egyéb nyelvben – egzisztenciális és kopuláris használata is létezik. Egzisztenciális használatban teljes értékű igeként viselkedik, létezést fejez ki ((1) példa). Kopuláris használat esetén az ige nem önállóan alkotja a mondat predikátumát, egy névszói rész is társul hozzá ((2) példa). A magyar nyelv – nem egyedi, de – érdekes és problémás tulajdonsága, hogy kopuláris használatban az igei paradigma bizonyos helyein (harmadik személy, jelenidő, kijelentő módban) a felszíni szerkezetben nem jelenik meg az ige ((3) példa).

(1) Sanyi a szobában van.
(2) Sanyi orvos volt.
(3) Sanyi orvos.

2.2. Funkció fej

A funkció fej elemzés a mondatban a funkciószavakat tekinti fejnek. Ilyen módon minden mondat feje a ragozott ige, legyen az akár teljes értékű ige, akár kopula. Ezt az elemzést támogatja Mel'čuk [3], aki a magyarhoz hasonló nyelvek esetén (ha a kopula csak bizonyos szám, személy, mód, idő esetén nem jelenik meg a felszíni szerkezetben) azt javasolja, hogy hiányzó ige esetén egy üres igealakot szúrjunk be a szerkezetbe.

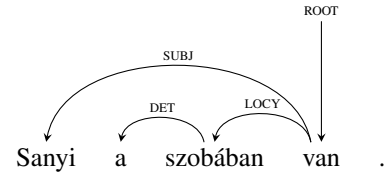
Ennek megfelelően lett létrehozva a Szeged Dependencia Treebank [4] szintaktikai annotációja. Jelen cikkben a treebank Népszava cikkekből álló részével dolgoztunk.

A treebankben minden mondat feje egy ige: ez lehet teljes értékű, jelentéses ige, jelentéssel nem bíró kopula vagy egy, a korpuszba kézzel beszúrt, üres 'VAN' fej a (3) példához hasonló mondatokban.

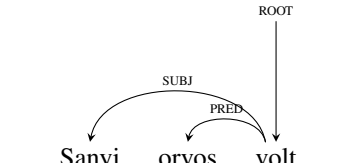
Az elemzés előnye az elemző számára, hogy az összes ige hasonlóképpen viselkedik a mondatokban és minden mondatban van ige. A módszer nagy hátránya viszont az előfeldolgozásban kézzel beszúrt VAN csomópont, ami életszerűtlenné teszi az automatikus elemzést.

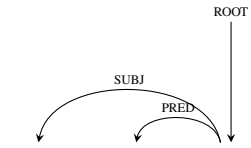
A (4) - (6) példák a (1) - (3) példamondatok elemzései ebben az elméletben.

(4)



(5)



(6)


2.3. Tartalmas fej

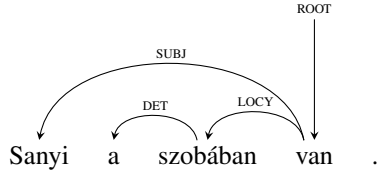
A tartalmas fej megközelítés alapja, hogy a mondatok fő elemeinek a jelentéses egységeket tekinti, a funkciószavak ezekhez kapcsolódnak. Ebben az elemzésben a *van* ige csak tartalmas igeként lesz fej, vagyis a (1) példához hasonló esetekben. A kopula *van* igét tartalmazó mondatok feje a névszói predikátum, az ige ehhez kapcsolódik.

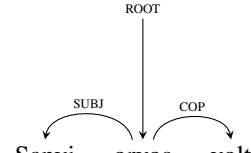
Az univerzális dependencia projektben [5] a különböző nyelvek könnyebb összehasonlíthatóságának érdekében tartalmas fej elemzést alkalmaznak. Mivel a funkciószavak rendszerei erősen különböznek az egyes nyelvekben, a funkció fej elemzésben sokkal nagyobb eltérések mutatkoznának egy-egy mondat különböző nyelvekre fordított változatának szintaktikai leírásában. Jelenlegi munkánkban a Szeged Dependencia Treebank Népszava alkorpuszának univerzális dependencia elveknek megfelelően átalakított változatát használtuk [6].

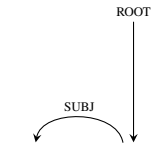
A treebankben a mondatok feje a ragozott, tartalmas ige (köztük a nem kopula *van*). Névszói predikátumot (is) tartalmazó mondatokban a fej mindig a névszó, a kopuláris létige (ha megjelenik) ehhez kapcsolódik, így nem okoz problémát a szerkezetben, ha az ige nem jelenik meg a felszínen.

Az elemzés előnye, hogy nincs szükség kézzel beszúrt igei csomópontokra, valamint elkülöníthető egymástól a *van* létige kopuláris és nem kopuláris használata. Hátránya, hogy kopuláris igéből jóval kevesebb tanítópélda található egy-egy korpuszban, mint jelentéses igéből, így nehézséget okozhat az elemzőnek megtanulni ezt a mintázatot, főleg olyan kis méretű korpusz esetén, mint amilyennel a jelenlegi kutatásban dolgoztunk.

A (1) - (3) példamondatok elemzései tartalmas fej elméletben a (7) - (9) példákban láthatóak. (Az egységes megjelenés miatt az univerzális dependenciában használt címkék helyett a cikkben a Szeged Dependencia Treebank címkéit használjuk.)

(7)


(8)


(9)


2.4. Komplex címke

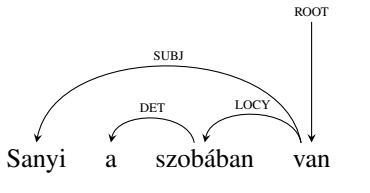
A harmadik megvizsgált elmélet a komplex címkés elemzés. Az elmélet létrehozásában a cél az üres csomópontok eltüntetése volt olyan módon, hogy az elemzésből látszódjon, honnan hiányzik az ige. A komplex címke elemzés alapjában véve egy funkció fej elemzés, attól csak a felszíni szerkezetben meg nem jelenő kopulát tartalmazó mondatok elemzésében tér el.

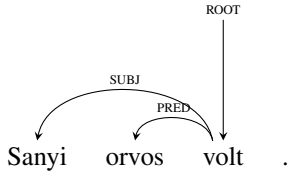
A meg nem jelenő, harmadik személy, jelenidő, kijelentő módú kopula helyett a névszói predikátum lesz a mondat feje, a hiányzó ige ennek a címkéjén és a belőle kiinduló címkéken van jelölve. A címkén szerepel ennek a hiányzó igének a címkéje, az ige és az a címke, amivel az adott szó az igéhez kapcsolódna.

A komplex címkés elemzéshez szintén a Szeged Dependencia Treebank Népszava részének egy átalakított változatát használtuk [7]. Ebből a treebankből automatikusan lettek eltávolítva a beszúrt VAN csomópontok, a hozzá kapcsolódó relációk pedig automatikusan átalakítva. A treebankben csak a meg nem jelenő kopulák elemzése tér el az eredeti, funkció fej változattól.

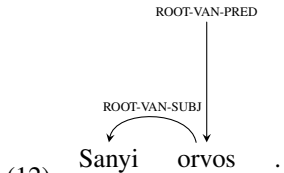
Ennek az elemzésnek szintén előnye, hogy nincs szükség VAN csomópontok beszúrására. Hátránya, hogy csak a megjelenő és a meg nem jelenő igéket különbözteti meg egymástól az elemzésben, valamint hogy potenciálisan végtelen sok címkére szükség lehet például meg nem jelenő kopulát tartalmazó tagmondatok koordinációja esetén.

Komplex címkéket tartalmazó elemzések a (10) - (12) példákban láthatóak a (1) - (3) példamondatokra, valamint a (13) példában két meg nem jelenő kopulát tartalmazó tagmondat koordinációja.

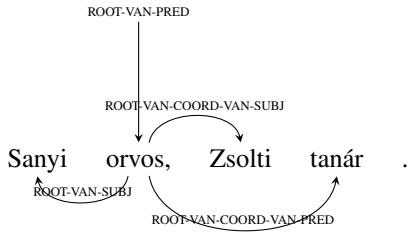
(10)


(11)


(12)



(13)



3. Hibaelemzés

A három, azonos szövegeken meglévő treebanken a Bohnet parsert [8] tanítottuk etalon morfológiai címkék használata mellett, majd a hagyományos, címkézetlen ULA és címkézett LAS értékeket számoltuk rajtuk. Ezek az eredmények a 1. táblázatban a teljes treebanken kiértékelve, illetve az *Egzisztenciális* sorokban csak a teljes értékű ige, jelentéses *van* igét tartalmazó mondatokra, a *„Megjelenő" kopula* sorokban a mondat felszíni szerkezetében megjelenő kopulát tartalmazó mondatokra, a *Virtuális kopula* sorokban a meg nem jelenő kopulát tartalmazó mondatokra számolt értékek vannak feltüntetve.

1. táblázat. ULA és LAS értékek a Funkció fej, Tartalmas fej és Komplex címkés elemzés teljesítményére.

	Funkció fej	Tartalmas fej	Komplex
Egzisztenciális - ULA	86,18	80,48	86,84
LAS	91,04	77,21	82,46
„Megjelenő" kopula - ULA	82,8	75,05	83,62
LAS	77,31	71,67	77,82
Virtuális kopula - ULA	84,42	78,39	77,5
LAS	79,17	75,15	69,59
Teljes anyag - ULA	85,75	84,41	84,76
LAS	81,24	81,2	79,89

Az eredmények alapján azt állapíthatjuk meg, hogy legjobban a funkció fej elemzés teljesít, azt szorosan követi a komplex címkés elemzés, legrosszabb pedig a tartalmas fej elemzés. Azonban ezek az eredmények nem tükrözik megfelelően a vizsgálni kívánt problémát: a teljes mondatokra számolt ULA és LAS számok nem fejezik ki megfelelően a *van* létigével kapcsolatos szintaktikai elemzési problémákat, mivel a mondat egyéb

részeiben ejtett elemzési hibák ugyanúgy befolyásolják az eredményt, mint a jelenleg vizsgálni kívánt, problémás szerkezet hibái.

Ebből kifolyólag kézi hibaelemzést végeztünk a *van*-nal kapcsolatos problémákra koncentrálva. A kézi hibaelemzés során 50-50 mondatot néztünk háromszor három kategóriában: Egzisztenciális, „Megjelenő" kopulás és Virtuális kopulás mondatokon mindhárom elemzésben. A hibaelemzés során négyféle hibakategóriát vettünk figyelembe: nem megfelelő az alany címkéje és/vagy kötése vagy egy egyéb szó kapott alanyi címkét; nem megfelelő a névszói predikátum címkéje és/vagy kötése vagy egyéb szó kapott predikátum címkét; az alany és a névszói predikátum alanyesetű NP-k egymás címkéit kapják meg az elemzésben; nem megfelelő szó lesz a *van* igét tartalmazó CP feje. Azokat a mondatokat, amelyekben ezek közül a hibák közül egyik sem jelent meg, helyesnek tekintettük a jelenlegi kutatás tekintetében, attól függetlenül, hogy volt-e egyéb hiba a mondat elemzésében. A helyes mondatok százalékos aránya a 2. táblázatban láthatóak, az alsó sorban a három kategóriára vonatkoztatott átlagos eredmény található.

2. táblázat. Helyes mondatok százaléka a kézi hibaelemzés alapján a Funkció fej, Tartalmas fej és Komplex címkés elemzésekben.

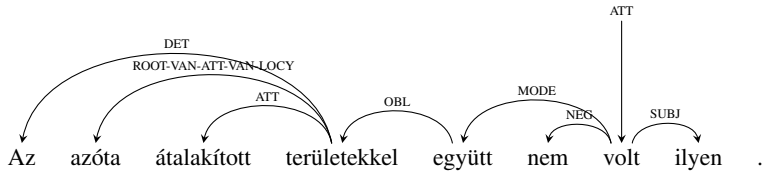
	Funkció fej	Tartalmas fej	Komplex
Egzisztenciális	78	80	80
„Megjelenő" kopula	62	42	52
Virtuális kopula	70	68	30
Összesen	70	63	54

Összességében megállapítható, hogy mindhárom elemző alanyi hibákat ejt legnagyobb számban. Ezek az alanyesetű főnévi csoportok a SUBJ helyett gyakran kapják meg más főnévi módosítók címkéit, illetve a névszói predikátum PRED címkéjét. Az alany és a névszói predikátum megkülönböztetése nehéz feladat: mindkét kifejezés alanyesetű főnévi csoport, és míg első vagy második személyű alany esetén az igével való egyeztetés miatt könnyebben elkülöníthetőek, harmadik személyű NP-nél (főként ha például mindkettő határozott névelős kifejezés) még az anyanyelvi beszélő ember számára sem egyértelmű a helyzet, mint a (14), (15) példák esetén.

(14) A kedvenc rajzfilmem a Mulán.
(15) A Mulán a kedvenc rajzfilmem.

A kézi hibaelemzésből kiderült, hogy a *van* ige elemzésével kapcsolatos hibatípusok a komplex címkés elemzés teljesít legrosszabbul: az Egzisztenciális mondattípus kivételével minden kategóriában ez az elemzés teljesít legrosszabban, valamint az alanyi, a predikátumi címkéket is ez elemzi legtöbbször hibásan. A (13) példában bemutatotthoz hasonló, összetett komplex címkéket (a belső logikájuk megértésének hiányában és az egyes ilyen típusú címkékre vonatkozó kevés tanítópélda miatt) nem tudja helyesen elemezni. Emellett a komplex címkés elemzésben előfordul, hogy a komplex címkék hibákat okoznak olyan helyeken, ahol nem kellene megjelenniük, mint a (1) ábrán, ahol a

TFROM (időhatározói) címke helyett egy komplex címke jelenik meg. Az is a komplex címkés elemzés ellen szól, hogy az elemző tanításának futási ideje több, mint duplája a másik két lehetőségnek, mivel míg a funkció fej 26, a tartalmas fej elemzés pedig 50 különböző lehetséges címkét tartalmaz, addig a komplex címkés elemzés a felhasznált treebankben 200-at, potenciálisan pedig végtelen sok címke tartozhat hozzá.



1. ábra. A komplex címkés szó a kézi annotációban TFROM címkével az *átalakított* szóhoz kötve.

A funkció fej és tartalmas fej elemzések a „Megjelenő” kopula kategóriában értek el a legalacsonyabb értékeket, valószínűleg azért, mert ebben a kategóriában a legnehezebb megkülönböztetni egymástól a teljes értékű ige *van*-t a kopulától. A két elemzés a különböző hibatípusokban is hasonló mennyiségű hibát produkált. Az eredmények értelmezésénél viszont érdemes figyelembe venni, hogy a funkció fej elemzés megfelelő működéséhez egy előfeldolgozó lépés is szükséges, amelyben a hiányzó VAN csomópontokat az egyes mondatokba illesztjük, míg a tartalmas fej elemzés nem igényel ilyet.

Jelen cikk kopulás szerkezetekre vonatkozó eredményeinek figyelembevételével a tartalmas fej alapú, univerzális dependencia nyelvtani elemzés tűnik a legmegfelelőbbnek a magyar számítógépes dependenciaelemzésre.

4. Összegzés

Cikkünkben megmutattuk, hogy a tartalmas fej típusú dependencia-nyelvtani elemzéssel előfeldolgozó lépés (üres VAN fejek beszúrása) és a címkék számának jelentős megnövelése nélkül érhetünk el versenyképes eredményeket a magyar *van* létigés szerkezetek elemzésében. Jelen cikk eredményei alapján a magyar nyelv dependenciaszintaxis elemzésére a tartalmas fej típusú, univerzális dependencia elemzést látjuk a legjobban alkalmazhatónak.

További terveink között szerepel jelen kísérlet megismétlése nagyobb tanítókorpuszokkal, hasonló kísérletek elvégzése más szerkezetek alapos megvizsgálására is, valamint a három elemzés tesztelése különböző alkalmazásokban való felhasználhatóság szempontjából: megvizsgálni, hogy milyen hatással van a három elemzés egyes feladatokra, amelyek erősen függnek a szintaktikai elemzéstől (például információkinyerés, véleménykinyerés, gépi fordítás).

Célunk volt emellett azt is megmutatni, hogy egyrészt a szintaktikai kereten kívül magának az alkalmazott, konkrét elméletnek is hatása van az elemzésre, és ez a számítógépes nyelvészetben számszerűen is kimutatható. Másrészt, bár nagyon hasznos és

viszonylag egyszerűen kinyerhető eredményeket ad a hagyományos ULA és LAS kiértékelése egy-egy elemzőnek főként nagyméretű treebankek esetén, a kézi hibaelemzés és nyelvészeti alapú vizsgálat sokkal informatívabb, mélyebb összefüggésekre világíthat rá.

Hivatkozások

1. Simkó, K.I.: Magyar kopulás szerkezetek az elméleti és a számítógépes szintaxisban. Master's thesis, Szegedi Tudományegyetem (2015)
2. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
3. Polguère, A., Mel'čuk, I.A., eds.: Dependency in Linguistic Description. Studies in language companion series. Amsterdam Philadelphia, Pa. J. Benjamins (2009)
4. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
5. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
6. Vincze, V., Farkas, R., Simkó, K.I., Szántó, Zs., Varga, V.: Univerzális dependencia és morfológia magyar nyelvre. In: XII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2016) 322–329
7. Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING 2012: Posters, Mumbai, India, The COLING 2012 Organizing Committee (2012) 1081–1090
8. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97

Szabályalapú szintaktikai elemző szintaktikai szabályok nélkül

Kovács Viktória[1], Simkó Katalin Ilona[2], Szécsényi Tibor[3]

Szegedi Tudományegyetem, Bölcsészettudományi Kar
[1] viki921015@hotmail.com
[2] kata.simko@gmail.com
[3] szecsenyi@hung.u-szeged.hu

Kivonat: Cikkünkben bemutatjuk az általunk készített Prolog alapú magyar nyelvi szintaktikai elemzőt. A szabályalapú szintaktikai elemzéshez szükséges, nagyméretű lexikon kiküszöbölésére a magyarlánc morfológiai elemzőjét használtuk. A szabályrendszer leegyszerűsítésére öt alapszabályra vezetjük vissza az elemzést. Célunk, hogy megmutassuk, hogy szabályalapú elemzővel, kevés szabállyal is kezelni tudjuk a magyar nyelv sok sajátosságát.

1. Bevezetés

A természetes nyelvi kifejezések szintaktikai elemzését alapvetően kétféleképpen végezhetjük el. Statisztikai módszerrel egy nagy nyelvi korpusz alapján megjósolhatjuk, hogy egy adott mondatnak mi a legvalószínűbb szintaktikai szerkezete. Ilyenkor a korpusz mérete és változatossága garantálja, hogy a célmondatban meglevő minden konstrukciót felismerjen az elemző. Szabályalapú elemzőt használva olyan expliciten megfogalmazott helyettesítési szabályokra támaszkodhatunk, amiket az emberi nyelvtudás leírásánál felmerülő elméleti szempontok figyelembevételével alkottunk meg. Ezek a szabályok nem (csak) a megfigyelhető jelenségeket, mondatokat írják le, hanem a lehetségeseket.

A számítógéppel összegyűjtött, valószínűségi jellegű szabályok is egyfajta újraíró szabályoknak tekinthetők, azonban ezek a szabályok sem önmagukban, sem összességükben nem adnak pontos képet, magyarázatot az emberi nyelvek jellemzőiről, csak annak a visszaadására képesek, amiknek létrejöttüket köszönhetik: egy mondatnak, vagy az ő szintaktikai szerkezetének a valószínűségét határozzák meg. Így ugyanúgy kis valószínűségi értéket rendelnek egy ritkán használt konstrukciót tartalmazó, de lehetséges mondatnak/mondatszerkezetnek, mint egy gyakorihoz hasonló, de nem lehetségesnek. Használatukkal a prototipikus mondatok prototipikus elemzéseit kaphatjuk meg.

A szabályalapú elemzők ezzel szemben az emberi szabályalkotás miatt nem képesek az adatok ugyanolyan széleskörű figyelembevételére, mint a számítógépes, statisztikai alapú elemzők. Az elméleti, nyelvész szakemberek leginkább csak az érdekes, a számukra érdekesnek tűnő jelenségekre koncentrálnak. Az egyedi szabálymegállapítások mozaikjából azonban nem feltétlenül jön létre egy koherens szabályrendszer a nyelvre vonatkozóan. A leírt jelenségek bővülésével a leíró szabályrend-

szer is bővül, ezeknek a szabályoknak az egymásra gyakorolt hatását viszont nem mindig veszik figyelembe. A szabályalapú elemzők egyik legnagyobb hibaforrása a szabályok nagy száma.

Előadásunkban egy olyan szintaktikai elemzőt mutatunk be, ami a valószínű mondatok/mondatszerkezetek helyett a lehetséges mondatokat/mondatszerkezeteket elemzi, ugyanakkor a lehető legkevesebb szintaktikai szabályt alkalmazza. Elemzőnk csupán öt szintaktikai szabályt használ, de ezzel a magyar mondatok jelentős részét képes elemezni, egyúttal értelmezni is, úgy, hogy közben figyelembe veszi a magyar nyelv diskurzuskonfiguracionalitását és szabad szórendűségét is.

Az elemző egy számítógépes szintaxis szeminárium eredménye, így nem volt és nem is lehetett célunk, hogy egy teljes, minden jelenséget hatékonyan leíró elemzőt hozzunk létre, hanem csak az, hogy egy ilyen kevés szabállyal dolgozó, szabályalapú elemző megvalósíthatóságát, számítógépes implementálhatóságát bemutassuk.

2. Az elemző felépítése

Az elemző megalkotásánál arra törekedtünk, hogy a lehető legkevesebb újraíró szabályt kelljen alkalmaznunk. Ennek a célnak az eléréséhez a természetes nyelveknek azt a tulajdonságát használtuk ki, hogy a nyelvi kifejezésekben az egy-egy összetevőben előforduló szószerkezetek számát és tulajdonságát mindig az összetevő egyik szava, az összetevő feje határozza meg: a főnévi csoportot a főnév, a névutós szerkezetet a névutó, a mondatot pedig az ige. A különböző lexikai egységek lexikai jellemzésében felsorolásként szerepel, hogy hány és milyen elemekkel kombinálható/kombinálandó. A fejjel kombinálódó más összetevők pedig a fejhez és egymáshoz viszonyítva viszonylag rögzített pozíciókat foglalhatnak el, és az így kialakítható szerkezetek száma az elméleti nyelvészeti kutatások szerint igen kicsi.

2.1. Kötött szórendű összetevők

A magyar nyelv rögzített szórendű szerkezeteinek – a főnévi csoportnak, a névutói csoportnak, a határozói és a melléknévi csoportnak – a leírására a Jackendoff [0] által javasolt X-vonás elmélet három szabálya elegendő. E három szabályból kettő a lexikai fej vonzatait helyezi el a szerkezetben, a specifikálót és a komplementumokat. A komplementumok az X fejhez jobbról csatlakoznak, az így kialakuló közbenső összetevőhöz, az X'-höz balról járul a specifikáló, így adva meg a teljes XP frázist. A kötött szórendű angol mondatszerkezet fejéül szolgáló *gives* igén mutatjuk be a két szabály működését:

A *gives* lexikai leírásában az ő szintaktikai tulajdonságainak a felsorolásán kívül (egyes szám harmadik személyű, ragozott, jelen idejű ige: V[fin, present, 3sg]) jelöljük, hogy milyen specifikálót, azaz igék esetében milyen alanyt kíván maga mellé (spec:⟨NP[nom,3sg]⟩), illetve hogy milyen komplementumokat (egy tárgyat és egy prepozíciós kifejezést: comp:⟨NP[acc], PP[to]⟩). A **komplementumszabály** a fej *gives* elemből és két neki komplementumként megfelelő elemből állítja össze a V' kifejezést:

$$X[\text{spec}:\alpha, \text{comps}:\beta] \rightarrow X[\text{spec}:\alpha, \text{comp}:\langle C\rangle \oplus \beta] \quad C \qquad (1)$$

ahol ⊕ a konkatenáció jele. Mivel előre nem rögzített, hogy egy lexikai elemnek hány komplementuma lehet, a szabály egyszerre csak eggyel kapcsolja össze a fejet, a komplementumlistán legelöl levővel (a szabályban a C változó jelöli), a komplementumlista maradékát (β) pedig továbbadja a létrejövő összetett kifejezés komplementumlistájának. Így a komplementumszabály kétszeri alkalmazásával a *gives a book to Mary* igei csoport szerkezete [[*gives* [*a book*]] [*to Mary*]] lesz, a kategóriája pedig V[fin, present, 3sg, spec:⟨NP[nom,3sg]⟩, comp:⟨ ⟩], ahol a komplementumlista egy üres lista. Látható, hogy a szabály bal oldalán levő elem szintaktikai kategóriája és egyéb szintaktikai tulajdonságai (X), valamint a specifikálólistájának az elemei (spec:α) megegyeznek a fej hasonló tulajdonságaival.

A másik X-vonás szabály a **specifikálószabály**, ami a fejből és a komplementumokból összeállított közbülső összetevőhöz csatolja a fej specifikálóját:

$$X[\text{spec}:\alpha, \text{comp}:\langle\ \rangle] \rightarrow A \quad X[\text{spec}:\langle A\rangle \oplus \alpha, \text{comp}:\langle\ \rangle] \qquad (2)$$

A szabály alkalmazásával a fej lexikai leírásában szereplő specifikálólistáról egyesével kapcsolódnak az összetevők a fejből és a komplementumokból álló közbülső összetevőhöz. A specifikálószabály ismételt alkalmazásával kaphatunk teljes frázisokat (maximális projekciókat), azaz NP főnévi csoportokat, PP névutós szerkezeteket stb., ezeknek mind a specifikálólistájuk, mind a komplementumlistájuk üres lista.

Az X-vonás elmélet harmadik szabályának az ismertetésére, az adjunktum-szabályra itt nem térünk ki, megvalósítása hasonló az előbbiekhez.

2.2. Mondatszerkezetek

A magyar mondatszerkezet az eddig említett szerkezetekkel szemben nem kizárólag az X-vonás szabályrendszer szerint épül fel, hanem az igéből és a komplementumaiból álló, a komplementumszabály segítségével létrehozott igei csoportot megelőzve különböző funkcionális pozíciók találhatóak, a Topikok és a Fókusz pozíciója. Az itt megjelenő elemek maguk is az igei fej bővítményei, de szintaktikai és szemantikai tulajdonságaikban szelektáltak, illetve jelentésükben módosultak. Elemzőnkben a Topik és a Fókusz argumentumok a szerkezet fejének, az igének a lexikai leírásából származóan a komplementumlistához hasonlóan egy topik- és egy (egyelemű) fókuszlistán kerülnek felsorolásra. A *Péter Marinak adott egy könyvet* mondat *ad* igéjének lexikai leírásában az *egy könyvet* kifejezésnek megfelelő egyelemű komplementumlistán kívül még szerepel egy (esetünkben) egyelemű topiklista és egy szintén egyelemű fókuszlista is (az *ad* igének nincs specifikálója): V[fin, present, indef, 3sg, spec:⟨ ⟩, comp:⟨NP[acc, indef]⟩, topic:⟨NP[nom, 3sg]⟩, focus:⟨NP[dat]⟩]. A csak igei fejű szerkezetek, azaz mondatok esetében alkalmazható **fókuszszabály** (3) és **topik-szabály** (4) a már ismertetett két szabályhoz hasonló felépítésű, csak nem a komplementumlista vagy a specifikálólista elemeit fogyasztják, hanem a topik- és a fókuszlista elemeit:

$$V[\text{spec}:\langle\ \rangle, \text{comp}:\langle\ \rangle, \text{topic}:\alpha, \text{focus}:\beta] \rightarrow \qquad (3)$$
$$F \quad V[\text{spec}:\langle\ \rangle, \text{comp}:\langle\ \rangle, \text{topic}:\alpha, \text{focus}:\langle F\rangle \oplus \beta]$$

$$\begin{array}{l} V[spec:\langle\,\rangle, comp:\langle\,\rangle, topic:\alpha, focus:\langle\,\rangle] \rightarrow \\ \quad T \quad V[spec:\langle\,\rangle, comp:\langle\,\rangle, topic: \langle T\rangle \oplus \alpha, focus:\langle\,\rangle] \end{array} \qquad (4)$$

Látható, hogy a (4) topikszabályt csak abban az esetben lehet alkalmazni, amennyiben az ige fej fókuszlistája üres, vagyis a topik-összetevők a mondatban megelőzik a fókuszt.

A bemutatott szabályok alkalmazásával, egy minimális szintaxis segítségével a magyar mondatok szintaktikai elemzését el tudjuk végezni. A mondatok szerkezete ezáltal az elméleti nyelvészeti szempontok szerint kialakított É. Kiss-féle [0] mondatszerkezeti modellnek egy egyszerűsített változata lesz.

A szintaktikai szabályok számának minimalizálásának a feltétele az volt, hogy a szabályokat ne konkrét terminális és nem terminális kategóriák felhasználásával adjuk meg, hanem csak néhány sematizált kategóriával. A sematizált szintaktikai szabályoknak az ára a szavaknak, vagyis a lexikai elemeknek a gazdagabb jellemzése. Ugyanakkor egy jól működő szintaktikai elemző esetében a szavaknak ezeket az egyedi tulajdonságait úgyis mindenképpen fel kell tüntetni a lexikonban.

2.3. A szintaktikai elemző megvalósítása

Az ismertetett szintaktikai implementációja prolog programozási nyelven történt, azonban nem a prolog saját DCG formalizmusa szerint. Ennek az oka az, hogy mivel a szabályok nagyon sematikusak, ezért szükséges a balrekurzió miatt az nem alkalmazható. Ehelyett a left corner recognizer Blackburn és Striegnitz [0] által kidolgozott prolog implementációját alkalmaztuk az általunk kidolgozott szintaktikai szabályokkal.

3. Lexikai elemek, lexikai szabályok

A kevés szintaktikai szabály érdekében részletesen kidolgozott lexikai leírásokat kellett adni a lexikai elemekhez. Nem csak a szintaktikai kategóriát kellett feltüntetnünk, hanem a lexikai elemek egyéb szintaktikai-morfológiai tulajdonságait is, továbbá a lexikai elemek kombinációs képességét is. Azonban a lexikonnak csak a szintaxis felől nézve kell nagynak lenni, mivel a lexikai elemek különböző morfológiai és argumentumszerkezeti variánsai a bázis lexikai elemekből lexikai szabályok segítségével levezethetőek. A bázis lexikon méretét egy morfológiai előelemző felhasználásával tovább csökkentettük.

3.1. Morfológia

A nagyméretű lexikon létrehozásának elkerülésére az elemzőnk a magyarlánc [0] morfológiai elemző modulját is felhasználja.

Az elemezni kívánt mondatokat először ezzel a külső, magyarlánc modullal elemezzük. A magyarlánc MSD-kódokkal látja el a mondat szavait a szófajuknak és

egyéb morfológiai jegyeiknek megfelelően. Az MSD-kód rengeteg morfológiai információt tartalmaz; ezekből a mi elemzőnk egyelőre nem mindet használja fel.

Az MSD-kódokból kinyerjük a szófaji meghatározást; főnevek esetén a számot, esetet és birtokviszonyt jelölő részeket; melléknevek esetén a számot, esetet és fokot; igéknél az igemódot, időt, számot, személyt és a tárgy határozottságára vonatkozó információt; ezek mellett az MSD-kód alapján elemezzük a névmások és határozott, valamint határozatlan determinánsok típusait is.

Manuálisan hozzáadott morfológiai információ csak az argumentumszerkezettel rendelkező elemek, azaz leginkább az igék argumentumszerkezetére vonatkozó részekben van az elemzőnkben. Mivel a magyarlánc a szavak szótövezését is elvégzi, a szótövek alapján egy kivétellista segítségével meghatározzuk, hogy melyik argumentumszerkezetes szó milyen argumentumszerkezet-típusba tartozik, majd a típusának megfelelő argumentumszerkezeti leírást rendelünk hozzá. A *látja* szóról a magyarlánc azt az információt adja vissza, hogy a töve *lát*, az MSD-kódja pedig Vmip3s---y. A *lát* tövű Vxxxxxxxxx MSD-kódú szavakhoz *vtr* tranzitív igei típust rendelünk, majd az MSD-kód alapján meghatározzuk a szó egyedi szintaktikai és morfológiai tulajdonságait (kijelentő módú, finit, jelen idejű, határozott ragozású, 3sg egyeztetési jegyű), majd a *vtr* típus és a már meghatározott jegyek generáljuk az argumentumszerkezetét: comp:⟨NP[nom, 3sg], NP[acc, def]⟩. Igék esetében a spec, topic és focus listák az alap lexikai leírásban mind üresek, azokat majd a később ismertetett lexikai szabályok töltik fel.

A kivétellistán nem szereplő, argumentumszerkezet nélküli vagy prototipikus argumentumszerkezettel rendelkező szavak lexikai leírását (főnevek, melléknevek, kötőszavak stb.) a szótő figyelembevétele nélkül, csak a szó MSD-kódjára támaszkodva hozzuk létre.

Az MSD-ből kinyert morfológiai információkat és az X-vonás elmélet alapszabályait összeillesztő egyszerű szabályok megalkotásával már képesek vagyunk egyszerű mondatok elemzésére intranzitív (1. példamondat), tranzitív (2. példamondat), ditranzitív igék (3. példamondat) és a kopula (4. példamondat) esetén is. Ezek a szabályok még kötött SVO szórenden működnek, ezeket bővítettük ki a nyelvspecifikus jelenségeket leíró szabályokkal.

1. A kutya fut.
2. Mari kergeti Pétert.
3. Mari kutyát ad Péternek.
4. A kutyák voltak pirosak.

3.2. Pro-drop

A magyar nyelv egyik tulajdonsága a pro-drop. Mivel az igén morfológiailag rengeteg információ megjelenik az alanyra és a tárgyra nézve, így azokat nem mindig szükséges expliciten megjeleníteni a mondatban. Megjelenhet az alanyi és a tárgyi összetevő is (5. példamondat); eltűnhet csak az alany (6. példamondat) vagy csak a tárgy (7. példamondat), de akár mindkét névszói összetevő is (8. példamondat).

5. Én látom őt.

6. Látom őt.
7. Én látom.
8. Látom.

Elemzőnket kibővítettük ennek a jelenségnek a kezelésére. Az igei csoportokat létrehozó szabályokat átalakítottuk, hogy alanyi és/vagy tárgyi összetevő hiányában is létre tudják hozni a mondatot. Egy-egy opcionális lexikai szabály segítségével a már létező igei lexikai elemek komplementumlistájáról törölhetjük az alanyi és a tárgyi főnévi csoportokat:

$$V[comp:\alpha \oplus \langle NP[nom] \rangle \oplus \beta] \Rightarrow V[comp:\alpha \oplus \beta] \quad (\mathbf{5})$$

Így a *látja* lexikai elemből, aminek a komplementumlistája comp:⟨NP[nom, 3sg], NP[acc, def]⟩, a lexikai szabály comp:⟨NP[acc, def]⟩ komplementumlistájú *levezetett* lexikai elemet képez.

3.3. Kopula

A magyarra jellemző, hogy a kopula harmadik személyű alany mellett jelen időben és kijelentő módban nem jelenik meg a mondat felszíni szerkezetében (9. példamondat), de egyéb esetben megjelenik (10. példamondat).

9. Ő katona.
10. Ő katona volt.

Ige nélküli mondatokban az elemzőnk képes alany esetű főnévből ige nélkül igei csoportot képezni. A főnév megtartja a komplementumlistáját, és ehhez hozzáadjuk az igei csoporthoz szükséges alanyesetű főnevet, vagyis az alanyt. Az így kapott igei csoport jelen idejű és harmadik személyű, számában pedig a főnévi predikátummal egyezik:

$$N[nom.NUM,comp:\alpha] \Rightarrow V[fin, present, 3NUM, comp: \langle NP[nom, NUM] \rangle \oplus \alpha] \quad (\mathbf{6})$$

Ilyen módon az *Ő katona*, ige nélküli mondatban a *katona* főnévből (N[nom, 3sg, comp:⟨ ⟩]) egyes számú igei csoportot (V[fin, present, 3sg, comp:⟨NP[nom, 3sg]⟩]) képezhetünk. Ennek a komplementumlistájára kerül fel az alany, ami az *ő* lesz a további mondatszerkezetben.

3.4. Topik és fókusz

A korábban bevezetett topik és fókusz szintaktikai szabály működését egy-egy lexikai szabály egészíti ki. Ezek a szabályok alakítják át az igék eredeti argumentumszerkezetét úgy, hogy a megfelelő elemek a topik és fókusz pozíciókba kerülhessenek.

A (7) fókusz szabály egy ige üres fókuszlistájára rakhat át egy elemet a komplementumlistáról, ezzel biztosítva azt, hogy egy mondatban csak egy fókusz jelenjen meg az ige előtt. Az így a fókuszlistára mozgatott elemet fogja a szintaktikai fókusz szabály az ige előtt pozícióban elemezni.

$$V[\text{focus: } \langle\,\rangle, \text{comp: } \alpha \oplus \langle A \rangle \oplus \beta] \Rightarrow V[\text{focus: } \langle A \rangle, \text{comp: } \alpha \oplus \beta] \quad (7)$$

A (8) topik szabály nem csak üres topik listán működhet, egy mondatban több topik is megjelenhet az ige előtt. A szabály a listához kapcsol még egy elemet a komplementum listáról. A szintaktikai topik szabály ennek a listának az elemeit elemzi az ige előtt.

$$V[\text{topic: } \alpha, \text{comp: } \beta \oplus \langle A \rangle \oplus \gamma] \Rightarrow V[\text{topic: } \alpha \oplus \langle A \rangle, \text{comp: } \beta \oplus \gamma] \quad (8)$$

Ilyen módon a *Péter Marinak ad egy kutyát* mondat lehetséges elemzése a következő: az *ad* ige (V[comp:⟨NP[nom, 3sg], NP[acc, indef], NP[dat]⟩, topic:⟨ ⟩, focus:⟨ ⟩]) komplementumlistájáról a lexikai fókusz szabály a *Marinak* összetevőnek megfelelő elemet a fókusz listára mozgatja (V[comp:⟨NP[nom, 3sg], NP[acc, indef]⟩, topic:⟨ ⟩, focus:⟨NP[dat]⟩]), a lexikai topik szabály pedig a *Péter* összetevőt rakja át a topik listára (V[comp:⟨NP[acc, indef]⟩, topic:⟨NP[nom, 3sg]⟩, focus:⟨NP[dat]⟩]). A komplementum listának így már csak egy eleme marad: az *egy kutyát*, ez az elem jelenik meg egyedül az ige után. A szintaktikai topik és fókusz szabályok a topik és fókusz listák tartalmát az ige előtt elemzik a megfelelő sorrendben.

3.5. Scrambling

Az ige utáni pozícióban az összetevők szabad sorrendben jelenhetnek meg a magyar mondatokban. Egy újabb lexikai szabállyal kezeltük ezt a jelenséget, ami biztosítja, hogy az itt lévő összetevők szabadon keveredhessenek.

11. Odaadta Mari Péternek a kutyát.
12. Odaadta Péternek Mari a kutyát.
13. Odaadta Mari a kutyát Péternek.

A scrambling szabály az igei csoportok komplementum listájának elemeit az összes lehetséges sorrendben előállítja, a lista elemeinek összes permutációját generálja:

$$V[\text{comp:}\alpha] \Rightarrow V[\text{comp:Perm}(\alpha)] \quad (9)$$

Ez a szabály hossza létre a magyarra jellemző ige utáni szabad szórendet. Így az *odaadta* ige komplementumlistáján szereplő három összetevőt: *Mari*, *Péternek*, *a kutyát* az összes lehetséges sorrendbe állítja az ige után, a három vonzat mind a hat lehetséges sorrendjével képes elemezni a mondatot.

3.6. A lexikai szabályok sorrendje

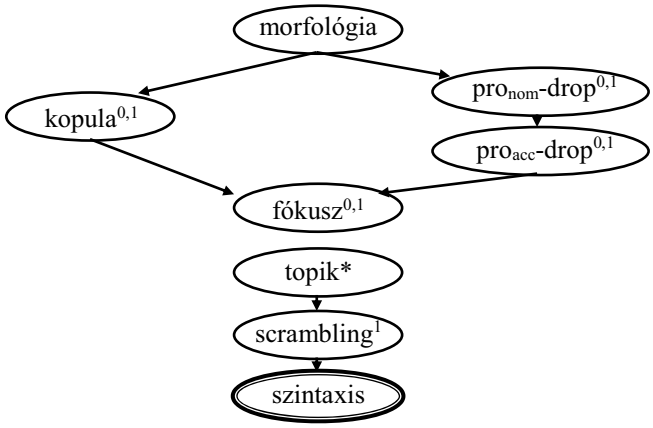
A 2. szakaszban megadott sematikus szintaktikai újraíró szabályok az így előkészített lexikai elemekből egyértelműen képesek felépíteni a mondatot (vagy képesek visszavezetni egy mondatot lexikai elemekre). A lexikai elemek viszont a lexikai szabályok miatt nem egyértelműek. Ugyanaz a lexikai elem többszörösen is elérhetővé válik a szintaktikai elemzésben, ha ugyanazon alap lexikai egységből különböző módon generáljuk. A *Péternek Mari ad egy kutyát* mondat *ad* igéjét például ugyanabból az *ad* alap lexikai egységből előállíthatjuk úgy, hogy először a (7) fókusz lexikai szabályt

alkalmazzuk, majd utána a (8) topik lexikai szabályt, vagy fordítva, először a topikot emeljük át a topiklistára, és csak utána a fókuszt.

Az ilyen jellegű üres lexikai többértelműséget a lexikai szabályok hierarchiába rendezésével lehet megszüntetni: véges automataként szabályozzuk, hogy egy lexikai szabályt hányszor lehet alkalmazni, két lexikai szabályt milyen sorrendben kell elvégezni, illetve egy lexikai szabály alkalmazása milyen más szabályok alkalmazhatóságát zárja ki: a *zéró kopula* lexikai szabály nem alkalmazható az *alanyi* és a *tárgyi pro-drop* szabállyal együtt. Ha az *alanyi* és a *tárgyi pro-drop* lexikai szabályokat is alkalmazzuk, akkor az *alanyi pro-drop* szabálynak meg kell előznie a *tárgyi pro-drop* lexikai szabály alkalmazását. Ezek a lexikai szabályok csak a *fókusz* és a *topik* lexikai szabályok előtt alkalmazhatóak, a *fókusz* és a *topik* lexikai szabályok együttes alkalmazása esetén a *fókusz* szabály megelőzi a *topik* szabályt. A *scrambling* lexikai szabályt legutoljára kell alkalmazni.

A *scrambling* szabályt pontosan egyszer kötelező alkalmazni, a *topik* szabályt akárhányszor, az összes többit legfeljebb egyszer.

A lexikai szabályok sematikus rendezése (^{0,1}: legföljebb egyszer; ^{1}: pontosan egyszer; *:akárhányszor alkalmazható):



4. Összegzés

Cikkünkben bemutattuk Prolog alapú magyar szintaktikai elemzőnk működését. Az elemző összesen öt szintaktikai szabállyal elemzi a mondatokat külső lexikon, a magyarlánc morfológiai elemző felhasználásával, így elkerülve a nagyméretű szótár használatát.

A jövőben tervezzük, hogy az elemzőt bővítjük. Terveink között szerepel az igék argumentumszerkezeteinek automatikus hozzáadása az elemzőhöz és összetett mondatok kezelése.

Hivatkozások

1. Blackburn, P., Striegnitz, K.: Natural Language Processing Techniques in Prolog. http://cs.union.edu/~striegnk/courses/nlp-with-prolog/html/index.html (2002)
2. É. Kiss, K: The Syntax of Hungarian. Cmabridge, Cambridge University Press (2002)
3. Jackendoff, R.: X-bar-Syntax: A Study of Phrase Structure. Linguistic Inquiry Monograph 2. Cambridge, MA: MIT Press (1977)
4. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács, A., Vincze, V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2013) 368–374

Mozaik nyelvmodell az AnaGramma elemzőhöz

Indig Balázs[1,2], Laki László[1,2], Prószéky Gábor[1,2,3]

[1] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[3] MorphoLogic
e-mail:{indig.balazs, laki.laszlo, proszeky.gabor}@itk.ppke.hu

Kivonat Cikkünkben bemutatjuk az elemző rendszerünkhöz a rendelkezésünkre álló nagyméretű magyar nyelvű korpuszok felhasználásával készített modult, amely szimulálni tudja az emberi elemzőkön megfigyelt jelenséget, miszerint bizonyos gyakori szerkezetek feldolgozása egyfajta gyorsítótárazás segítségével az átlagosnál gyorsabb. Létrehoztunk egy olyan rendszert, amellyel 3-nál magasabb gramok esetén, több faktor kombinálásával gyakori mintákat tud előállítani. Megvizsgáltuk a keletkezett mintákat, a szintaktikai elemzés gyorsításának szempontjából, beleértve az őket alkotó példák különböző teljes kifejtésű eloszlásait. Az ilyen minták megfigyelésével a szakértő szemlélő további ötleteket nyerhet, a korpuszon megfigyelhető jelenségek keresésére. Felsorolunk továbbá néhány az elemző szempontjából érdekes példákat is.

Kulcsszavak: nyelvmodell, korpuszminták, nyelvi elemző, big data

1. Bevezetés

Az MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport a létező megközelítésektől merőben eltérő nyelvelemzőrendszer létrehozását tűzte ki célul. Az általuk létrehozott AnaGramma nevű [1] pszicholingvisztikai indíttatású elemzőrendszer lényege, hogy működésével az emberi nyelvelemzést modellezi.

Ennek megfelelően a következő főbb tulajdonságokkal rendelkezik: (1) Szigorúan balról jobbra, szavanként dolgozza fel a szöveget, így nincs mód az elemzéshez a mondat azon részének a felhasználására, amely még nem került be a rendszer látókörébe. (2) Az elemző több, független, párhuzamosan futó modulból áll, amik kommunikálnak egymással. (3) Performanciaalapú rendszerként minden olyan nyelvi jelenséget megpróbál feldolgozni, ami leírt szövegekben előfordul, viszont a csak elméleti szinten létező szerkezetek elemzése nem tartozik az elsődleges céljai közé.

2. Mozaik n-gramok

Munkánk során a fent vázolt rendszer egy olyan modulját készítettük el, amely a rendszer egésze által támasztott kritériumoknak megfelel. A modul feladata egy olyan adatbázis felépítése és lekérdezése korpuszok felhasználásával, amely

tárolja a nyelvben előforduló *gestaltokat*, azaz olyan gyakori szerkezeti mintákat, melyeket az emberi elemző a teljes elemzés helyett *egészleges feldolgozás* segítségével gyorsan, „egy lépésben" kezel. Az előre eltárolt, megelemzett „mentális reprezentáció" egészben hívódik elő a memóriából a szerkezet valósidejű felépítése helyett.

Ilyen minták lehetnek például az *állandósult szókapcsolatok*, a *többszavas kifejezések*, *főnévi frázisok* vagy bármilyen más gyakori, összefüggő szerkezetek, amelyek a korpusz alapján megfigyelhetők. A következő felsorolásban néhány példamondatot gyűjtöttünk össze:

- **Többszavas kifejezések**: „az ördög ügyvédje"
- **Szólásmondás**: „Itt van a kutya elásva."
- **Udvariassági formula**: „Jó [NAPSZAK][ACC]!"
- **Merev szerkezetek**:
  „Az országgyűlés a javaslatot [SZN|DIGIT][NOM] igennel... elfogadta."
- **Igei szerkezetek**: „lemma:es(ik) szó [DEL]"
- **Névelem**: „Petőfi/[VEZ.NÉV] Sándor/[KER.NÉV] utcai/[KÖZT.TÍPUS] általános/[INT.TÍPUS] iskola/[INTÉZMÉNY]"
- **Név + titulus**:
  „Orbán/[VEZ.NÉV] Viktor/[KER.NÉV] Magyarország/[ORSZÁG|SZERVEZET] miniszterelnöke/[TITULUS]"

Az általunk létrehozott modul legfontosabb tulajdonsága, hogy a nyelvi intuíciónkat „utánozza", miszerint kategoriális helyettesítéseket alkalmazva mintákkal leírhatók a nyelvi jelenségek. Ennek alátámasztására nagy korpuszokon (MNSZ 1-2, lásd 1. táblázat) számszerűsítve vizsgáltuk különböző hosszúságú, összefüggő n-gramok számosságát, mert előzetes tapasztalatainkban úgy találtuk, hogy bizonyos minták esetén egyes szóalakok *kategóriáikkal (faktorokkal)* való helyettesítése olyan számszerű összefüggéseket tár fel, amik a hagyományos n-gramok esetén nem látszódnak. Így olyan gyakori mintákat generáltunk a korpuszokból, amelyekben „szükség szerint" a szóalakok helyettesítve lettek a lemmájukkal, illetve a szófaji címkéjükkel. Az így létrejött heterogén felépítésű n-gramokat nevezzük *mozaik n-gramoknak*[4].

1. táblázat. Különböző korpuszok jellemzői

Neve	Mondatok száma	Tokenek száma	Mondatokban az átlagos tokenszám
Szeged Korpusz 2	70 990	1 194 348	16,824
MNSZ1	18 657 302	264 465 825	14,175
MNSZ2	28 777 590	444 760 553	15,455
Szószablya	24 991 306	462 024 888	18,487
huTenTen12	-	3 184 161 466	-

[4] Bár a módszer és az elkészült rendszer, képes más jellegű, illetve több független faktor együttes kezelésére, cikkünkben csak a fent említett faktorokat használtuk.

A mozaik n-grammokkal leírt minták elemzés nélkül, egészben történő feldolgozása analóg a számítógépeknél ismert *gyorsítótárazáshoz*, idegen szóval *cacheeléshez*. Az AnaGramma alapgondolata szerint, az emberi feldolgozás gyorsasága nagyrészt a gyakori esetek egészleges feldolgozásának tudható be, mivel az előre tárolt minták segítségével nagymértékben csökkenthető a mondatszintű elemzés komplexitása, aminek következtében javulhat az elemzőrendszer minősége.

A modul egy másik fontos feladata, hogy a gyakori szerkezetek ismeretében képes megjósolni az elemzés során a szerkezeti határokat. Mivel a gyakori szerkezetekből és a Grice-i maximák [2] alapján arra számít az elemző, hogy a szöveg folytatása kiszámítható és az előzetes tapasztalatoknak megfelel. Ennek köszönhetően az elemzőrendszer többi moduljának képes jelezni, hogy az elemzés valószínűleg elért egy szerkezeti határt, vagy éppen hogy egy nagyobb szerkezeti egység közepén tart. Így képesek vagyunk támogatni az elemzőrendszer azon moduljait, amelyek a különböző elemzési szintek szerkezeteit keresik (pl. szintaktikai elemző, NP/VP chunker). A rendszer előnye továbbá, hogy *nyelvi modellként* képes lesz számszerű becslést adni a soron következő szóra és/vagy kategóriára.

Munkánk során nagy hangsúlyt fektettünk arra, hogy az általunk létrehozott rendszer valós időben tudjon a keresett mintákra példákat adni nagyméretű korpuszokból trigramnál magasabb rendben is. Ez különösen fontos, ha figyelembe vesszük, hogy az elemzés komplexitása exponenciálisan nő a különböző kategóriák számával, valamint az elemzett mondat hosszával.

A fent említett minták keresése azért kevéssé kutatott téma, mert nagyon nagy tárhelyet igényel, illetve napjainkig a számítási kapacitások szűkösek voltak. A keresési tér nagyon rosszul skálázódik (lásd 2. táblázat), ezért szükséges számtalan premissza figyelembevétele, úgy hogy a zajokat csökkentsük a rendszerben, míg a keresett elemeket megtartsuk.

2. táblázat. A korpuszokban mért kifejezések mennyisége

	Szeged Korpusz 2		MNSZ1		MNSZ2	
	szó	WLT	szó	WLT	szó	WLT
1-gram	129 273	181 279	6 297 534	9 519 354	7 208 999	8 650 798
2-gram	578 642	2 962 756	57 770 805	236 296 463	73 919 408	299 373 602
3-gram	918 915	17 641 621	135 616 024	2 028 400 881	191 820 777	2 589 580 489
4-gram	1 019 316	67 750 636	184 815 630	10 241 065 746	280 556 568	12 498 795 104
5-gram	998 515	213 126 488	197 430 850	28 785 417 930	314 801 331	42 073 197 888
6-gram	946 278	618 181 519	192 819 805	88 556 351 842	310 102 954	131 117 731 010
7-gram	887 086	1 742 852 595	182 743 426	259 778 917 230	305 349 214	400 011 439 879
8-gram	826 405	4 825 618 452	171 459 179	731 213 387 722	289 872 274	1 179 148 233 622
9-gram	766 638	13 429 864 821	160 185 064	2 207 045 830 298	273 095 868	3 493 974 880 398

3. Kapcsolódó munkák

Korpuszbeli minták keresésével a *Mazsola* nevű rendszer [3] is foglalkozik, de az igei vonzatkeretek detekciójánál, az igék sajátosságaiból fakadóan, a szavak

sorrendje sokkal szabadabb, mint az általános, főleg főnévi csoportokat és gyakori szekvenciáikat tartalmazó szerkezeteké. További jellemvonása a módszernek, hogy szintaktikailag elemzett bemenettel dolgozik, mely feltétel a mi megközelítésünknek ellentmond.

A legközelebbi hasonló implementáció a *SRILM* [4] nevű eszközben megvalósított faktoros nyelvmodell [5], mely módszer használható bigramra, de legfeljebb trigramra. Magasabb nyelvmodell kezelésére viszont a magas erőforrásigény és futásidő miatt nem alkalmas. Ezért a tár és számítási kapacitások figyelembevételével eltérünk a faktoros nyelvi modellektől. Ahogy az emberi elemzés során is, mi is csak a gyakori szerkezetek vizsgálatára hagyatkozunk, nem kívánunk teljes modellt adni a nyelvhez. Ezzel a feladat a létező kapacitások határain belül tartható, de alapjaiban más megközelítésre van szükség.

A Sketch Engine [6] a napjainkban elérhető legátfogóbb korpuszkezelő rendszer, melynek a nyílt forráskódú változatát (NoSketchEngine [7]) használtuk a korpuszokban való keresésre, illetve az általunk megtalált mintákhoz példák és gyakorisági eloszlások generálásához. Számtalan funkciója között megtalálható az n-gramok generálása is, de sebességben tapasztalataink szerint közel azonos teljesítményt nyújt a mi rendszerünkkel, továbbá nem képes mozaikgramok generálására. Ezért szükségszerű volt egy saját rendszer fejlesztése, mely kiegészítőként szolgál ehhez a maga területén rendkívül hatékony az eszközhöz.

4. Módszerek

Alapvetően a Humor kódokkal [8] elemzett és a Szeged korpuszon [9] tanított PurePOS 2.1-gyel [10] egyértelműsített korpuszokon vizsgálódtunk és három faktort vettünk figyelembe, a szóalakot, a lemmát és a szófaji címkét. Egy token az ezekből alkotott hármasokkal (WLT) volt reprezentálva és a hagyományos csak szóalakot vagy csak címkét tartalmazó n-gramok helyett a szótöveket is és a három faktor tetszőleges kombinációit is megvizsgáltuk. A bevezetett kategoriális megkülönböztetések segítségével elénk tárulnak olyan valóban gyakori esetek is, melyek csak a kategóriájuknál fogva képezik gyakori minták részét. A modul képes további kategóriák kezelésére is. Így lehetnek akár szemantikai jellegű megszorítások is (élő, intézmény, nyelv stb.).

Egy külön mérésben kíváncsiak voltunk továbbá arra, hogy ha egy egyszerű binárisan eldönthető kérdéssel „*Főnévi csoport (NP)* vagy nem főnévi csoport része az adott szó?" géppel felcímkézett nagy méretű korpuszok esetében az NP-k és az egymás mellett álló NP-k (mivel ezek nincsenek megkülönböztetve az egyszerűség kedvéért) belső szerkezete milyen tipikus mintázatokat mutat.

4.1. A nagy adatok problémája

A nagy korpuszok gyakran esnek olyan hibába, hogy mivel nincs emberi kapacitás kézzel elvégezni az annotációt, ezért gépi eszközöket futtatnak rajta, amik a pipeline architektúra miatt, felnagyítják a már meglevő hibákat. Például a nagy

korpuszok vizsgálata során találtunk olyan esetet, amikor egy táblázat egyes mezői, amik számokat tartalmaztak, külön-külön mondatokká lettek alakítva egy szavas, számokból álló mondatokat alkotva. Ezért alaprendszerként két egyszerű n-gram alapú nyelvfelismerőt[5] futtatunk a korpuszok mondatain és ahol mindkettő magyar nyelvűnek ítélte az adott mondatot, azt tekintettük jó mondatnak. Ezzel a nagyon durva méréssel a korpuszok kb. 30%-át találtuk használhatónak a modellalkotásunk céljára. Tudjuk, hogy a nyelvfelismerők a rövid (3-4 szavas) mondatok esetében nem mindig rendelkeznek elég információval a döntéshez, ezért tévednek, ám az általunk készített modellek esetében a rövid mondatok nem rendelkeznek elég információval, így a kihagyásuk nem okoz problémát. A mérést finomítani szeretnénk a jövőben a fordításminőségbecslő algoritmusok [12] korpuszminőségbecslő algoritmussá alakításával.

4.2. Felhasznált eszközök és technikák

A fenti számítások (2. táblázat) alapján úgy találtuk, hogy memóriában tárolni nem tudjuk egyben a szükséges adatokat, ezért lemezre írva kell tárolni és ennek a kritériumnak megfelelően feldolgozni minden köztes adatot. Választásunk az egyszerűbb, standardabb feladatok esetén *Unix Coreutils* parancsaira mint a *sort*, *uniq*, *(e)grep*, stb. esett, mert ezek lemezorientáltan nagyon hatékonyak és több tíz éves fennállásuk óta sokszor sikeresen alkalmazták őket hasonló területeken. Míg a bonyolultabb számításokat az *AWK* nyelv különböző variánsaival végeztük, mivel előzetes méréseinkben úgy találtuk, hogy már kicsi korpuszméreten is kiemelkedően jól teljesítenek, a gyors prototípus alkotást lehetővé tevő szkript nyelvekhez (mint a Pyhton és Perl) képest megtartva a gyors változtatások lehetőségét. Az AWK nyelv különböző implementációira azért volt együttesen szükség, mert az Unicode karaktereken történő változtatásokat nem igénylő feladatok, az MAWK[6] variánssal sokkal gyorsabban futottak le, mint a szabványnak tekinthető GNU AWK-val[7], míg az utóbbi segítségünkre volt az Unicode kisbetűsítés gyors elvégzésében.

4.3. A Zipf-görbe vágása

A korpusznyelvészetben jól ismert Zipf-görbe [13], ami egy szó előfordulási gyakoriságának és a gyakorisági táblában levő rangjának függvénye. Ez a görbe „emberi szemmel nézve" nagyon hasonlít a reciprok függvény egy változatára. A főbb alaktani tulajdonságai megmaradnak akkor is, ha nem szavakon, hanem lemmákon, címkéken vagy éppen ezekből alkotott n-gramokon nézzük a gyakoriságokat. A görbe lecsengése minden esetben nagyon hosszú, ezért a nagyon magas számú *Hapax Legomenonok*, illetve nagyon ritka elemek tárolására, amik statisztikailag nem bírnak információtartalommal, nincs szükség. Ezért szükséges egy *alsó küszöb* meghatározása, amivel a keresési tér méretét csökkentjük. A célunk az

[5] langid.py (https://github.com/saffsd/langid.py) és A textCat nyelvfelismerőt [11]
[6] http://invisible-island.net/mawk/
[7] https://www.gnu.org/software/gawk/

ember fejében alkalmazásra készen álló gyakori szerkezetek megtartása, a passzív nyelvtudást nem kívánjuk modellezni, ezért szükséges egy *felső korlát* meghatározása, ami a gyakori, aktívan behelyettesíthető mintákat elválasztja a passzív nyelvismerettől. A fenti kettő küszöb megállapítása szükségszerűen automatikus kell, hogy legyen és a korpusz méretétől függetlenül azonos eredményt kell produkáljon. Továbbá a kísérletezésnek teret hagyva kellően finoman változtatható kell, hogy legyen.

Célszerűnek látszott a görbéhez egy megfelelő meredekségű érintő húzása, mely a szóban forgó számoktól függetlenül megadja azt a pontot, ahol a szavak gyakorisága és rangja a kívánt arányt éri el. Felső korlátnak heurisztikusan a 45 fokot választottuk, míg alsónak a 10 fokot választottuk kiindulásként. Tapasztalataink szerint az első az elemek kevesebb mint 1%-ánál metsz, míg az alsó korlát hozzávetőlegesen az elemek 50%-át tartja meg. Így a memória és tárbeli korlátainkba beleférünk.

A „görbe hasonlat" viszont csak akkor alkalmazható, amíg nem számítógépnek kell feldolgoznia az adatot, ugyanis közelebbről megnézve a görbét láthatjuk, hogy az nem folytonos és nem is görbe, hanem lépcsőzetes egyenes vonalakból áll. Ha élünk azzal a közelítéssel, hogy az egyes ugrásait a függvénynek összekötjük, ezzel nulla meredekségű egyenes vonalakból különféle meredekségű egyeneseket képezve, akkor is csak egy (szabálytalan) törtvonalat kapunk. Továbbá mivel nem áll végtelen adat rendelkezésünkre, ezért bizonyos helyeken az adathiány miatt, egynél nagyobb ugrások is gyakran előfordulnak, így tovább csúfítva a „görbénket". Az így kapott függvényt tovább kéne interpolálni, hogy az érintő numerikus deriválással kiszámolható legyen.

Ezért a görbe további alakítása helyett az érintő megkeresését egy egyszerű konvex lineáris programozási feladatra (minimalizálásra) redukáljuk. A kapott egyenesek meghatároznak egy félsíkot és a megfelelő meredekségű egyenes mint célfüggvény minimalizálásával (balra tolásával) megkapjuk azt a pontot, ahol „utoljára metszi a görbénket". Ezt a pontot egészre kerekítve kapjuk meg a keresett küszöb értéket, ahol vágni szeretnénk. Az eljárás előnye, hogy gyors és az adatok minőségétől és a görbe meredekségétől függetlenül működik.

4.4. A megtalált minták súlyozása

A kigyűjtött mintáknál sok olyan eset fordul elő, hogy azonos gyakorisággal egy adott minta többször több formában is előfordul. Ezeket összevonjuk egy mintává a legspecifikusabbat megtartva, hogy ne befolyásolják kedvezőtlenül az igazi minták rangját. A gondos eljárás ellenére, a taggelési hibák miatt kicsit eltérő gyakorisági számú, ám azonos minták is előfordulhatnak. Jelenleg, ezekkel a kis számuk és rendezetlenségük miatt nincs értelme foglalkozni. Többségük a korpusz hibáiból adódik.

A minták relevanciájának megállapításában az előfordulási gyakoriságuk mellett fontos szempont, hogy mekkora mértékben részei nagyobb mintáknak. Egy minta fontosságát nagy mértékben csökkenti, ha az gyakran része egy nagyobb szerkezetnek. Ezt oly módon vettük figyelembe, hogy a kisebb részminta előfordulási gyakoriságát csökkentettük az őt tartalmazó nagyobb minta súlyozott

gyakoriságával. Ezt a súlyozási technikát Frantzi et al. [14] c-value-nak nevezték el. A módszer lényege, hogy miután az összes, a feltételeinknek megfelelő n-gramot kigyűjtöttünk ($1 < \mathrm{n} < 12$), mindegyikre meghatároztuk a hozzá tartozó *c-value*-t, ami az adott n-gram *kifejezés voltára utaló mérőszám* mely pontosabb képet ad a gyakorisági értékeknél. Ez az érték a következő képlettel írható le:

$$C_{value(a)} = log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) \quad (1)$$

, ahol

- a: a vizsgált kifejezésjelölt
- $f(a)$: a kifejezésjelölt gyakorisága
- $|a|$: a vizsgált kifejezés hossza
- $P(T_a)$: annak a gyakorisága, hogy a jelölt hányszor fordul elő hosszabb kifejezés részeként
- $f(b)$: az ilyen, hosszabb kifejezések száma

5. Eredmények

Kivettük a korpuszból a [PUNCT] POS-taget és a hozzátartozó lemmát, mert úgy találtuk, hogy a szempontunkból fölösleges és nem mond többet, mint a „szóalak". Illetve hasonlóan azokat a lemmákat is kiszűrtük, amelyek megegyeztek a szótővel, ezzel növelve a minták átláthatóságát és csökkentve az állapotteret. Ennek eredményeképpen a számszerű gyakoriságok megváltoztak, a fölösleges minták egy része eltűnt, de az egész rendszer viselkedése érdemben nem változott.

Az MNSZ 2.0 korpusz vizsgálatakor úgy találtuk, hogy az 5-nél hosszabb gyakori szerkezetek kizárólagosan a parlamenti doménből származtak és általános nyelvi információk nem voltak kinyerhetőek belőlük. Azt találtuk, hogy a doménektől nagyon erősen függenek az ilyen szerkezetek. A szaknyelvi zsargonok és fordulatok használata épp úgy, mint az etikett és az egyéb udvariassági konvenciók betartásával keletkező „részben merev" ismétlődő szerkezetek nagy eltéréseket okoznak a domének között a gyakori mintákban.

Ha egy gyakori mintákhoz megnézzük, a konkrét előfordulásokat, amit a 3. táblázat mutat, láthatjuk, hogy bár az első szó meghatározása [FN][NOM] – biztosabbat nem tudunk mondani róla – a szófaji címkék finomítása ebben az esetben kívánatos lenne, mert a példák alapján látható, hogy a főnevek egy speciális alosztálya statisztikailag szignifikánsan megfigyelhető. Kivételek persze akadnak elenyésző számban, de ezek még mindig elemezhetők a hagyományos elemzéssel, míg a meghatározott gyakori osztály előelemezhető és mintaként beilleszthető. Továbbá a példában szereplő *azt* egy speciális esete az „azt" szónak, ahol a „mondta , hogy" szerkezet következik és nincs szükség visszamenőleges koreferenciafeloldásra, ami úgy állapítható meg, hogy egy szóval előre tekintünk az aktuális állapothoz képest, hogy felismerjük a „gyorsítási lehetőséget". Vegyük észre továbbá, hogy bár elméleti lehetősége van, az „ezt" szó nem szerepel hasonló kontextusban a korpuszban.

3. táblázat. *„[FN][NOM] [FN|NM][ACC] lemma:mond , [KOT]”* mintákhoz tartozó szóalakok és azok előfordulási gyakorisága

Vizsgált minta					Gyakoriság
[FN][NOM]	**[FN|NM][ACC]**	**lemma:mond**	**,**	**[KOT]**	**11918**
úr	azt	mondta	,	hogy	906
úr	azt	mondja	,	hogy	304
törvény	azt	mondja	,	hogy	176
miniszterelnök	azt	mondta	,	hogy	168
miniszter	azt	mondta	,	hogy	158
asszony	azt	mondta	,	hogy	126
államtitkár	azt	mondta	,	hogy	118
ember	azt	mondja	,	hogy	117
kormány	azt	mondja	,	hogy	108
gábor	azt	mondta	,	hogy	104
istván	azt	mondta	,	hogy	102
viktor	azt	mondta	,	hogy	98
lászló	azt	mondta	,	hogy	97
péter	azt	mondta	,	hogy	97
ferenc	azt	mondta	,	hogy	91
túlzás	azt	mondani	,	hogy	86
...					

A második példán a 4. táblázatban látható gyakori mondat esetén csak a számot és személyt kell behelyettesíteni két esetben. Ha egy olyan „félig elemzett szerkezetet” tartunk készenlétben a memóriában, amiket csak ezekkel a kérdéses részekkel kell paraméterezni, a mondat többi elemzési lépését teljes egészében megspórolhatjuk a gyorsítótárazással, így növelve az elemzés sebességét.

A harmadik példánkban a 5. táblázatban látható, hogy a „gondolom , hogy” szerkezetek a mondatokban az esetek túlnyomó részében „azt”-al kezdődnek, habár a lehetséges névmások tárháza sokkal nagyobb, a korpusz a nyelvhasználat statisztikával ezt nem igazolja vissza. Ezzel elkerülhetjük, hogy az elemzőnkben fölöslegese eseteket is számba vegyünk, akkor ha azok „egyszerűbb elemzési heurisztikákkal” gyorsan megelemezhetők, kikerülve az elemzési „tévutakat” és a kombinatorikus robbanást.

Az utolsó példán a 6. táblázatban láthatjuk, hogy a triviális mintában a kötőszavak eloszlása mennyire eltérő, akkor ha NP-n belüli vagy pedig általánosan tekintett mintákról beszélünk. Ebből levonhatjuk azt a következtetést, hogy ha az NP elejét tudjuk detektálni, akkor „átállíthatjuk az agyunkat, egy másfajta elemzési módba”, ahol a triviális minták is másképpen viselkednek az NP-k végéig. Ezzel leszűkítve az állapotteret, gyorsítva az elemzést a tipikus szövegeken. Ez a viselkedés igazolni látszik az AnaGramma elemző elvét, miszerint a „különböző modulok elemzés közben hatással vannak egymásra, kommunikálnak”.

4. táblázat. *„lemma:köszön a lemma:figyelem .”* mintához tartozó szóalakok és azok előfordulási gyakorisága

Vizsgált minta				Gyakoriság
lemma:köszön	**a**	**lemma:figyelem**	**.**	**14582**
köszönöm	a	figyelmüket	.	7654
köszönöm	a	figyelmet	.	6762
köszönöm	a	figyelmét	.	142
köszönjük	a	figyelmüket	.	32
köszönjük	a	figyelmet	.	12
köszöni	a	figyelmüket	.	5
köszönöm	a	figyelmünket	.	3
köszönöm	a	figyelmeteket	.	2
köszöni	a	figyelmet	.	1
köszönjük	a	figyelmét	.	1

5. táblázat. *„[FN|NM][ACC] gondolom , hogy [HA]”* minta esetén az „[FN|NM][ACC]” címkéhez tartozó szavak és azok előfordulási gyakorisága

[FN|NM][ACC]	**7067**
azt	7056
ezt	8
.azt	1
amit	1
-azt	1

6. táblázat. *„[KOT] [DET] [MN][NOM] [FN][NOM]”* minta esetén a kötőszavak száma a teljes korpuszon, illetve az NP-k esetén

Általános szövegekben	gyakoriság	NP-ken belül	gyakoriság
hogy	185 102	és	6 236
és	27 556	illetve	489
mint	20 069	valamint	472
valami	17 791	azaz	33
de	16 480	és/vagy	30
illetve	13 072	avagy	27

6. Konklúzió

Munkánk során létrehoztunk egy olyan rendszert, amely szöveges korpuszból különböző faktorok kombinációinak segítségével képzett n-gramok gyakoriságából előállít mintákat. A mintákhoz lekérdezhetőek gyakorisággal együtt a teljesen kitöltött példák, amelyekre az adott minták illeszkednek. Ezen mintákból válogatott példákon bemutattuk, hogy a leendő elemző rendszer elemzéseit a minták gyorsítótárazásával gyorsítani tudjuk, az állapottér alkalmas leszűkítésével a tipikus szövegek esetében, ezzel utánozva az emberi agy gyorsaságát hasonló helyzetekben. Továbbá bemutattuk, hogy adott információk ismeretében, az elemző belső állapota átállítható egy specifikus almintacsoport elemzésére, ami nagyobb léptékekben (például ilyen a különböző doménbe tartozó esetleg roncsolt szövegek feldolgozása) könnyen megfigyelhető az emberi elemzőnél. Bár sok triviális minta is keletkezett, a létrejött minták számtalan ötletet adhatnak, a korpuszokon megfigyelhető jelenségek keresésére annak, aki hajlandó végigböngészni azokat.

Hivatkozások

1. Prószéky, G., Indig, B., Miháltz, M., Sass, B.: Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2014) 79–88
2. Grice, H.P.: Logic and conversation. na (1970)
3. Sass, B.: "Mazsola" - eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi, T., ed.: Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, Budapest, MTA Nyelvtudományi Intézet (2009) 117–129
4. Stolcke, A.: Srilm - an extensible language modeling toolkit. (2002) 901–904
5. Bilmes, J.A., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2, Association for Computational Linguistics (2003) 4–6
6. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The sketch engine. In: Proceedings of the Eleventh EURALEX International Congress. (2004) 105–116
7. Rychlỳ, P.: Manatee/bonito-a modular corpus manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing, within MU: Faculty of Informatics Further information (2007) 65–70
8. Novák, A.: What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
9. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In Sojka, P., Kopecek, I., Pala, K., eds.: Text, Speech and Dialogue. Volume 3206 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004) 41–47
10. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria (2013) 539–545
11. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. Ann Arbor MI **48113**(2) (1994) 161–175

12. Yang, Z.Gy., Laki, L.J., Prószéky, G.: Gépi fordítás minőségének becslése referencia nélküli módszerrel. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Egyetem (2015) 3–13
13. Zipf, G.K.: Human behavior and the principle of least effort. (1949)
14. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries **3**(2) (2000) 115–130

VII. Poszterek

Discovering Utterance Fragment Boundaries in Small Unsegmented Texts

László Drienkó

dri@t-online.hu

Abstract: We propose an algorithm for inferring boundaries of utterance fragments in relatively small unsegmented texts. The algorithm looks for subsequent largest chunks that occur at least twice in the text. Then adjacent fragments below an arbitrary length bound are merged. In our pilot experiment three types of English text were segmented: mother-child language from the CHILDES database, excerpts from *Gulliver's travels* by Jonathan Swift, and *Now We Are Six*, a children's poem by A. A. Milne. The results are interpreted in terms of four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability. We find that i) Inference Precision grows with merge-length, whereas Alignment Precision decreases – i.e. the longer a segment is the more probable that its two boundaries are correct; ii) Redundancy and Boundary Variability also decrease with the merge-length bound – i.e. the less boundaries we insert, the closer they are to the ideal boundaries.

1. Introduction

The problem of how to segment continuous speech into components dates back at least to Harris [2]. Harris used "successor frequencies", i.e. statistics, to predict boundaries between linguistic units. [8], using syllable-based artificial languages, demonstrated that statistical information is indeed available for infants acquiring language. Results in language acquisition research indicate that speech segmentation is affected by various lexical and sub-lexical linguistic cues (see e.g. [4]). Computational models of speech segmentation typically seek to identify the computational mechanisms underlying children's capacity to segment continuous speech (see [1] for a review). [6] outlines an integrated theory of language acquisition where the learner uses various cognitive heuristics to extract large chunks from the speech stream and the 'ultimate' units of language are formed by segmenting and fusioning the relevant chunks. The philosophy behind our boundary inference algorithm is, broadly speaking, similar in that we first identify "large" utterance fragments in unsegmented texts, i.e. character sequences, and then apply 'fusion' – 'merging', in our terminology – to see how precision changes.

Our heuristic for identifying utterance fragments is based on the assumption/intuition that recurring long sequences are more informative of segment boundaries than recurring shorter ones. Individual characters, for instance, can be followed

by, practically, any other character even in a short text. Reoccurring words or, notably, word combinations, on the other hand, are less likely to be followed/preceded by a word beginning/ending with the same letter as on the first occurrence. If a word combination still happens to be followed/preceded by a word beginning/ending with the same letter as on a previous occurrence, this reoccurring word can be considered as being part of an even larger word combination which, in turn, is less likely to reoccur in the same context. Thus we intuit that largest chunks represent sequence boundaries more reliably than shorter ones. Naturally, the ultimate largest chunk is the whole text itself with its two 100%-certain boundaries.

In our pilot investigation three types of English text were segmented: mother-child language from the CHILDES database, excerpts from *Gulliver's travels* by Jonathan Swift, and *Now We Are Six*, a children's poem by A. A. Milne. The texts are relatively short, with 60 - 7741 word tokens, 186 - 32,859 characters.

2. Description of the algorithm

The basic, CHUNKER, module of our algorithm looks for largest character sequences that occur more than once in the text. Starting from the first character, it concatenates the subsequent characters and if a resultant string s_i only occurs in the text once, a boundary is inserted before its last character in the original text since the previous string, s_{i-1}, is the largest of the i strings. Thus the first boundary corresponds to s_{i-1}, our first tentative speech fragment. The search for the next fragment continues from the position after the last character of s_{i-1}, and so on. As can be seen from our results, in terms of sequence length, the fragments output by this module broadly correspond to words.

The MERGE component of the algorithm concatenates fragments s_i and s_{i+1} if s_{i+1} consists of less than k characters. In other words, the boundary between s_i and s_{i+1} is deleted if s_{i+1} is shorter than k, an arbitrary length bound. In our experiments we had $1 \leq k \leq 11$.

The EVALUATE module computes four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability.

Inference Precision (IP) represents the proportion of correctly inferred boundaries (cib) to all inferred boundaries (aib), i.e. IP = cib / aib. The maximum value of IP is 1, even if more boundaries are inferred than all the correct (original) boundaries (acb).

Redundancy (R) is computed as the proportion of all the inferred boundaries to all the correct (original) boundaries, i.e. R = aib / acb. R is 1 if as many boundaries are inferred as there are boundaries in the original text, i.e. aib=acb, R is less than 1 if less boundaries are inferred than acb, and R is greater than 1 if more boundaries are inferred than optimal. Note that 1/R = acb\aib specifies how many words are grouped together on average in an inferred segment, i.e. the average fragment length in words.

Alignment Precision (AP) is specified as the proportion of correctly inferred boundaries to all the original boundaries, i.e. AP = cib / acb. Naturally, the maximum value for AP is 1.

Boundary Variability (BV) designates the average distance (in characters) of an inferred boundary from the nearest correct boundary, i.e. BV = (Σdf_i)/aib.

The above measures are not totally independent, since Inference Precision × Redundancy = Alignment Precision, but emphasise different aspects of the segmentation mechanism. Obviously, IP =AP for R=1.

3. The experiments

3.1. Experiment 1

In this experiment the first Anne file, *anne01a.xml*, of the Manchester corpus, [9], in the CHILDES database, [3], was investigated. The files were converted to simple text format, annotations were removed together with punctuation symbols and spaces. Mother and child utterances were not separated, so the dataset constituted an unsegmented (written) stream of 'mother-child language'. The original text consisted of 1815 word tokens and the average word length was 3.75 characters. The unsegmented version of the text consisted of 6801 characters. Initially, *k=1*, the CHUNKER module of our algorithm inserted 1129 boundaries, i.e. 1129 segments were identified with average segment length 6.02 characters. This means that the inferred fragments were, on average, 2.27 characters longer. The precision values were as follows: Inference Precision = 0.66, Redundancy = 0.62, Alignment Precision = 0.41, Boundary Variability = 0.53. In the second part of the experiment we let the merge-length bound *k* change from 2 to 11. For instance, *k = 3* means that, given the segmentation as provided by the CHUNKER module (the *k = 1* case, with no merging), fragment f_{i+1} is glued to the end of f_i if f_{i+1} consists of less than 3 characters, i.e. if f_{i+1} is one- or two-character-long. That is, the maximum merge-length is 2 for *k=3*. Figure 1 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 2 plots how the precision values change. For *k=11*, the values were IP=0.78, R=0.07, AP=0.05, BV= 0.3, and 121 boundaries were inserted.

3.2. Experiment 2

In this experiment the first part of Chapter 1 from *Gulliver's travels* by Jonathan Swift, [7], was investigated. The original text consisted of 1634 word tokens and the average word length was 4.05 characters. The unsegmented version of the text consisted of 6621 characters. The CHUNKER module inserted 1565 boundaries. The average segment length was 4.23 characters, which is quite close to the 4.05 average for the original text. The precision values were the following: IP = 0.5, R = 0.96, AP = 0.48, BV = 0.9. Figure 3 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 4 plots how the precision values change. For *k=11*, the values were IP=0.86, R=0.02, AP=0.02, BV= 0.17, and 30 boundaries were inserted.

3.3. Experiment 3

In this experiment Chapter 1 from *Gulliver's Travels* was investigated. The original text consisted of 4034 word tokens and the average word length was 4.17 characters. The unsegmented text consisted of 16,821 characters. The CHUNKER module inserted 3307 boundaries. The average segment length was 5.09 characters, about 1 character longer than the 4.17 value for the original text. The precision values were the following: IP = 0.53, R = 0.82, AP = 0.43, BV = 0.84. Figure 5 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 6 plots how the precision values change. For *k=11*, the values were IP=0.71, R=0.03, AP=0.02, BV= 0.35, and 133 boundaries were inserted.

3.4. Experiment 4

In this experiment Chapters 1 and 2 from *Gulliver's travels* were merged into a single text. The two chapters contained 7742 word tokens and the average word length was 4.24 characters. The unsegmented text consisted of 32,859 characters. The CHUNKER module inserted 5802 boundaries. The average segment length was 5.66 characters, about 1.5 characters longer than the 4.24 value for the original text. The precision values were the following: IP = 0.53, R = 0.75, AP = 0.4, BV = 0.8. Figure 7 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 8 plots how the precision values change. For *k=11*, the values were IP=0.75, R=0.04, AP=0.03, BV= 0.29, and 323 boundaries were inserted.

3.5. Experiment 5

Finally, we examined the children's poem *Now We Are Six* written by A. A. Milne, [5]. The poem consists of 60 word tokens and the average word length was 3.1. The unsegmented poem consists of 186 characters. The CHUNKER module inserted 79 boundaries, cf. Box 1. The average segment length was 2.35 characters, about 0.7 character shorter than the 3.1 value for the original text. The precision values were the following: IP = 0.45, R = 1.32, AP = 0.6, BV = 0.77. Figure 9 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 10 plots how the precision values change. For *k=11*, the values were IP=1 R=0.0167, AP=0.0167, BV= 0, i.e. the original sequence was restored with a 100% inference precision due to the single boundary inserted at the end of the text.

4. Discussion and conclusions

For all the texts that we looked at, the following pattern could be observed:

Inference Precision (the proportion of correctly inferred boundaries of all inferred boundaries) grows (45-66% to 70-100%) with maximum merge-length (0 to 10), whereas Alignment Precision (the proportion of correctly identified boundaries of all

the original, correct boundaries) decreases: i.e. the longer a segment is the more probable that its two boundaries are correct.

Redundancy (the proportion of all the inferred boundaries to all the correct boundaries) and Boundary Variability (the average distance from the closest correct boundary) also decrease with the merge-length bound: i.e. the less boundaries we insert, the closer they are to the ideal boundaries.

Our data suggest, most explicitly in Experiment 5, that, as the merge-length bound grows, Inference Precision approaches 1, Boundary Variability 0, Redundancy and Alignment Precision 1/n, where n is the number of word tokens in the original text.

The utterance fragments that our algorithm can detect are not necessarily individual words or syntactic phrases. The possible strengths of our method lie, on the one hand, in its potential to provide empirical insights into the statistical structure of natural language i) on the basis of small texts ii) without previous training corpora or iii) explicit probability values. On the other hand, the utterance fragments detected by our algorithm can serve as input for subsequent segmenting mechanisms to break down text into ultimate components, practically, into words.

Our results also suggest that looking for largest recurring chunks may be a powerful cognitive strategy. Statistically, the lengths of the fragments that our CHUNKER identified are quite close to the original word lengths. Note also, that all BV values were less than 1, which means that, for a given R value, a learner could obtain an optimal segmentation – i.e. where all inferred boundaries are correct – by shifting the inferred boundaries less than 1 character, on average, to the right or to the left. In other words, language learning could be based on memorizing tentative chunks that could be "finalised" later, as cognitive development progresses.

Figures and boxes

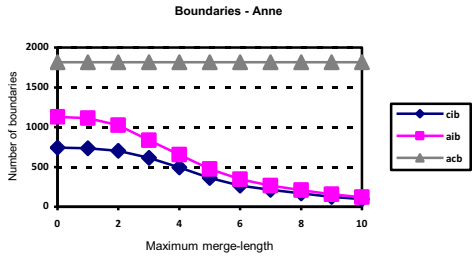


Fig. 1. Number of boundaries as function of maximum merge-length – Anne file from CHILDES (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

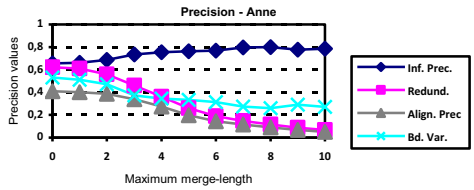


Fig. 2. Precision values changing with maximum merge-length – Anne file from CHILDES.

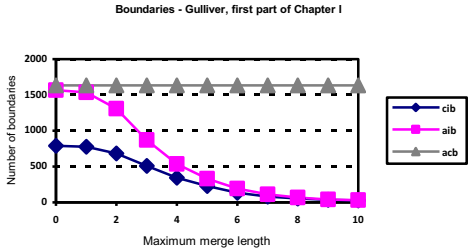


Fig. 3. Number of boundaries as function of maximum merge-length – first part of Chapter 1, *Gulliver's travels*. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

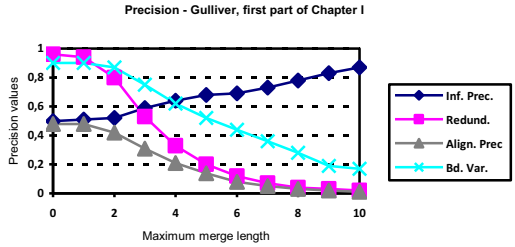


Fig. 4. Precision values changing with maximum merge-length – first part of Chapter 1, *Gulliver's travels*.

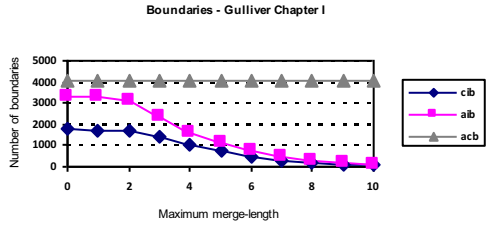


Fig. 5. Number of boundaries as function of maximum merge-length – *Gulliver's travels*, Chapter 1. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

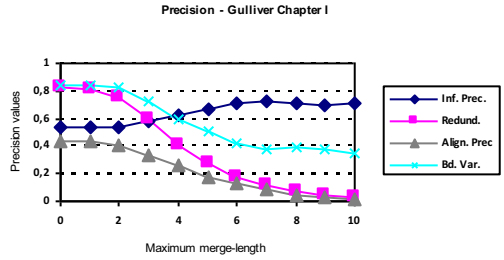


Fig. 6. Precision values changing with maximum merge-length – *Gulliver's travels*, Chapter 1.

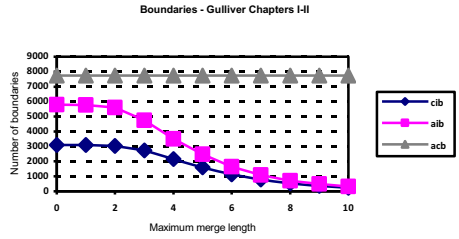


Fig. 7. Number of boundaries as function of maximum merge-length – *Gulliver's travels*, Chapters 1 and 2. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

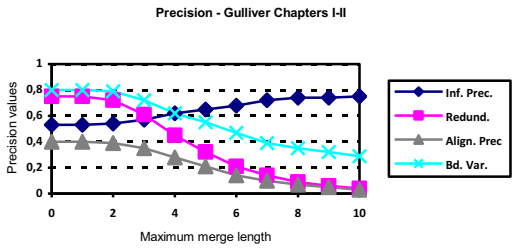


Fig. 8. Precision values changing with maximum merge-length – *Gulliver's travels*, Chapters 1and 2.

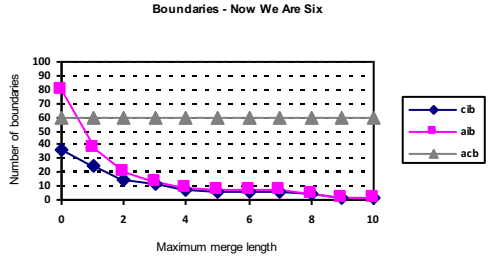


Fig. 9. Number of boundaries as function of maximum merge-length – *Now we are six.* (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

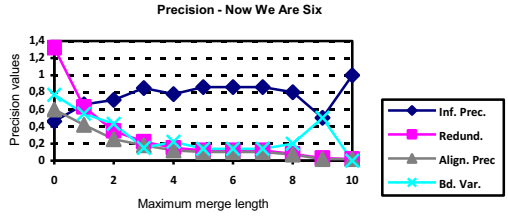


Fig. 10. Precision values changing with maximum merge-length – *Now we are six*.

Box 1

WHENIWAS:O:NE:I:HA:D:JUST:BE:G:U:N:WHENIWAST:W:OI:WASN:E:AR:LY:NE
:W:WHENIWAST:H:RE:EIWAS:HA:R:D:LY:M:EWHENIWASF:O:U:R:IWASN:O:T:M
:U:C:H:M:ORE:WHENIWASF:IVE:IWAS:JUST:A:L:IVE:B:U:T:NOW:I:A:M:SIX:I':M:
ASCLEVER:ASCLEVER:SO:I:TH:I:N:K:I':L:L:BE:SIX:NOW:A:N:D:FO:RE:VER:

References

1. Brent, M. R.: Speech segmentation and word discovery: a computational perspective. Trends Cognitive Sciences 3(8) (1999) 294–301
2. Harris, Z. S.: From phoneme to morpheme. Language 31 (1955) 190–222
3. MacWhinney, B.: The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates (2000)
4. Mattys, S. L, White, L., Melhorn, J.F.: Integration of multiple speech segmentation cues: a hierarchical framework. Journal of Experimental Psychology: General 134(4) (2005) 477–500
5. Milne, A.A. Now we are six (poem). Source: http://www.familyfriendpoems.com/poem/now-we-are-six-by-a-a-milne#ixzz3lkEs6IVU
6. Peters, A. (1983). The units of language acquisition. Cambridge, Cambridge University Press.
7. Swift, J.: Gulliver's Travels. The Project Gutenberg eBook. http://www.gutenberg.org/files/829/829-h/829-h.htm
8. Saffran, J. R., Aslin, R. N., Newport, E.L.: Statistical learning by 8-month-old infants. Science 274(5294) (1996) 1926–8
9. Theakston, A. L., Lieven, E. V., Pine, J. M., Rowland, C. F.: The role of performance limitations in the acquisition of verb-argument structure: an alternative account. J. Child Lang. 28(1) (2001) 127–52

Magyar nyelvű orvosi szakcikkek hivatkozásainak automatikus feldolgozása

Farkas Richárd[1], Kojedzinszky Tamás[1], Sliz-Nagy Alex[1],
Tímár György[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
rfarkas@inf.u-szeged.hu
[2] Comfit kft.
gyorgy.timar@comfit.hu

Kivonat: Cikkünkben bemutatunk egy szakirodalmi hivatkozások feldolgozására kidolgozott nyelvtechnológiai rendszert. A rendszer két legfontosabb modulja egy közelítő illesztésen alapuló visszakereső modul és egy szekvenciajelölő modul, ami a hivatkozások egyes elemeit azonosítja. Ez utóbbi megoldásnál egy újszerű kétlépcsős újrarangsoroló technikát is ismertetünk.

1. Bevezető

A magyar nyelvű orvosi szaklapok egy zárt rendszert alkotnak, a publikációk többsége nem érhető el publikusan, így azokat nem indexelik a sztenderd citációs adatbázisok (mint például Web of Science, Scopus vagy Google Scholar).

A Comfit kft. és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának közös projektjében azt a célt tűztük ki, hogy a cég magyar nyelvű orvosi szaklap adatbázisában (körülbelül 70 000 publikáció) szereplő hivatkozásokat automatikus eszközökkel feldolgozzuk, hogy az később alkalmas legyen tudománymetrikai mutatók számítására.

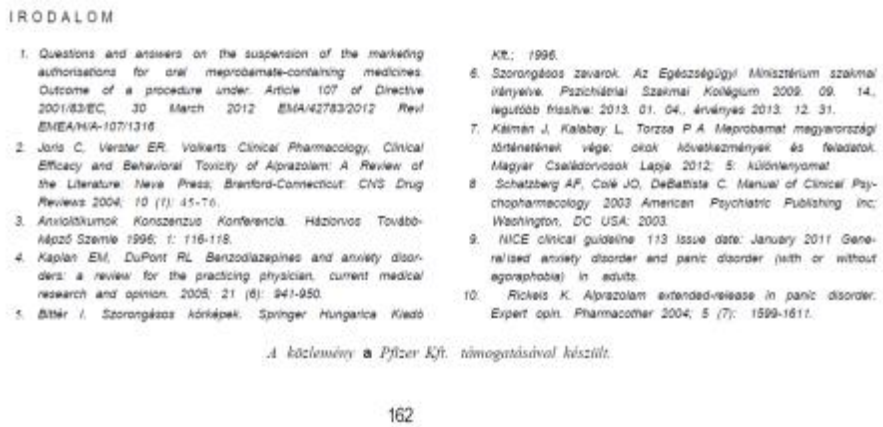
A feldolgozásra egy háromlépéses rendszert dolgoztunk ki. Először azonosítani kell a hivatkozásblokkokat a publikációkban és az egyes hivatkozásrekordokat szegmentálni kell. Ezután egy modul bejárja a rekordokat és megvizsgálja, hogy ismert publikációs adatbázisokban szerepel-e a hivatkozás. Végül a nem illesztett hivatkozásrekordokat elemezzük. Ehhez egy osztályozó eldönti, hogy újságcikkről van-e szó és ha igen, akkor a hivatkozás elemeit (szerzők nevei, cím, újság, évszám stb.) azonosítjuk. Ennek segítségével a meglévő publikációs adatbázis egy kézi jóváhagyás után gyorsan bővíthető.

2. Hivatkozásrekordok azonosítása

Első lépésben a pdf formátumban lévő újságcikkekből kellett a szöveges tartalmakat kinyerni. Itt komoly gondot okozott a többhasábos szerkesztés és a grafikonok, hirdetések nagy száma. A rendszer formázott szöveges bemenetét végül egy manuális szabályrendszer segítségével állítottuk elő. Ez a bemenet folyószöveges részeket tartalmazott. Az előfeldolgozás főbb lépései a hasábok azonosítása és azok összekötése valamint a sorvégek detektálása, sorvégi elválasztások helyreállítása.

A hivatkozás blokkok felismerésére ezután egy Conditional Random Fields alapú szekvenciajelölő módszert [1] fejlesztettünk ki, ami a szöveg egyes soraihoz REFERENCIA vagy NEMREFERENCIA címkét rendel. Ennek tanításához 150 cikkben kézzel bejelöltük a hivatkozásblokkokat. A rendszer egyes sorokat leíró jellemzőkészlete tartalmi (pl. az „irodalom" vagy „referencia" sztringeket tartalmazza), formai (pl. milyen hosszú, digitek és egyéb karakterek aránya), valamint környezeti (pl. a következő és megelőző 5 sorban hány évszám szerepel) jegyeket tartalmazott.

A hivatkozásblokkokat végül reguláris kifejezések és egyéb szabályok segítségével bontjuk rekordokra. Ez a rekordra bontás kiaknázza a hivatkozásblokkok azon tulajdonságát, hogy valamilyen módon sorszámozva vannak azok. Gyakran előfordul ugyanis, hogy egy sorszámmal kezdődik, de az csak a hivatkozásrekord része (pl. oldalszám), és nem egy új rekord sorszáma (lásd például a 1. ábrán).



IRODALOM

1. Questions and answers on the suspension of the marketing authorisations for oral meprobamate-containing medicines. Outcome of a procedure under Article 107 of Directive 2001/83/EC, 30 March 2012 EMA/42783/2012 Rev1 EMEA/H/A-107/1316
2. Joris C, Verster ER. Volkerts Clinical Pharmacology, Clinical Efficacy and Behavioral Toxicity of Alprazolam: A Review of the Literature. Neva Press; Branford-Connecticut: CNS Drug Reviews 2004; 10 (1): 45-76.
3. Anxiolitikumok Konszenzus Konferencia. Háziorvos Továbbképző Szemle 1996; 1: 116-119.
4. Kaplan EM, DuPont RL. Benzodiazepines and anxiety disorders: a review for the practicing physician. current medical research and opinion. 2005; 21 (6): 941-950.
5. Bitter I. Szorongásos kórképek. Springer Hungarica Kiadó Kft.; 1996.
6. Szorongásos zavarok. Az Egészségügyi Minisztérium szakmai irányelve. Pszichiátriai Szakmai Kollégium 2009. 09. 14., legutóbb frissítve: 2013. 01. 04., érvényes 2013. 12. 31.
7. Kálmán J, Kalabay L, Torzsa P A. Meprobamat magyarországi történetének vége: okok következmények és feladatok. Magyar Családorvosok Lapja 2012; 5: különlenyomat
8. Schatzberg AF, Cole JO, DeBattista C. Manual of Clinical Psychopharmacology 2003 American Psychiatric Publishing Inc; Washington, DC USA: 2003.
9. NICE clinical guideline 113 Issue date: January 2011 Generalised anxiety disorder and panic disorder (with or without agoraphobia) in adults.
10. Rickels K. Alprazolam extended-release in panic disorder. Expert opin. Pharmacother 2004; 5 (7): 1599-1611.

A közlemény a Pfizer Kft. támogatásával készült.

162

1. Ábra. Hivatkozási blokk a Ferencz Cs.: Hogyan tovább? Meprobamat után… Háziorvosi Továbbképző Szemle. 2013. 18 pp 160-162 cikkből.

3. Közelítő illesztések adatbázisban

Rendelkezésünkre állt a Medline adatbázis[1] 26 millió nemzetközi orvostudományi publikációs adatbázisa és a Comfit kft. magyar publikációs adatbázisa 70 ezer elemmel. Ezek az adatbázisok strukturált formában tartalmazzák a publikációk metaadatait (szerzők, cím, újság stb). Az adatbázisban történő pontos kereséshez számos közelítő heurisztika implementálására volt szükség, ugyanis a hivatkozások hemzsegnek az elgépelésektől, rövidítésektől és hibáktól.

A keresést a SolR rendszerben[2] implementáltuk. Ez hatékony közelítő keresést biztosított több tízmillió rekord felett is. Az illesztés elfogadására egy küszöbértéket határoztunk meg, ami a tokenszintű TF-IDF-el súlyozott koszinusz távolság és egy karakteralapú szerkesztési távolságból képzett aggregált hasonlósági mértékre vonatkozott. A két megközelítés együttes alkalmazására azért volt szükség, mert a tokenalapú metrika képes az egyes szavak (tipikusan a hivatkozás elemeinek) sorrendbeli különbségének kezelésére, míg a karakteralapú szerkesztési távolság képes kezelni az elírásokat, de az egyes tokenek sorrendjének felcserélését nem.

4. Kétlépcsős módszer tulajdonnév-felismerésre

Azokat a hivatkozásokat, amelyeket nem sikerült az adatbázisokban azonosítani, osztályoztuk újságcikk/könyv/könyvfejezet/URL/egyéb kategóriákba. A feladatra egy kézi szabályrendszert dolgoztunk ki, ami különböző reguláris kifejezéseken, valamint a leghasonlóbb adatbázisrekordból a szövegrészletre visszailleszthető mezők számosságán alapul.

Végül kidolgoztunk egy szekvenciajelölő algoritmust, ami az ismeretlen újságcikk hivatkozásrekordok egyes elemeit azonosítja (szerzők, cím, év, újság, oldalszámok). Ez a módszer lehetőséget biztosít az adatbázisokban nem szereplő, új hivatkozások összegyűjtésére és az adatbázis bővítésére. Ennek tanítására az ún. távoli felügyelet módszerét követtük, a sikeresen illesztett adatbázisrekordok egyes mezőit visszajelöltük az eredeti szövegrészletekre. Ez egy elég zajos, de nagyméretű (52 ezer rekord) tanító adatbázist eredményezett.

Maga a szekvenciajelölő módszer egy újszerű kétlépcsős megközelítés. Itt először egy tanított Maximum Entrópia Markov-modell [1] megadja a 100 legvalószínűbb szekvenciát, majd egy második felügyelt tanuláson alapuló újrarangsoroló lépés, frázis- és szekvenciaszintű jellemzők kiaknázásával, kiválasztja a legjobb szekvenciát. Ennek motivációja az, hogy a sztenderd szekvenciajelölők (MEMM, CRF stb.) tipikusan csak a lokális környezet leírására alkalmas jellemzőkkel dolgoznak, mint például a megelőző és a rákövetkező 3-4 token. De a jellemzőkészlet nem kódol az egész szekvencia jelölésére vonatkozó *nem lokális* információkat. A legegyszerűbb ilyen információ az lehet, hogy az egyes címkékből hány összefüggő címkesorozat predikálódott.

[1] *www.ncbi.nlm.nih.gov/pubmed*
[2] *http://lucene.apache.org/solr/*

A referenciaelemek azonosításánál például triviális megkötés, hogy legfeljebb *egy cikkcím* és egy *újságnév* kerülhet jelölésre. Egy ilyen jellegű megkötést nem lehet az egyszerű szekvencia jelölőkbe bevezetni, az csak speciális dekóder esetén lenne lehetséges, ami a keresési tér robbanásához vezetni.

Számos struktúrapredikciós probléma esetén bevált megoldás [2], hogy egy egyszerű(bb) és gyors első fázisa a rendszernek *n* darab lehetséges jó megoldást ad, majd egy második lépésben a lehetséges megoldásokat leírjuk nem lokális jellemzőkkel hiszen itt már rendelkezésre állnak az egész jelölésszekvenciára mint megoldásra vonatkozó információk is. Ezen jellemzők alapján újrarangsorolhatjuk az első fázis által megadott jelölteket. Az újrarangsorolás történhet a felügyelt tanulási paradigma keretein belül. Ekkor a tanító adatbázist lehetséges jelöltek és a legelőre rangsorolandó elem vagy elemek alkotják. Jelen rendszerben a 100 legvalószínűbb címkesorozatot írtuk le jellemzőkkel, majd a *max P(y|Y)* célfügvényre optimalizáló újrarangsoroló implementációt alkalmaztunk [3].

5. Eredmények

A rendszerrel a 2012 és 2013 alatt megjelent 13367 db magyar nyelvű orvosi szakcikkeket dolgoztunk fel. Ezek közel egynegyedében azonosítottunk hivatkozásblokkot, ami 66766 hivatkozásrekordot tartalmazott. Ezek közül 52353 rekordot tudtunk azonosítani a rendelkezésre álló adatbázisban. A fennmaradó rekordok közül az osztályozónk szerint 5621db elem van, amely az adatbázisban nem szereplő újságcikkre hivatkozik.

Az 52 ezer illesztett rekordon tanítottuk és kiértékeltük (80-20% arányban megbontva azt tanító és kiértékelő adatbázisra) a hivatkozáselem felismerő, kétlépcsős szekvenciajelölőnket. Ennek eredményeit az 1. táblázat foglalja össze. Megjegyezzük, hogy mivel a hivatkozásokban nagyon ritkán fordulnak elő egyik címkéhez sem tartozó tokenek, ezért az *O* címke szerepeltetése a kiértékelési metrikában életszerű.

1. táblázat. Címkékkénti eredmények újrarangsorolással.

	Pontosság	Fedés	F-mérték
Szerző	94.2933	99.0349	96.6059
Cím	95.0974	97.5388	96.3027
Újság	96.2294	95.1563	95.6898
Év	97.1175	99.5097	98.2991
Oldalszám_mettől	95.3211	99.2425	97.2423
Oldalszám:_meddig	92.9677	98.9209	95.8520
O	97.3818	94.3603	95.8473

6. Összegzés

Poszterünkön bemutattuk tudományos folyóiratok hivatkozásblokkjainak feldolgozására kialakított rendszerünket. A rendszer több modulból épül fel, amelyek rendre a számítógépes nyelvészet vívmányait aknázzák ki. Ez a rendszer akkor tud helyesen működni, ha rendelkezésre áll egy nagyméretű strukturált citációs adatbázis. Ennek felhasználásával – az ún. távoli felügyelet módszerét követve – építhetünk automatikusan annotált tanító adatbázist a gépi tanulási eljárásoknak.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning (4) (2012)
2. Farkas, R., Schmid, H.: Forest Reranking through Subtree Ranking. In: Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2012 (2012) 1038-1047
3. Charniak, E., Johnson, M. Coarse-tofine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL ’05 (2005) 173–180

Többsávos, zajtűrő beszédfelismerés mély neuronhálóval

Kovács György[1], Tóth László[2]

[1] KU Leuven, Department of Electrical Engineering
Leuven, Kasteelpark Arenberg 10, e-mail: gkovacs@esat.kuleuven.be
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: tothl@inf.u-szeged.hu

Kivonat Miközben az automatikus beszédfelismerés terén jelentős előrelépések történtek az elmúlt években, a beszédfelismerő rendszerek eredményessége spontán vagy zajjal szennyezett beszéd esetén továbbra sem kielégítő. Ezen probléma kiküszöbölésére számos módszert javasoltak, melyek közül több jól kiegészíti egymást. Jelen cikkünkben három ilyen módszer, az ARMA spektrogram, a neuronháló-tanítással egyidejűleg optimalizált spektro-temporális jellemzőkinyerés, és a többsávos feldolgozás kombinációját vizsgáljuk az Aurora-4 beszédfelismerési feladaton.[1]

Kulcsszavak: zajtűrő beszédfelismerés, mély neuronháló, többsávos feldolgozás, ARMA, Aurora-4

1. Bevezetés

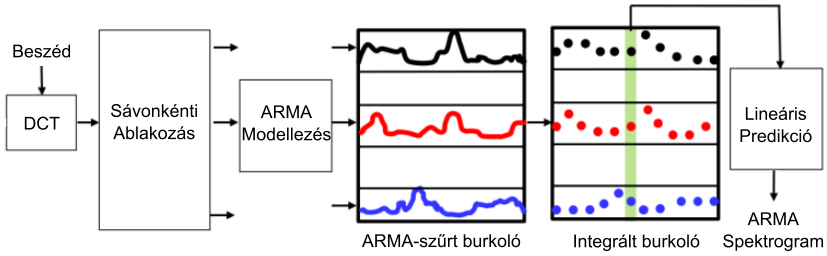
A beszédfelismerésben számos módszer létezik zajjal szennyezett beszéd pontosabb felismerésére. Jobb felismerési eredményeket érhetünk el például azáltal, ha a beszédnek olyan reprezentációját állítjuk elő, amely a hagyományos melskála szerinti sávszűrőknél (filter-bank) kevésbé érzékeny a beszédjelbe keveredő zajra. Ilyenre példa az itt használt ARMA spektrogram [1]. A felismerési pontosságot úgy is javíthatjuk, ha a spektrogramból zajtűrő jellemzőket próbálunk kinyerni. Ez motiválta a spektro-temporális jellemzőkinyerési módszerek, pl. Gábor-szűrők [2] bevezetését. Ezen módszerek egyik előnye, hogy a jellemzők előállításánál a spektrum korlátozott frekvenciatartományára támaszkodnak, így egy sávhatárolt zaj nem rontja el az összes jellemzőt. Hasonló megfontolásra épít a több-adatfolyamos (multi-stream) beszédfeldolgozás [3], amely a beszédjel különböző reprezentációit egymástól független taszkokban dolgozza fel, majd ezen taszkok eredményeit kombinálja. A többsávos (multi-band) beszédfeldolgozás a több adatfolyamos beszédfeldolgozás speciális esete, ahol az adatfolyamokat a beszédjel különböző frekvenciatartományaiból kinyert jellemzők jelentik. Jelen kísérleteinkben ezen megközelítéseket (zajtűrő reprezentáció használata, spektro-temporális jellemzők kinyerése, és többsávos feldolgozás) kombináljuk.

[1] A szerzők köszönetüket fejezik ki Sriram Ganapathy-nak az ARMA spektrogramokért, valamint Deepak Baby-nek az Aurora-4 használatában nyújtott segítségért.

2. Zajtűrő beszédfelismerési technikák

2.1. ARMA spektrogram

Ezen spektrogram előállítása az autoregresszív mozgóátlag (AutoRegressive Moving Average - ARMA) modell használatával történik. Ez a hagyományos AR modellezés [4] általánosítása. Az ARMA modellt a diszkrét koszinusz transzformáció (DCT) részsáv komponenseire alkalmazzák, a temporális burkológörbe becslésére. Az így kapott görbék rövid szakaszokon történő integrálásának eredménye egy spektrális reprezentáció. Az ARMA spektrogram ebből a spektrális reprezentációból lineáris predikción alapuló spektrális simítással áll elő. Ezt a folyamatot szemlélteti az 1. ábra. A cikkben felhasznált spektrogramokat Sriram Ganapathy (IBM T.J. Watson Research Center) bocsátotta rendelkezésünkre.



1. ábra. Az ARMA spektrogram előállítása (részletes leírásért, lásd [1]).

2.2. Spektro-temporális jellemzőkinyerés

A spektro-temporális jellemzőkinyerés során egy-egy jellemző előállításához a spektrális reprezentáció időben és frekvenciában korlátozott tartományát használjuk. Ennek egyik oka a megfigyelés, hogy az agykérgi neuronok a hangjel frekvencia- és időbeli (spektro-temporális) változásaira egyidejűleg reagálnak [5]. Így használatukkal az automatikus beszédfelismerésnél zajtűrőbb emberi hallás működéséhez közelíthetjük rendszerünket (ahogy azt tesszük a mel-skála használatával is). A spektro-temporális jellemzők másik előnye, hogy egy-egy jellemző a beszédjelnek csak egy korlátozott frekvenciatartományára támaszkodik, így egy sávhatárolt zaj jelenléte a beszédjelben nem érinti az összes jellemzőt.

2.3. Többsávos beszédfelismerés

A spektro-temporális feldolgozáshoz hasonlóan a többsávos beszédfelismerés esetében is az első lépés egymástól független akusztikus jellemzők kinyerése a különböző frekvenciasávokból. Ezért bizonyos esetekben a spektro-temporális beszédfelismerésre jellemzőrekombináció (feature recombination) néven, mint a többsávos feldolgozás egy speciális esetére hivatkoznak [6]. Ez is mutatja, hogy a többsávos beszédfelismerés módszerek széles skáláját fedi le. Az általunk követett módszer [7] esetében a különböző frekvenciatartományokból származó jellemzőkön különálló neuronhálókat tanítunk, majd ezen neuronhálók kimenetét egy kombinációs neuronháló bemeneteként használva végzünk újabb tanítást.

3. Mély neuronhálók

A mély neuronhálók jelentős előrelépést hoztak a beszédfelismerésben. Ezek olyan neuronhálók, melyek a hagyományosnál több rejtett réteget tartalmaznak. Ez nehézségekkel jár a háló tanítása során, melyek kiküszöbölésének érdekében változathatunk a tanítási algoritmuson, vagy a felhasznált neuronok típusán [8]. Cikkünkben ez utóbbi megoldás mellett döntöttünk, és neuronhálónkat rectifier ($max(0, x)$ függvényt megvalósító) aktivációs függvényt alkalmazó neuronokból (ún. ReLU-kból) építettük fel, miközben a tanítást továbbra is a hagyományos hibavisszaterjesztési (backpropagation) algoritmussal végezzük.

3.1. Együttes neuronháló-tanítás és jellemzőkinyerés

A spektro-temporális jellemzőkinyerést végrehajtó szűrők megvalósíthatók speciális, lineáris aktivációs függvényt alkalmazó neuronok formájában [9]. Könnyen belátható, hogy a lineáris neuronok alkalmasak erre, ha összehasonlítjuk a lineáris neuronok kimenetét és a szűrés eredményét meghatározó egyenleteket

$$\begin{aligned} o &= \left(\sum_{i=1}^{L} x_i \cdot w_i + b \right), \\ o &= \sum_{f=0}^{N} \sum_{t=0}^{M} A(f,t) F(f,t), \end{aligned} \tag{1}$$

ahol x a neuron bemeneti vektora, w a neuron súlyvektora, amelyek a megfelelő indexeléssel (továbbá feltételezve, hogy $L = N \cdot M$) megfeleltethetők A és F változóknak, ahol A az ablak amire az F szűrőt alkalmazzuk. Ekkor a b biast nullának választva, a két egyenlet egymás megfelelője. Erre az alapötletre építve mutattuk meg korábban, hogy a szűrők paramétereinek optimalizálása és a neuronháló betanítása egyetlen közös optimalizálási lépésként is elvégezhető [9]. Jelen cikkben is ezt a módszert alkalmazzuk.

3.2. Konvolúció

Az időbeli konvolúció egy megoldást ad arra, hogy a spektro-temporális ablakok méreténél hosszabb időintervallumból tudjunk információt kinyerni az ablakok megnövelése (és így további neuronháló-súlyok) bevezetése nélkül. Esetünkben ennek megvalósítása oly módon történik, hogy a lokális jellemzőket a neuronháló több, egymást követő időpontban nyeri ki, az egymást követő ablakokra azonos súlyokat alkalmazva, majd az eredmények közül csak minden negyediket ad át a következő rétegnek. Ezzel a konvolúció mindhárom követelménye, a lokális ablakok használata (local windows), a súlyok megosztása (weight sharing), és eredmények csoportjainak egyetlen értékkel történő reprezentációja (pooling) teljesül [10]. Arra vonatkozóan, hogy mely ablakokból származó eredményeket adjuk tovább, és mely ablakok eredményét nem vesszük a későbbiekben figyelembe, korábban több kísérletet végeztünk [11]. Az általunk használt neuronháló architektúra részletesebb leírása is megtalálható ebben a munkában.

4. Kísérleti beállítások

4.1. Előfeldolgozás

A bemenő beszédjelből először egy 39 csatornás ARMA spektrogramot számoltunk. A kapott ARMA spektrogramokat bemondásonként normalizáltuk oly módon, hogy átlaguk nulla, varianciájuk pedig egy legyen. Továbbá Ganapathy nyomán [1] derivatív jellemzőket számoltunk a spektrogramból, minden csatornára a szomszédos csatornák alapján ($band(K+1) - band(K-1)$). Így minden időpillanathoz egy 78 (39·2) komponensű jellemzővektort rendeltünk.

4.2. Aurora-4

Az Aurora-4 a Wall Street Journal beszédadatbázis zajjal szennyezett változata [12]. Elérhető (az általunk használt) 16 kiloherzes, valamint 8 kiloherzes mintavételezéssel. Az adatbázis két 7138 mondatból álló tanító halmazt, és 14, egyenként 330 mondatból álló teszthalmazt tartalmaz. A tanító halmaz első (tiszta) változata a mondatok zaj nélküli változatát tartalmazza Sennheiser mikrofonnal rögzítve, míg a második változatban az egyes mondatok különböző zajokkal szennyezettek, illetve rögzítésük eltérő mikrofonnal történt. Mivel cikkünkben azt kívántuk megvizsgálni, hogy a beszédfelismerő rendszer hogy teljesít általa nem ismert zajtípusok (illetve átviteli karakterisztika) esetén, ezért csak a tiszta tanító halmazt használtuk. A tanító halmaz kilencven százalékán tanítottuk a neuronhálók súlyait, míg a fennmaradó részt megállási feltételként (és validációs halmazként) használtuk.

A kiértékelést mind a 14 teszthalmaz felhasználásával végeztük. Ezen teszthalmazok ugyanazt a 330 mondatot tartalmazzák különböző verziókban: az első hét teszthalmazban lévő hangfájlok rögzítése a Sennheiser mikrofonnal történt, míg a második hét teszthalmazban ettől eltérő mikrofonnal rögzített felvételeket találunk. Mindkét (különböző mikrofonnal felvett) hetes csoport belső felosztása azonos: az első halmaz zajjal nem szennyezett beszédet tartalmaz, míg a következő hatban hat különböző zajjal (autó, csevegés, étterem, utca, repülőtér, vonat) szennyezett beszéd található.

4.3. Kaldi

Beszédfelismerési kísérleteinkben a Kaldi [13] Aurora-4 receptjéből indultunk ki, melyet a saját neuronhálónkkkal kombináltunk. A Kaldi recept egy HMM/GMM-et tanít be, majd kényszerített illesztést végez, ami minden adatvektorhoz az 1997 kontextusfüggő trifón állapot valamelyikét rendeli. A kombinációs neuronhálókat úgy tanítottuk be, hogy valószínűségeket biztosítsanak minden kerethez. A neuronháló kiértékelése után ezeket a valószínűségeket a Kaldi dekóderének inputjaként használtuk. A dekódolás során emellett egy trigram modellt, valamint egy 5000 szót tartalmazó szótárat használtunk.

4.4. Korábbi módszereink adaptálása

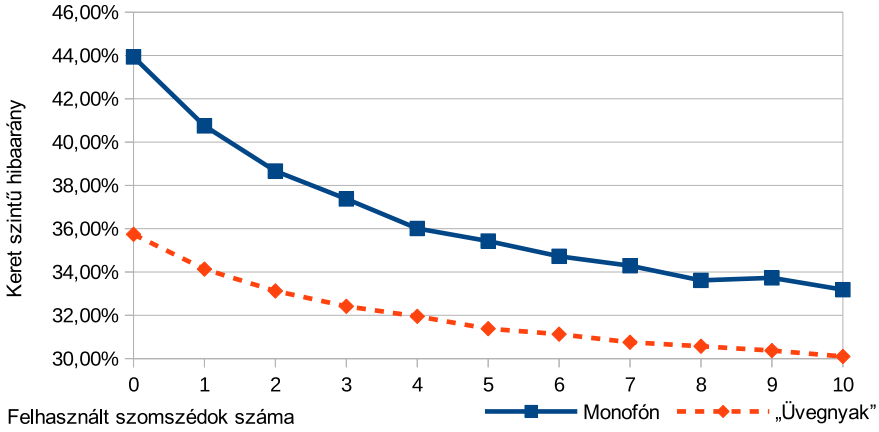
A korábbi munkáinkban felhasznált módszereket [7,11] jelen munkában adaptálnunk kellett az új spektrogramhoz, valamint adatbázishoz. A felhasznált architektúrát bemutató korábbi munkánkban [11] egy tükrözéssel 30 sávosra bővített 26 sávos spektrogramot használtunk a háló bemeneteként, így az ott leírt ablakméret és átfedés használatával hat, a szűrők alkalmazását megvalósító rétegre volt szükségünk a neuronhálóban. Jelen munkában ezen rétegek száma az ablakméret és átfedés megtartásával (de a tükrözés elhagyásával) 16-ra bővült (8-8 az eredeti spektrogramon, és annak derivatív változatán). Mivel az adatbázison végzett korábbi kísérleteinkben a szűrők irányított inicializálása nem javította a felismerést, jelen kísérleteinkben a szűrőket véletlen számokkal inicializáltuk. Adaptálnunk kellett továbbá a többsávos beszédfeldolgozás módszerét is: korábbi cikkünkben [7] a spektrogramot (a tükrözést nem számítva) 5 sávra bontottuk az ablakozásnak megfelelően. Jelen esetben mind a spektrogramból, mind a derivatív változatból egy-egy időszakaszon 8-8 ponton nyerünk ki ablakokat, így legjobban úgy tudtuk közelíteni a korábban használt paramétereket, hogy mind a kettőből 2-2 ablakot foglaltunk egy sávba, bemenetünket 4 sávra osztva.

További eltérés korábbi, a TIMIT beszédadatbázison végzett kísérleteinkhez képest, hogy az Aurora-4 esetén 42 monofón címkével ellátott címkézéssel rendelkeztünk, így a különböző sávokon tanított neuronhálók 42 osztálycímkére tanultak. (A kísérletek során ezen hálók kombinációjára „monofón" néven hivatkozunk). Mivel azonban a kombinációs hálók tanítása jelen kísérletek során 1997 állapotra történt, problémát jelenthet, hogy az eltérő szinteken lévő neuronhálók eltérő célfüggvényekkel tanulnak. Így további kísérleteket végeztünk, melyek során a különböző sávokon tanított neuronhálókat szintén 1997 állapotra tanítjuk. Ily módon azonban ha továbbra is ezen neuronhálók kimeneteit használnánk a kombinációs háló bemeneteként, jelentős dimenziószám-növekedéssel kéne szembenéznünk (168 helyett közel 8000 jellemző), ezért az alsó rétegben tanított neuronhálókba a kimeneti réteg elé egy második (50 lineáris aktivációs függvényt alkalmazó neuronból álló) „üvegnyak" (bottleneck) réteget illesztettünk, és ezen rétegek kimenetét használtuk a kombinációs háló bemeneteként. (A kísérletek során ezen hálók kombinációjára „üvegnyak" néven hivatkozunk).

4.5. Neuronháló paraméterek

Kísérleteinkben kétfajta neuronhálót használtunk: egyet, mely a különböző sávokon tanult (illetve összehasonlítási alapként ezeknek egy olyan változatát, mely az összes sávon tanult), valamint az előbbi hálók kimenetét bemenetként használó kombinációs hálókat. Ez utóbbiak ReLu hálók két, egyenként ezer neuronból álló rejtett réteggel, és 1997 kimeneti neuronnal. A különböző sávokon tanított hálók architektúrja a korábban leírtaknak megfelelő[2]. Az összehasonlítási alapként, az összes sávon egyszerre tanított (a kísérletek során jellemzőrekombinációsként

[2] A 4.4 szekcióban leírt módosításoktól eltekintve a hálók megegyeznek a korábbi cikkünkben [11] utolsóként leírt hálóval.



2. ábra. Keretszintű hibaarányok a validációs halmazon, a kombinációs háló által használt szomszédos keretek számának függvényében.

hivatkozott módszerhez tartozó) háló annyiban tér el ettől, hogy 4 helyett 16 szűrő réteggel rendelkezik, második bottleneck rétege 200 neuront tartalmaz, és minden köztes réteg kétszer annyi neuronból áll, mint az egyes sávokon külön tanított társaié (így a két változat paraméterszáma közel azonos).

5. Kísérletek és eredmények

Első lépésben a többsávos módszerek felismerési eredményeit hasonlítottuk össze, a kombinációs hálóban felhasznált szomszédok szempontjából. Ehhez betanítottuk a sávokat feldolgozó hálókat, valamint ezek kimenetén 3-3 kombinációs hálót minden konfiguráció esetére. A validációs halmazon kapott keretszintű hibaarányok leolvashatók a 2. ábráról. Megfigyelhető, hogy az „üvegnyak” módszerrel kapott eredmények konzisztensen felülmúlják a monofón módszer eredményeit. Bár a keretszintű eredmények a tizedik szomszéd felhasználásával is javulnak, mivel a görbe ellaposodik, valamint figyelembe véve azt a megfigyelést, hogy a kontextus kiterjesztése a keretszintű eredményeken akkor is javíthat, amikor a szószintű eredményekre már negatív hatással van [14], a teszteket a négy szomszédos keretet felhasználó változaton végeztük. (A monofón verzión végzett előzetes teszteredményeink igazolták a feltételezésünket, hogy négynél több szomszéd alkalmazása nem javítja tovább a szófelismerési pontosságot).

Hogy a szószintű felismerés hatékonyságát is elbírálhassuk, a kiválasztott, négy szomszédot használó hálókat kiértékeltük a teszthalmazokon szószintű hibákra. Ezt tettük a jellemzőrekombinációs háló kimenetén tanított kombinációs hálóval is (amely az összehasonlíthatóság miatt szintén négy szomszédot használt). Az összehasonlítás teljességéért hozzáadtuk a táblázathoz Sriram Ganapathy [1] eredményeit is (aki szintén az ARMA spektrogramot használva, de

1. táblázat. Szófelismerési-hibaszázalékok az Aurora-4 teszthalmazain.

Mikrofon	Zaj	Jellemzőrekombináció	Monofón	Üvegnyak	Ganapathy[1]
Azonos mikrofon	Tiszta	3,9%	3,8%	3,7%	**3,0%**
	Autó	6,4%	6,1%	5,8%	**5,0%**
	Csevegés	13,6%	13,9%	**12,6%**	13,0%
	Étterem	17,7%	18,8%	**17,2%**	17,3%
	Utca	**14,0%**	15,5%	**13,7%**	**13,6%**
	Repülőtér	14,0%	14,5%	**13,5%**	**13,7%**
	Vonat	**14,6%**	17,1%	**14,5%**	**14,5%**
	Átlag	12,0%	12,8%	**11,6%**	**11,4%**
Eltérő mikrofon	Tiszta	14,3%	**11,9%**	**11,6%**	**11,7%**
	Autó	21,7%	18,6%	**17,7%**	18,4%
	Csevegés	30,1%	29,8%	**27,5%**	29,6%
	Étterem	30,6%	31,9%	**29,6%**	31,1%
	Utca	29,8%	30,9%	**27,8%**	28,3%
	Repülőtér	29,4%	29,4%	**27,3%**	29,5%
	Vonat	30,3%	31,0%	**27,3%**	29,1%
	Átlag	26,6%	26,2%	**24,1%**	25,4%
Átlag		19,3%	19,5%	**17,8%**	18,5%

eltérő jellemzőkkel, és eltérő architektúrájú neuronhálóval dolgozott), melyek az általunk ismert legjobb olyan publikált eredmények az Aurora-4 adatbázison, melyeket a tiszta tanító halmaz felhasználásával értek el. A kapott eredmények kiolvashatók az 1. táblázatból.

Az „üvegnyak" és monofón oszlopokat összehasonlítva azt látjuk, hogy a validációs halmazon mutatott különbség a szószintű felismerésben is megjelenik. Nem csupán az átlagolt felismerési eredményben múlja felül az „üvegnyak" a monofón módszert, de (három kivétellel) minden teszthalmazon szignifikánsan jobb annál. Érdekesebb megfigyeléseket tehetünk a jellemzőkombináció összehasonlításával az „üvegnyak", illetve monofón módszerekkel. Itt egy kettősséget fedezhetünk fel, a Sennheiser mikrofonnal készült teszthalmazok és a többi teszthalmaz között. Míg az első esetén az „üvegnyak" módszer eredménye alig múlja felül a jellemzőkombinációs módszerét (miközben a monofón módszer alatta is marad), a második esetben a monofón módszer is egy kevéssel túlteljesíti a jellemzőkombinációs változatot, míg az „üvegnyak" módszer előnye jóval jelentősebb (9,3 százalékos hibaarány-csökkenés, szemben a korábbi 3,7 százalékkal). Úgy tűnik tehát, hogy miközben a többsávos módszerrel javult a felismerés eredményessége additív zaj jelenlétében (ez akár szignifikáns különbséget is jelenthet), igazán akkor profitáltunk belőle, amikor a beszédfelismerés a tanítóhalmaztól eltérő mikrofonnal rögzített beszéden történik. Ezt láthatjuk akkor is, ha az „üvegnyak" módszer eredményeit Ganapathy eredményeivel vetjük össze: miközben a Sennheiser mikrofonnal rögzített teszthalmazok esetén nem látunk érdemben jobb felismerési eredményeket az „üvegnyak" módszernél (sőt, néhány esetben rosszabbul is teljesít), a többi teszthalmazon jelentősen jobb eredményeket kaphatunk a használatával.

6. Konklúzió és jövőbeni tervek

Kísérleteink azt mutatták, hogy korábban bemutatott módszereink nem csak sikeresen kombinálhatók, de új környezetben is alkalmazhatók. Az is világossá vált, mennyire fontos a módszerek megfelelő adaptációja (az „üvegnyak" bevezetése előtt nem tudtunk javulást elérni a felismerési eredményekben). Ismételten megtapasztaltuk továbbá, hogy mennyire körülményes lehet az eltérő sávokat feldolgozó neuronhálók eredményeinek egyesítése. Így a jövőben megpróbálunk olyan módszert találni, amely a jelenleg két lépéses módszert egyszerűsítheti.

Hivatkozások

1. Ganapathy, S.: Robust speech processing using ARMA spectrogram models. In: Proceedings of ICASSP. (2015) 5029–5033
2. Kleinschmidt, M., Gelbart, D.: Improving word accuracy with Gabor feature extraction. In: Proceedings of ICSLP. (2002) 25–28
3. Hermansky, H., Timbrewala, S., Pavel, M.: Towards ASR on partially corrupted speech. In: Proceedings of ICSLP. (1996) 464–465
4. Atal, B.S., L, H.S.: Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America **50**(2) (1971) 637–655
5. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America **118**(2) (2005) 887–906
6. Okawa, S., Bocchieri, E., Potamianos, A.: Multi-band speech recognition in noisy environments. In: Proceedings of ICASSP. (1998) 644–644
7. Kovács, Gy., Tóth, L., Grósz, T.: Robust multi-band ASR using deep neural nets and spectro-temporal features. In: Proceedings of SPECOM. (2014) 386–393
8. Grósz, T., Kovács, Gy., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY. (2014) 3–13
9. Kovács, Gy., Tóth, L.: The joint optimization of spectro-temporal features and neural net classifiers. In: Proceedings of TSD, Springer (2013) 552–559
10. Veselý, K., Karafiát, M., Grézl, F.: Convolutive bottleneck network features for LVCSR. In: Proceedings of ASRU. (2011) 42 – 47
11. Kovács, Gy., Tóth, L.: Joint optimization of spectro-temporal features and deep neural nets for robust automatic speech recognition. Acta Cybernetica **22**(1) (2015) 117–134
12. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). (2000) 29–32
13. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Veselý, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. (2011)
14. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.: Reverberation robust acoustic modeling using with time delay neural networks. In: Proceedings of Interspeech. (2015) 3214 – 3218

Statisztikai koreferenciafeloldó rendszer magyar nyelvre – első eredmények

Munkácsy Gergely, Farkas Richárd

Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
rfarkas@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatjuk az első statisztikai (gépi tanulás alapú) módszert koreferenciafeloldásra magyar nyelvű szövegekben. Ehhez a SzegedKoref korpuszon [1] tanítottuk a HOTCoref rendszert [2], majd a rendszer egyes moduljait alakítottuk át a magyar korpusznak megfelelően.

1. Bevezetés

A természetes nyelvű szövegekre jellemző, hogy kevés bennük az ismétlés, a szerzők törekednek a változatosságra, többféle kifejezést használnak ugyanarra az entitásra hivatkozásnál, pl. *kutya*, *eb*. Ezeknek a szövegeknek a megértéséhez szükség van arra, hogy tudjuk, hogy a kifejezések melyik másik kifejezésre utalnak, vagy épp melyek azok a szövegrészek, melyek azonos egyedre utalnak. Ez a koreferencafeloldás feladata.

Cikkünkben bemutatjuk az első statisztikai (gépitanulás-alapú) módszert koreferenciafeloldásra magyar nyelvű szövegekben. Ehhez a SzegedKoref korpuszon [1] tanítottuk a HOTCoref rendszert [2], majd a rendszer egyes moduljait alakítottuk át a magyar korpusznak megfelelően.

2. Korpusz

A SzegedKoref korpusz [1] egy magyar nyelvű, teljes mértékben kézzel annotált koreferenciakorpusz, mely azzal a céllal készült, hogy alapjául szolgáljon különböző statisztikai (adatvezérelt) algoritmusok tanításának és kiértékelésének. Az annotálás alapja a Szeged Korpusz volt, melyet egy újabb réteggel bővítettek. Ezek közül is azokat a szövegeken választották, amik viszonylag hosszabbak, az egy-két mondatos dokumentumokon kevésbé érdekes a koreferenciafeloldási feladat.

A SzegedKoref folyamatosan bővül, munkánkhoz a 2015 eleji változatot használtuk fel, mely két részből állt össze:

- Iskolai fogalmazások, elbeszélések
- Újsághírek

Az adatbázisból az egyszerűség kedvéért kísérleteinkhez töröltük a zéró névmásokat. Az így létrejövő adatbázisban lévő 400 dokumentum összesen 9 565 mondatot és 123 971 tokent tartalmaz. Ezekből 18 854 szerepel koreferencialáncban.

3. Koreferenciafeloldó rendszer

Munkánk alapjául a stuttgarti egyetemen fejlesztett HOTCoref (Higher Order Tree Coreference) program [2] szolgált, mely a CoNLL (Conference on Computational Natural Language Learning) Shared Task [3] adatain a legjobb eredményeket adja jelenleg. A HOTCoref első lépésben szabályok alapján kiválasztja a lehetséges anaforajelölteket, majd felügyelt gépi tanulási módszertant követve alakítja az említési láncokat, azaz azokat az említéscsoportokat amelyek egy entitásra vonatkoznak. A gépi tanulási megközelítés az egyes csoportokat látens faszerkezettel reprezentálja.

A HOTCoref magyar nyelvhez igazításának első lépése a jelöltek azonosítására szolgáló algoritmus honosítása volt. Ehhez statisztikákat gyűjtöttünk a tanító adatbázisban előforduló anaforákról (szófajok és konstituens-elemzésbeli nem terminális címkék).

A gépi tanulási rész itt is a jól megválasztott jellemzőkészleten áll vagy bukik. Ennek magyarra átalakításához a Szeged Treebank morfológiai és szintaktikai leírói alapján átírtuk a tulajdonnév, határozottság, számosságra utaló jegyeket. Egy másik fontos átalakítás a magyar ún. headFinder szabályok implementálása, mely a kifejezések fejét keresi meg. Ez egy olyan szabályrendszer, amely megadja, hogy ha egy utalás több szóból áll, akkor abból melyik a legfontosabb. Ezt a mondat konstituenselemzése alapján döntöttük el. Például a *nagy piros labda* kifejezés esetén a *labda* token a fej. Végül töröltük a magyarban nem értelmezhető jellemzőket (pl. gender).

Megvizsgáltuk továbbá, hogy a ragozott alakok hordoznak-e hasznos információt a koreferenciafeloldási feladatban. Azt tapasztaltuk, hogy az eredmények drasztikusan (átlagos 15 százalékponttal) emelkedtek, ha szótöveket használunk a lexikai jellemzők kinyerésekor. Ennek magyarázata kettős. Egyrészt a tanító adatbázis viszonylag kicsi, nagyon alacsony az egyes ragozott alakok gyakorisága, ami a jellemzők extrém ritkaságát vonja maga után. Másrészt az említések összerendelése elsősorban szemantikai, és nem morfoszintaktikai alapon dönthető el.

4. Eredmények

Ellentétben más problémáknál használt leszámolós módszerekkel, egy koreferenciajelölés pontosságának értékelése vitatott feladat. Sokfajta metrika létezik, melyek különbözően jellemzik az egyes mintákat. Viszont az, hogy melyiket érdemes használni, az nem nyilvánvaló. A nemzetközi trendet követve négy különböző metrika (MUC, BCUC, CEAFM, CEAFE) mellett is kiértékeltük[1] a koreferenciafeloldót [3].

[1] a kiértékelő szkript elérhető: http://conll.cemantix.org/2011/software.html

Az átalakításokkal az alábbi eredményeket éri el a rendszer tökéletes (gold standard) morfoszintaktikai jelölések mellett:

	Fedés	Pontosság	F1
MUC	35,69	49	41,3
BCUB	34,21	49,09	40,32
CEAFM	40,47	54,45	46,43
CEAFE	39,11	50,72	44,16
Átlag	37,37	50,815	**43,0525**

Összehasonlításként egy ugyanekkora tanító adatbázist használó angol rendszer eredményei:

	Fedés	Pontosság	F1
MUC	64,85	64,43	64,64
BCUB	50,44	52,39	51,4
CEAFM	54,98	54,94	54,96
CEAFE	47,38	48,21	47,79
Átlag	54,4125	54,9925	**54,6975**

5. Összegzés

Ezek az eredmények első, de megismételhető empirikus eredmények. Az átalakított HOTCoref rendszert kérésre bárkinek odaadjuk. Számos ponton javítható még a rendszer a jövőben, például szemantikai információk (WordNet, tulajdonnév-kategorizáció stb.) beépítésével.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. Farkas R., Vincze V., Hegedűs K.: SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2015) 312–319
2. Björkelund, A., Kuhn, J. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2014) 47–57

Angol-magyar többszavas kifejezések szótárának automatikus építése párhuzamos korpuszok segítségével

Nagy T. István[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail: nistvan@inf.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat Jelen tanulmányunkban bemutatjuk gépi tanulási módszeren alapuló megközelítésünket, melynek segítségével félig kompozicionális szerkezetek (FX) fordításait tudjuk automatikusan megadni. A feladat nehézségét többek közt az adja, hogy a félig kompozicionális szerkezetek jelentése nem teljesen kompozicionális, vagyis azok elemeinek egyenkénti fordításával nem, vagy csak nagyon ritkán kapjuk meg az aktuális szerkezet idegen nyelvű megfelelőjét. A probléma megoldásához a SzegedParallelFX korpuszon a már korábban manuálisan annotált FX-ket kézzel megfeleltettük egymásnak. Az így létrejött korpuszon bináris osztályozó segítségével automatikusan választottuk ki a megfelelő magyar és angol nyelvű FX-párokat.
Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, szintaktikai elemzés

1. Bevezetés

A félig kompozicionális szerkezetek (FX-ek) olyan többszavas kifejezések, melyek egy főnévből és egy igéből állnak, ahol jellemzően a főnév a szemantikai fej, míg az ige csupán a szerkezet igeiségért felelős, mint például *figyelembe vesz* („take into account") vagy *támogatást nyújt* („grant support"). Korábbi munkáink során már több olyan korpuszt is bemutattunk, ahol FX-ek manuálisan vannak jelölve. Ezek közül több esetben párhuzamos korpuszokat hoztunk létre [1,2], ahol az FX-ek különböző nyelven is jelölve vannak. Ezen korpuszokon lehetőségünk nyílt olyan adatvezérelt gépi tanuló megközelítések [3,4] megvalósítására, melyek automatikusan képesek félig kompozicionális szerkezeteket azonosítani folyó szövegekben különböző nyelveken.

Mivel a félig kompozicionális szerkezetek jelentése nem teljesen kompozicionális, ezért azok elemeinek egyenkénti fordításával nem, vagy csak nagyon ritkán kapjuk meg az aktuális szerkezet idegen nyelvű megfelelőjét. Ezen szerkezetek viszonylag gyakoriak a nyelvekben, ezért megfelelő fordításuk elengedhetetlen, ám mivel szintaktikai, lexikai, szemantikai, pragmatikai vagy statisztikai szempontból idioszinkratikus tulajdonságokkal bírnak, ezért ezen szerkezetek idegen

nyelvi megfelelőinek automatikus megadása meglehetősen nehéz feladat [5]. Ezért jelen kutatásunk során kísérletet teszünk arra, hogy létrehozzunk egy olyan gépi tanuló megközelítést, mely képes a párhuzamos korpuszokon különböző nyelveken előforduló FX-ek automatikus megfeleltetésére. Ehhez a SzegedParallelFX [6] korpuszon, ahol a folyó szövegekben előforduló FX-ek már manuálisan annotálva vannak angol és magyar nyelven, manuálisan jelöltük az egy fordítási egységen belül előforduló FX-ek fordítási megfelelőit. Így például a következő fordítási egységben

> Látták, hogy bemászik az ablakon, úgyhogy nem lehetett titokban tartani.
> She was seen climbing through the window, so it couldn't be kept a secret.

a *titokban tartani* és *climbing through the window* egy negatív, míg a *titokban tartani* és *kept a secret* pedig egy pozitív példát jelöl.

Jelen munkában elsősorban az egyes FX-ek idegennyelvű FX-megfelelőit kerestük, ezért az annotálás során csupán FX-ket feleltettük meg egymásnak, nem foglalkoztunk azokkal az esetekkel, amikor egy adott szerkezet idegennyelvű fordítását egyetlen ige jelentette.

Az így létrejött korpuszon felszíni jellemzőket, valamint morfológiai, szintaktikai és lexikai információkat felhasználva tanítottuk gépi tanuló megközelítésünket, amely ezáltal képes párhuzamos korpuszokon félig kompozicionális szerkezetek idegennyelvű megfelelőinek automatikuson detektálására és így egy automatikus szótár építésére. A módszer előnyei közé tartozik továbbá, hogy amennyiben rendelkezésünkre áll egy adott nyelvű FX-azonosító rendszer, és az adott nyelvre léteznek párhuzamos korpuszok, akkor automatikusan tudunk FX-szótárakat generálni az adott nyelv és a párhuzamos korpusz többi nyelve alkotta párokra.

2. Kapcsolódó munkák

Az összetett kifejezések gépi fordító megközelítések általi automatikus fordításainak hatékonyságát vizsgáló kutatások [7] azt mutatjak, hogy számos nyelvpáron ezen szerkezetek automatikus fordítása meglehetősen nehéz feladat. Ennek megfelelően többek közt a gépi fordítórendszerek támogatása céljából jelenleg számos aktív kutatás foglalkozik az összetett kifejezések automatikus fordításával [8]. Ezen módszerek többsége [9,10] először valamilyen automatikus megközelítés segítségével azonosítja az összetett kifejezéseket eltérő nyelveken, majd a lehetséges fordítási párok kiválasztására adnak különböző megoldásokat. Jelen munkában egy hasonló elvekre épülő megközelítést mutatunk be magyar nyelvű félig kompozicionális szerkezetek angol megfelelőinek automatikus azonosítására.

3. Félig kompozicionális szerkezetek fordításainak automatikus azonosítása

Jelen munkában elődleges célunk magyar nyelvű félig kompozicionális szerkezetek angol megfelelőjének automatikus azonosítása párhuzamos korpuszokból. Vizsgálatainkat alapvetően a SzegedParalellFX [1] párhuzamos korpuszon végeztük, ahol az FX-ek magyar és angol nyelven is manuálisan jelölve vannak. Ugyanakkor méréseink elvégzéséhez még szükséges volt az egyes fordítási egységekben előforduló különböző nyelvű FX-ek manuális megfeleltetése is. Jelen munkában csak a magyar nyelvű FX-ek angol nyelvű megfelelőinek megtalálása a célunk, oly módon, hogy minden olyan magyar fordítási egységben, ahol előfordult egy manuálisan annotált FX, akkor az egység angol nyelvű megfelelőjéből a korábban már bemutatott jelöltkinyerő algoritmus [4] segítségével automatikusan kinyertük a lehetséges angol nyelvű FX-ket. Az annotátornak ezen potenciális FX-ek közül kellett kiválasztania a magyar nyelvű FX angol nyelvű megfelelőjét. A SzegedParalellFX magyar részében összesen 1377 FX van manuálisan annotálva, a hozzájuk tartozó angol nyelvű fordítási egységekben összesen 446 FX-nek találtuk meg a megfelelő fordítását, ezenkívül további 4635 egyéb lehetséges FX-et generált az automatikus jelöltkinyerő rendszer.

Mivel a megközelítésünk erősen támaszkodik az FX-ek morfoszintaktikai tulajdonságaira is, ezért szükségesnek bizonyult a párhuzamos korpusz nyelvi elemzése. Ennek során a magyar szövegek nyelvi elemzéséhez a `magyarlanc 2.0`-t [11] alkalmaztuk, míg az angol nyelvű szövegek elemzését a Stanford elemző [12] segítségével valósítottuk meg.

4. Gépi tanuló megközelítés félig kompozicionális szerkezetek fordításainak azonosítására

Az egyes FX-ek párok automatikus azonosítására egy gépi tanuló megközelítést alkalmaztunk. Ehhez alapvetően az FX-ek automatikus azonosításához korábban már ismertett [4] felszíni, morfológiai, szintaktikai és lexikai jellemzőkre támaszkodtunk, valamint új jellemzőket is definiáltunk. A korábban már ismertett, ebben a feladatban is felhasznált jellemzők a következők voltak:

- Felszíni jellemzők: a **végződés** jellemző azt vizsgálja, hogy a szerkezet főnévi tagja bizonyos bi- vagy trigramra végződik-e. Ezen jellemző alapja, hogy az FX-ek főnévi komponense igen gyakran egy igéből képzett főnév. A szerkezetet alkotó **tokenek száma** szintén jellemzőként lett felhasználva.
- Lexikai jellemzők: A **leggyakoribb ige** jellemző az FX-ek azon tulajdonságára támaszkodik, hogy a leggyakoribb igék sokszor funkcióigeként is szerepelhetnek (például *ad*, *vesz*, *hoz* stb.). Ezért az FX-jelöltek igei komponensének lemmáját vizsgáltuk, hogy az megegyezik-e az előre megadott leggyakoribb igék egyikével.
- Morfológiai jellemzők: A **szótő** jellemző alapvetően a főnévi komponens szótövét vizsgálja. Ez a jellemző az FX-ek azon már említett tulajdonságát kívánja kihasználni, hogy a félig kompozicionális szerkezetek főnévi tagja igen

gyakran egy igéből származik, ezért azt vizsgáltuk, hogy a főnév tag szótövének van-e igei elemzése. Továbbá mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért néhány magyarspecifikus morfológiaalapú jellemzőt is alkalmaztunk. Így megnéztük a magyar funkcióigék **MSD-kódját** felhasználva az ige módját (`Mood`), valamint a főnévi komponens típusát (`SubPos`), esetét (`Cas`), a birtokos számát (`NumP`), a birtokos személyét (`PerP`), valamint a birtok(olt) számát (`NumPd`).
– Szintaktikai jellemzők: korábbi kutatásaink azt mutatták [13], hogy az FX-ek igei és főnévi tagja közt csupán néhány szintaktikai osztályba tartozó él fordulhat elő, mint például alanyi vagy tárgyi. Ezen **szintaktikai osztályokat** szintén felhasználtuk jellemzőként.
– Szemantikai jellemzők: ebben az esetben is az FX azon tulajdonságát használtuk fel, hogy a főnévi tag igen gyakran egy igéből származik. Ezért a Magyar WordNetet [14] valamint a Princeton WordNet 3.1-et[1] felhasználva `tevékenység` vagy `esemény` **szemantikai jelentést** keresünk a főnévi tag felsőbb szintű hipernimái közt.

A jellemzőkészlet kialakítása során megvizsgáltuk az egyes jellemzők értékeit külön-külön a magyar FX-re és a hozzá tartozó angol FX-jelöltre, valamint megnéztük, hogy értékeik egyszerre is igazak-e az adott FX párra. Vagyis például amikor a leggyakoribb ige jellemzőt néztük az adott FX-párra, megvizsgáltuk, hogy a magyar FX igei komponense szerepel-e a magyar leggyakoribb igék közt, valamint hogy a potenciális angol szerkezet igei tagja szintén gyakori ige-e. Végül pedig megvizsgáltuk, hogy a két ige egymás fordításai-e. A fentebb ismertett, FX-ek automatikus azonosítására már korábban használt jellemzőket további attribútumokkal egeszítettük ki. Vagyis megvizsgáltuk, hogy a két szerkezet főneveinek fordításai megegyeznek-e a szótárban, illetve hogy a szerkezet főnévi tagjának van-e szintaktikai bővítménye az adott mondatban, és amennyiben igen, annak címkéjét is felvettük.

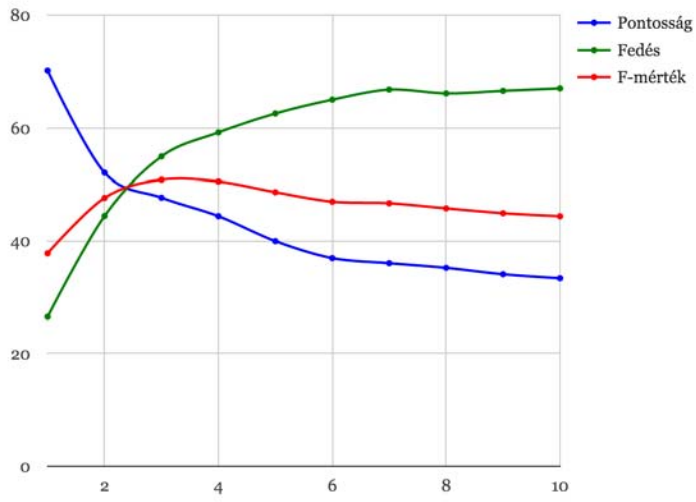
Az így létrejött tanítókorpuszon a WEKA gépi tanuló csomagban [15] található C4.5 döntési fa algoritmust implementáló J48 tanuló algoritmust alkalmaztuk. A kiértékelés során tízszeres keresztvalidációt felhasználva számítottunk pontosság, fedés és F-mérték metrikákat. Mivel a tanító korpuszon a negatív példák jelentősen felülreprezentáltak a pozitív példákhoz képest, ezért a tanítás során szükségesnek találtuk a pozitív példák felülsúlyozását. A legjobb F-mértéket eredményező súly megtalálához megvizsgáltuk módszerünk hatékonyságát különböző súlyozások mellett, melynek eredményét az 1. ábra mutatja.

Baseline megoldás szerint akkor tekintettük azonosnak egy FX-párt, amennyiben a szerkezetek főneveinek jelentése a szótár szerint megegyezik. Ezen megközelítések eredményei az 1. táblázatban láthatók.

5. Az eredmények értékelése, összegzés

Jelen munkánkban bemutattuk a gépi tanuláson alapuló rendszerünket, amely automatikusan képes magyar-angol párhuzamos korpuszokból magyar nyelvű

[1] http://wordnet.princeton.edu



1. ábra. Pozitív elemek súlyozásának hatása a gépi tanuló megközelítés hatékonyságára.

1. táblázat. Baseline, valamint a gépi tanult megközelítés eredményei

Megközelítés	**Pontosság**	**Fedés**	**F-mérték**
Baseline	73,68	15,69	25,88
Döntési fa	47,63	54,93	50,81

félig kompozicionális szerkezetek angol nyelvű megfelelőit azonosítani. Ehhez először egy manuális annotált korpuszt hoztunk létre a SzegedParallelFX korpuszon, ahol a magyar nyelvű FX-ekhez potenciális FX-eket generáltunk. Az így létrejött korpusz nem csak összetett kifejezések automatikus megfeleltetésére használható, hanem segítségével megvizsgálhatjuk, hogy mennyire hatékonyan képesek a különböző gépi fordító megközelítések folyó szövegekben az összetett kifejezéseket automatikusan fordítani.

A feladat megoldása során először a lehetséges fordítási párokat automatikusan azonosítottuk a párhuzamos szövegekben, majd gépi tanuló megközelítés segítségével választottuk ki a helyes fordításokat.

Eredményeink részletesebb vizsgálata alapján elmondhatjuk, hogy elsősorban azok az esetek jelentettek nehézséget a gépi tanulónak, amikor egy adott fordítási egységen belül az angolban és a magyarban is megtalálható volt egy FX, ezek azonban nem voltak egymás fordítási egységei, lásd pl.:

> *Háromévi várakozás után William Prichard kapitány, az Antilop gazdája, ki a déli vizekre volt indulóban, előnyös* ***ajánlatot tett*** *nekem, és én elfogadtam.*

After three years expectation that things would mend, I accepted an advantageous offer from Captain William Prichard, master of the Antelope, who was ***making a voyage*** *to the South Sea.*

Ahogy láthatjuk, a fenti esetben az *ajánlatot tett* kifejezést a rendszer megfeleltette a *making a voyage* kifejezésnek, ez azonban nem bizonyul helytállónak. A rendszernek nehézséget okozott továbbá a ritkábban előforduló igéket tartalmazó FX-ek sikeres azonosítása is (például *(nehéz) életet élnek – lead (difficult) lives*).

Ahogy azt az 1. ábra mutatja, amennyiben a tanítás során a pozitív elemek súlyát növeltük, a gépi tanuló megközelítés pontossága folyamatosan csökkent, míg a fedése növekedett. Ezen tendenciák mellett akkor kaptuk a legjobb F-mértéket, amikor a pozitív példák 3-as súlyt kaptak. Ugyanakkor a súlyozás segítségével a létrejövő automatikus szótár minőségét az alkalmazástól függően tudjuk parametrizálni. Amennyiben elsősorban pontos szótár építése a célunk, akkor a gépi tanulás során alacsonyabb súly rendelése szükséges a pozitív példákhoz, míg ha minél több lehetséges fordítási párra vagyunk kíváncsiak, akkor a pozitív példák nagyobb súlyt kívánnak a gépi tanulás során. Ugyanakkor a feladat nehézségéből fakadóan minden esetben szükséges lehet az automatikusan létrejött szótár manuális validációja.

A generált szótárakat oktatási és kutatási célra ingyenesen elérhetővé tesszük.

Hivatkozások

1. Vincze, V., Felvégi, Zs., R. Tóth, K.: Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 91–101
2. Rácz, A., István Nagy, T., Vincze, V.: 4FX: Light verb constructions in a multilingual parallel corpus. Proc. of LREC (2014) 710–715
3. Nagy T., I., Vincze, V., Zsibrita, J.: Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 47–58
4. Rácz, A., Nagy T., I., Vincze, V.: 4FX: félig kompozicionális szerkezetek automatikus azonosítása többnyelvű korpuszon. In Tanács, A., Vincze, V., Varga, V., eds.: MSzNy 2014 – X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2014) 317–324
5. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 102–110
6. Vincze, V.: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In: Proceedings of LREC-2012, Isztambul, ELRA (2012) 2381–2388
7. Seretan, V.: Multi-word expressions in user-generated content: How many and how well translated? evidence from a post-editing experiment. In: Proceedings of the Second Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015), Malaga, Spain (2015)

8. Monti, J., Mitkov, R., Pastor, G.C., Seretan, V.: Multi-word units in machine translation and translation technologies (2013)
9. Monti, J., Sangati, F., Arcan, M.: Multi-word expressions in a parallel bilingual spoken corpus: data annotation and initial identification results (2015) Poszter. PARSEME 5th General Meeting.
10. Wehrli, E., Villavicencio, A.: Extraction of multilingual mwes from aligned corpora (2015) Poszter. PARSEME 5th General Meeting.
11. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, INCOMA Ltd. Shoumen, BULGARIA (2013) 763–771
12. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. (2014) 55–60
13. Vincze, V., Nagy T., I., Farkas, R.: Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers. (2013) 255–261
14. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, Szegedi Tudományegyetem (2008) 311–320
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18

A magabiztosság-krízis skála alkalmazása idegen nyelvű megnyilatkozásoknál[1]

Puskás László[1], Pólya Tibor[2]

[1]Pécsi Tudományegyetem Bölcsészettudományi Kara, Pszichológia Doktori Iskola
laszlopuskas@gmail.com
[2]Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu

Kivonat: Tanulmányunkban korábbi kutatásaink eredményeit kívánjuk megvizsgálni idegen nyelvű megnyilatkozásokon. 2011-ben bemutatásra került a Magyar Számítógépes Nyelvészeti Konferencián az az új narratív pszichológiai eljárás, amelyben összekapcsoltuk a narratív pszichológiai tartalomelemzést és a vokális mintázatok pszichológiai „tartalomelemzését" [12]. Ennek az eljárásnak a részeként mutattuk be a magabiztosság-krízis indexet, amely a megnyilatkozás nyelvi, tartalmi elemeit és az elhangzottak fonetikai struktúráját vizsgálva von le következtetéseket a közlő lelkiállapotára vonatkozóan. Azt a feltételezésünket igyekeztünk adatokkal is alátámasztva igazolni, hogy a krízishelyzet nyelvi-fonetikai mintázata jól körülhatárolható, és ezen jegyek alapján a közlő lelkiállapotára vonatkozóan pszichológiailag értékelhető megállapítások tehetők. Vizsgálatunk nyelvi anyagát akkor Shakespeare Lear királya első és utolsó monológjának magyar nyelvű változata alkotta. 2014-ben a Magyar Számítógépes Nyelvészeti Konferencián mutattuk be a magabiztosság-krízis skála spontán megnyilatkozásokon történő alkalmazását [11]. A mostani kutatásunk a 2011-ben elhangzott előadáshoz kíván visszanyúlni, hasznosítva az időközben szerzett tapasztalatokat, egy olasz nyelvű Lear király megnyilatkozásával gazdagítva a korábbi kutatások eredményeit.

1. Bevezetés

Az emberek és csoportjaik a történetek révén saját identitásuk és pszichológiailag érvényes valóságuk számos lényeges vonását alkotják meg. Ezek a történetek vallanak az elbeszélők várható viselkedési adaptációjáról és megküzdési képességeiről egyaránt. A tudományos narratív pszichológia az elbeszélést komplex pszichológiai tartalmak hordozójának tekinti, melynek tanulmányozása révén eredményesen vizsgálható az emberi társas alkalmazkodás. Azt a szoros kapcsolatot hangsúlyozza, amely a pszichológiai folyamatok, az elbeszélés és az identitás között van [8].

[1] A tanulmány az OTKA K 109009 számú pályázat támogatásával készült.

A narratív pszichológiai kutatások módszerét az írott szövegek elemzésére dolgozták ki. Bár sok esetben a kutatások hanganyagát rögzítették, ez kizárólag azt a célt szolgálta, hogy az elhangzottakat lejegyezzék, magát a hanganyagot nem használták fel. A cél meghatározta az eszközöket is, hiszen nem volt szükség arra, hogy a fonetikai struktúra elemzésére alkalmas, jó minőségű felvételek készüljenek, elég volt, ha a felvétel lejegyezhető minőségű, így ezen felvételek jelentős része valószínűleg nem is lett volna alkalmas komolyabb fonetikai vizsgálatok lefolytatására. A tudományos narratív pszichológia és a fonetikai jelenségek összekapcsolásának lehetősége még viszonylag új keletű, de izgalmas és komoly kutatási eredményekkel kecsegtető terület.

2. A tudományos narratív pszichológia és a fonetikai elemzések összekapcsolásának előzményei

A megnyilatkozások nyelvi tartalmi jegyeinek és fonetikai struktúrájának párhuzamos vizsgálatáról szóló első tudományos előadás bemutatására 2005-ben került sor az Alpok-Adria Pszichológiai Konferencián, Zadarban. Ebben az előadásban fogalmazódott meg először az igény a tudományos narratív pszichológiai megközelítés kereteinek kibővítésére. Az előadás anyaga később egy konferenciakötetben is megjelent [13]. A kezdeti időszakban a kutatás egy olyan program alapjainak a kifejlesztésére irányult, amely a nyelvi és a fonetikai jegyek együttes és korlátlan vizsgálatára ad lehetőséget. Ekkor még hangsúlyosabb szerep jutott ennek az eszköznek a kifejlesztésére, mint egy minden részletre kiterjedő eljárás kidolgozására. Elsősorban a narratív pszichológiai tartalomelemzés módszereit és az elhangzott szöveg fonetikai struktúrájának elemzését kívántuk integrálni. Ez az időszak 2007-ig tartott [11].

2008-tól egyre inkább áttevődött a hangsúly az eljárás kidolgozására, amely képes megvalósítani a nyelvi tartalmi elemek és a fonetikai jegyek integrációját. (Később a programfejlesztés végleg elvetésre került.) A változást indokolta, hogy nagyjából erre az időre a tartalomelemzés területén a NooJ fejlettsége [18], valamint a hozzá kapcsolódó modulok fejlesztése olyan szintre jutott, amely megkérdőjelezte egy önálló program megvalósításának létjogosultságát, amely ezzel a programmal már nem vehette volna fel a versenyt, miközben a Praat programmal [1] a fonetikai elemzések elvégezhetők voltak [11].

Az új eljárással lefolytatott vizsgálatok eredményeinek bemutatására 2011-ben került sor, a Magyar Számítógépes Nyelvészeti Konferencián. Shakespeare Lear királyának első és utolsó monológja volt az elemzés tárgya, amely a színész modellálta helyzet fonetikai és a színműíró szövegének nyelvi tartalmi jegyeit elemezte [12]. A vizsgálat célja a magabiztosság és krízis nyelvi jegyeinek vizsgálata volt, amit részben a Pennebaker és Ireland [10], valamint a László János és munkatársai [7] által kidolgozott módszer segítségével végeztünk el. Ehhez kapcsolódott a krízis fonetikai jegyeinek vizsgálata, amelyhez részben Scherer [15] korábbi összefoglaló tanulmányát használtuk fel, amely harminckilenc korábbi tanulmány fonetikai vizsgálatait összegezte, részben saját feltevéseinkkel kiegészítve, melyek a Scherer-féle koncepcióból levezethetők. A nyelvi és a fonetikai markerek összekapcsolásával, melyek együttesen

jelzik a krízis jelenlétét, illetve mértékét, létrehoztuk a magabiztosság-krízis indexet [11, 12, 14].

2014-ben részben a 2011-es konferencián bemutatásra került eredmények [12], részben pedig a 2012-ben László Jánossal és Fülöp Évával közösen írt tanulmány [14] eredményeinek továbbfejlesztésével, spontán megnyilatkozásokon is igazoltuk korábbi kutatási eredményeink érvényességét, és a fonetikai elemzések vizsgálatának létjogosultságát egy komplex narratív pszichológiai eljárásmód keretében [11].

Ez a tanulmány a 2011-ben elhangzott előadáshoz kíván visszanyúlni, hasznosítva az időközben szerzett tapasztalatokat, egy olasz nyelvű Lear király megnyilatkozásaival gazdagítva a korábbi kutatások eredményeit. Az olasz nyelvű megnyilatkozásokat is a magabiztosság-krízis indexszel vizsgáltuk meg, melynek értéke a nyelvi markereket és a vokális jelzéseket egyaránt figyelembe veszi. Az új kutatás elsődleges célja, az index használhatóságának ellenőrzésén túl, annak alátámasztása, hogy a krízisnek, illetve a magabiztos lelkiállapotnak létezik egy olyan nyelvfüggetlen mintázata, amely mind a nyelvi tartalmi elemekben, mind pedig a fonetikai struktúrában megjelenik, és amelyet együttesen határoznak meg. Ennek a mintázatnak a vizsgálatára pedig a magabiztosság-krízis index megfelelő eszköznek bizonyul.

3. A vizsgálat

3.1. A vizsgálati anyag

Vizsgálatunk nyelvi anyagát Lear király 1960-ban készült olasz televíziós adaptációjából választottuk ki [17]. A korábbi vizsgálathoz hasonlóan Lear első és utolsó monológját elemeztük. A szöveg nyelvezetének eltéréseit is figyelembe véve, nem törekedtünk teljes megfelelésre. Mivel teljesen azonos krízist mutat be a Lear király olasz változata is, feltételeztük – ahogy a magabiztosság-krízis indexszel lefolytatott eredménynél is –, hogy a krízisjegyek a teljes szövegbeli egyezéstől függetlenül megjelennek. Természetesen nem azt vártuk, hogy az index értékei azonosak lesznek a korábbi vizsgálatban kapott értékekkel, hanem az értékeknek a skála „megfelelő" pólusán történő elhelyezkedését.

A vizsgálati anyag kiválasztásával kapcsolatban felmerül a kérdés, hogy az irodalmi alkotások, illetve azok színészi megjelenítése mennyiben adhat valós képet a krízishelyzet és a magabiztosság nyelvi tartalmi jegyeiről és fonetikai sajátosságairól. A kérdés nyelvészeti és pszichológiai megközelítése eltérő. Bár a nyelvészek is használnak felolvasott szöveget vagy színészi játékot elemzéseik lefolytatásához, nehezen lenne vitatható, hogy a spontán beszéd és a színészi játék a hétköznapi ember számára is nagy pontossággal megkülönböztethető. (Például a spontán megnyilatkozásokat egy sor megakadási jelenség is kíséri, míg a színészi játéknak ez nem jellemzője.) A szöveg életszerűségén túl, felmerül az is, hogy vajon mennyire adhatja vissza egy irodalmi fikció azokat a nyelvi tartalmi és fonetikai jegyeket, amelyek a hétköznapi beszédben például a krízis jellemzői. Ha a nyelvészek szempontjából nézzük, akkor mind a fonetikai jegyek, mind pedig a szöveg jellemzőit vizsgálva egy *technikai* jellegű problémával találjuk magunkat szemben.

Miben különbözik a kérdés tudományos narratív pszichológiai megközelítése a nyelvészet *technikai* jellegű problémafelvetésétől? A narratív pszichológia az elbeszélésekben fellelhető pszichológiai tartalmak felől közelít. Fülöp és munkatársai [4] például négy jelentős történelmi regény elemzését végezték el. A nemzeti identitást az irodalom, a film, a média és a népművészet is közvetíti. A történelmi regények és elbeszélések többnyire egy nemzet identitásképzésének és az identitás fenntartásának fontos eszközei [5, 9]. Az egyes szereplőknek tulajdonított érzelmek külön is értelmezhetők, az adott jellem érzelmi működésmódjának elemeiként, miközben a csoportot jellemző érzelmi dinamikát a csoporthoz tartozó szereplők érzelmi reakciói jelenítik meg [8].

Pennebaker és Ireland [10] nemcsak Shakespeare Learjét vizsgálta, de Rudolph Giuliani New York-i polgármester 1993-as megválasztását követő, valamint 2000-es megnyilatkozásait is, aki egy kéthetes időszak alatt bejelentette, hogy elválik, nyilvánosságot látott szerelmi ügye, prosztata-rákot állapítottak meg nála, és visszalépett a szenátusi megmérettetéstől. Pennebaker és Ireland vizsgálatában a magabiztosság és a krízis mintázata Giuliani esetében nagyon hasonlított a Learnél talált jegyekkel.

A fonetikai jegyek vizsgálatánál pszichológiai szempontból nem az az érdekes, hogy vajon a spontán beszéd megkülönböztethető-e a színészi játéktól, sokkal inkább az, hogy ha intenzívebben is, de a krízisre jellemző fonetikai mintázatokat tanulmányozhassuk, amely a közlő egyéni adottságaitól függetlenül is bizonyos keretek között jelenik meg. Ezeknek az elvárásoknak pedig a vizsgálati anyag megfelel.

3.2. Módszer és eredmények

3.2.1. A fonetikai paraméterek vizsgálata

Az érzelmi állapotok fonetikai paraméterekre gyakorolt hatását a témában korábban lefolytatott vizsgálatok alapján tanulmányoztuk [11, 12, 14], melyekben Scherer [15] összefoglaló tanulmánya is felhasználásra került. Scherer harminckilenc korábbi tanulmány tapasztalatait összegezte, amelyek a fonetikai paraméterek és az érzelmi állapotok közötti kapcsolatot vizsgálták. Ezek a kutatások több évtized munkáját rendszerezték, és foglalták egységes fogalmi keretbe. (A vizsgált kutatások nem használtak egységes fogalmakat az érzelmi állapotok leírására.) A vizsgálatok során megvizsgáltuk a fonetikai paraméterek „finomhangolásának” lehetőségét is.

Az monológok akusztikai változásait a Praat [1] fonetikai programmal vizsgáltuk, amit az Amszterdami Egyetemen fejlesztettek ki.

3.2.2. A magabiztosság-dominancia és a krízishelyzet skálázása, a magabiztosság-krízis index

A krízis, Caplan meghatározása szerint, olyan lelkiállapot, amely külső események hatására alakul ki, amikor az egyének olyan problémákkal találják magukat szemben, amelyek mindennél fontosabbá válnak számukra, és amelyeket sem elkerülni, sem pedig a szokásos eszközökkel megoldani nem tudnak [2]. A krízis meghatározásából következik, hogy egy meglehetősen sokszínű, és intenzitásában nagymértékben eltérő jelenségcsoportot ölel fel a meghatározás.

A krízishelyzet vizsgálatánál, annak időbeli elhúzódását is figyelembe kell vennünk, illetve a feldolgozás időtartamát. A krízishelyzet és hatásainak feldolgozása, illetve azok kezelhetővé válása, jelentősen csökkentheti a krízishelyzetre utaló jeleket a közlő elbeszélésében, amiből legfeljebb a krízis feldolgozottságának mértékére következtethetünk [11].

A magabiztosság-krízis index paramétereinek tartalmát a korábbi kutatások változatlanul határozták meg [11, 12, 14]. Az új vizsgálatok lefolytatásakor a korábban vizsgált paraméterek pontosításának, „finomhangolásának" lehetőségeit is megvizsgáltuk. (Például a rövid beszédszakaszok hosszának meghatározása.) Az eljárás lefolytatásával tehát nemcsak a korábbi vizsgálatok nyelvfüggetlen felhasználhatóságát kívántuk bebizonyítani, hanem a korábban felhasznált paraméterek pontosításának lehetőségét is. Az index kiszámításához, a korábbi vizsgálatoknak megfelelően, hat arányszámot használtunk fel, melyek értékét egymással összeadtuk. Az arányszámokat értelmeztük és felcímkéztük [11, 12, 14]:

1. Rövid beszédszakaszok: kettő másodperc alatti beszédszakaszok száma osztva a vizsgált szöveg szószámával. (Az esetszámok növekedésének tapasztalatai alapján valószínűleg érdemes megvizsgálni annak lehetőségét, hogy ennek értékét csökkentsük 1,6 másodpercet meg nem haladó értékre. Az index értékét ezzel az alternatív lehetőséggel is kiszámoltuk.)

2. Magas hangerő: a hangerőcsúcsokat tartalmazó beszédszakaszok száma osztva a vizsgált szöveg szószámával. (Ebbe a kategóriába tartozik minden nyolcvan dB-t meghaladó beszédszakasz, de a megnyilatkozótól függően ennek mértéke a beszélőhöz mérten csökkenthető.) Ezen a paraméteren nem kívántunk változtatni, ugyanakkor a beszélő személyes adottságaihoz mérten, figyelembe véve a 78 és 80 dB közé eső határértékek viszonylag nagy előfordulási arányát az utolsó monológban, a korábbi meghatározásnak megfelelően a legalább 78 dB-s értékeket soroltuk ebbe a kategóriába.

3. „Monoton" beszéd: az alacsony hangerő-intervallumokat tartalmazó beszédszakaszok száma (amelyek nem haladják meg a húsz dB-t) osztva a vizsgált szöveg szószámával. Ezen a paraméteren nem változtattunk, de hosszabb távon, kellő számú vizsgálat lefolytatása esetén, érdemes felülvizsgálni, hogy ez a paraméter változatlan formában az index részét képezze-e.

4. Szelf-referencia: a szelf-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával.

5. Tagadás: a tagadásra vonatkozó szavak száma osztva a vizsgált szöveg szószámával.

6. Mi-referencia (negatív korrekciós index): a mi-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával, negatív előjellel.

4. A magabiztosság-krízis index segítségével nyert eredmények

A magabiztosság-krízis indexszel kapott eredmények azt mutatják, hogy Lear első monológját a magabiztosság, kiegyensúlyozottság jellemzi, míg utolsó monológja

erőteljes krízishelyzetet mutat. Az olasz nyelvű előadás [17] eredményeit összevetettük a magyar előadás [16] eredményeivel [12, 14], melyet az 1. táblázat mutat.

1. táblázat: A magabiztosság-krízis index értéke a hat felhasznált mérőszám alapján

Mérőszámok	1	2	3	4	5	6	Összesen
Lear 1. monológ – olasz	0,2118	0,0059	0,0294	0,0000	0,0059	-0,1176	0,1353
Lear 2. monológ – olasz	0,3003	0,0924	0,0231	0,0858	0,0726	0,0000	0,5743
Lear 1. monológ – magyar	0,0540	0,0270	0,0135	0,0000	0,0135	-0,2162	-0,1082
Lear 2. monológ – magyar	0,3200	0,2533	0,1333	0,1200	0,0133	0,0000	0,8399

A táblázatban szereplő értékek, a korábbi vizsgálatoknak megfelelően, továbbra is arról tesznek tanúbizonyságot, hogy a vokális és az írott szövegben mért paramétereket külön-külön összesítve, eltérő mérőszámokat kapnánk, és együttesen határozzák meg a krízishelyzet mértékét, illetve a közlő magabiztos lelkiállapotát.

Megvizsgáltuk a magabiztosság-krízis index alakulását úgy is, ha a rövid beszédszakaszok hosszát 1,6 másodpercben határoznánk meg. Az olasz nyelvű monológoknál a magabiztosság-krízis index 0,0882-re és 0,5314-re változna, azaz a korábbi eljárással kapott valamivel több mint négyszeres eltérés hatszorosra. A magyar nyelvű változatnál az első monológban az index értéke nem változna, míg a másodikban 0,7831-re csökkenne, ami viszont még így is az eredeti érték több mint 93 százaléka.

A fonetikai paraméterek vizsgálatánál azt a korábbi vizsgálatoknál alkalmazott elvet követtük, hogy a kiválasztott beszédszakaszok szószámát osztottuk el a vizsgált szöveg szószámával. Ha egy beszédszakasz több vizsgált fonetikai paraméternek is megfelelt, akkor valamennyi fonetikai paramétert külön számítottuk be, mintha annyi megjelölt szó lenne az adott beszédszakaszban, ahány az általunk vizsgált fonetikai paraméternek megfelel, függetlenül attól, hogy hány szóból állt a beszédszakasz. Erre azért volt szükség, mert ha több kiugró értéket tartalmaz egy beszédszakasz, akkor intenzívebb a megnyilatkozó lelkiállapota [11, 12, 14].

A rövid beszédszakaszok relatív előfordulási gyakoriságát azért használtuk fel az indexhez, mert azt feltételeztük, hogy a beszédszakaszok hosszából következtethetünk a beszélő gondolatainak összeszedettségére, fájdalmára, és hogy az adott helyzetre milyen korábban konstruált sémával rendelkezik. Rövid beszédszakaszok magabiztos megnyilatkozásokban is találhatók, de feltételezésünk szerint kisebb arányban [11, 12, 14]. A vizsgálati anyag magabiztos megnyilatkozásában előforduló magas arányuk arra hívja fel a figyelmet, hogy a két másodpercben meghatározott határértéket a további vizsgálatok függvényében valószínűleg csökkenteni kell.

A hangerőcsúcsokat tartalmazó beszédszakaszok a korábbi vizsgálatok tapasztalatai alapján fontos szerepet töltenek be a krízis meghatározásában [11, 12, 14].

Az alacsony hangerő-intervallumok gyakorisága, feltételezésünk szerint, olyan monotonitást kölcsönöz a megnyilatkozásnak, amely erő és magabiztosság hiányára, rossz lelkiállapotra utal [11, 12, 14]. Ugyanakkor ennek a változónak az alkalmazása, ha már nagyobb adatbázissal rendelkezünk újraátgondolásra szorul.

A szelf-referencia és a tagadás előfordulási gyakoriságát nemcsak Pennebaker és Ireland [10] vizsgálta, de László és munkatársai [7] is, akik ezek relatív gyakoriságát nézték meg a szövegben. Az énre való túlzott utalás a befelé fordulás jele, míg a 'mi'-re történő utalás a mások irányába való nyitást fejezi ki. Patológiás esetben a magas én-referencia összefüggést mutat a depresszióval, a szuicid tendenciákkal. A tagadást pszichodinamikai szempontból az egészséges emberi környezethez és morális mércékhez való alkalmazkodásra, illetve a világ értéktelenítésére, a destrukcióra és öndestrukcióra való hajlamra vonatkozóan vizsgálták [3]. Krízishelyzetben a megváltozott környezethez való alkalmazkodás problémás, fokozottan fordulhat elő tagadás az elbeszélésben.

A mi-referencia a magabiztosság-krízis indexnél negatív korrekciós mérőszámként került felhasználásra, mivel értéke pont a kiegyensúlyozott megnyilatkozásoknál a legmagasabb, így az indexet alkotó többi paraméterrel szemben ellenkező hatást fejt ki [11, 12, 14].

5. Összegzés

Korábbi kutatásainkban egy új narratív pszichológiai szemlélet meghonosítására tettünk kísérletet, amely összekapcsolja a narratív pszichológiai tartalomelemzést és a fonetikai jegyek vizsgálatát egy *összetett tudományos narratív pszichológiai eljárás* keretében. Vizsgálataink bebizonyították, hogy mind a színészi játékban, mind pedig a spontán megnyilatkozásokban eredményesen alkalmazható az az eljárás, amit a krízis mérésére dolgoztunk ki, és amelynek számszerűsítésére a magabiztosság-krízis indexet alkalmaztuk [11, 12, 14].

Mostani kutatásunkban a kidolgozott eljárás nyelvfüggetlen felhasználását vizsgáltuk. Vizsgálatunk nyelvi anyagát Shakespeare Lear királyának magyar és olasz nyelvű monológjai alkották, amelyeknél a színészi játék modellálta helyzetben vizsgáltuk a krízis és a magabiztosság nyelvi tartalmi jegyeinek fonetikai paraméterekkel összekapcsolt mintázatát. Kutatási eredményeink arról tesznek tanúbizonyságot, hogy a magabiztos lelkiállapot és a krízishelyzet nyelvfüggetlenül is eredményesen vizsgálható, a magabiztosság-krízis indexszel nyert eredmények a skála „megfelelő" pólusánál helyezkedtek el. Ez azonban nem zárja ki annak lehetőségét sem, hogy az esetszámok és a tapasztalatok függvényében az indexet finomhangoljuk.

Összességében elmondhatjuk, hogy a nyelvi tartalmi jegyek és a fonetikai paraméterek összekapcsolása a lelkiállapot-változásokkal eredményesen alkalmazható technika, amely a krízis esetében nyelvfüggetlenül is vizsgálható, és adatokkal alátámasztható. A kutatások további iránya lehet az esetszámok növelése, és a tapasztalatok növekedésével az index finomhangolása, illetve más jelenségek hasonló technikával történő vizsgálata, valamint a magabiztosság-krízis indexszel lefolytatott vizsgálat más nyelvre, illetve nyelvekre történő kiterjesztése.

Hivatkozások

1. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer [computer program]. Forrás:http://www.praat.org/ (2013)
2. Caplan, G. (1964). Principles of preventive psychiatry. New York, Basic Books.
3. Hargitai, R. Naszódi, M., Kis, B., Nagy, L., Bóna, A., László, J. (2005). A depresszív dinamika nyelvi markerei az én-elbeszélésekben. A LAS VERTIKUM tagadás és szelfreferencia modulja. Pszichológia, 2 (2005) 181-199.
4. Fülöp, É., Péley, B., László, J.: A történelmi pályához kapcsolódó érzelmek modellje magyar történelmi regényekben, Pszichológia, 31, 1 (2011) 47-61.
5. László J.: A történetek tudománya: Bevezetés a narratív pszichológiába. Pszichológiai Horizont; Budapest, Új Mandátum Könyvkiadó. (2005)
6. László J.: Előszó. Forrás: László J., Thomka B. (szerk.): Narratív pszichológia. Narratívák 5. Budapest, Kijárat Kiadó. 7-15. (2001)
7. László, J.: The science of stories.: An introduction to narrative psychology. London; New York: Routledge. (2008)
8. László, J.: Történelemtörténetek – Bevezetés a narratív pszichológiába. Budapest, Akadémiai Kiadó. (2012)
9. Liu, J. H., László, J.: A narrative theory of history and identity: Social identity, social representations, society and the individual. In: Moloney, G., Walker, I. (eds.): Social representations and history: Content, process and power. London, Palgrave-Macmillan (2007) 85-107.
10. Pennebaker, J. W., Ireland, M.: Analyzing Words to Understanding. In: Jan Auracher, William van Peer (Eds.): New Beginnings to Literary Studies. Cambridge Scholar Publishing. 24-48. (2008)
11. Puskás, L.: A magabiztosság-krízis skála gyakorlati alkalmazása. Forrás: Tanács A., Varga V., Vincze V. (szerkesztették): X. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (JATE Press), Szeged (2014) 155-163.
12. Puskás L.: Paralingvisztikai jegyek a narratív pszichológiai tartalomelemzésben: a magabiztosság-krízis skála. Forrás: Takács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (JATE Press) Szeged (2011) 231-239.
13. Puskás, L., Karsai, B.: A New Method in Narrative Psychology. In: Cognition and Interpretation. Pécs Studies in Psychology. Edited by Beatrix Lábadi. PTE BTK Pszichológiai Intézet. (2008)
14. Puskás, L., László, J., Fülöp, É.: Lear király lelkiállapot-változása első és utolsó monológjának szövegbeli és akusztikai jegyei alapján. Pszichológia, 2 (2012)
15. Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99, 143-165. Magyarul: Vokális érzelemkifejezés. Áttekintés és egy modell az eljövendő kutatásokhoz. Fordította: Bodor Péter. Forrás: Barkóczi I., Séra L. (szerk.): Érzelmek és érzelemelméletek. Budapest, Tankönyvkiadó, 1989.
16. Shakespeare, W.: Lear király. Vörösmarty Mihály fordítását Mészöly Dezső dolgozta át. Rendező: Vámos László. Magyar Televízió, 1978. (Tévéjáték, 156 perc.) (1978)
17. Shakespeare, W.: Re Lear. Gino Chiarini fordítását Sandro Bolchi dolgozta át és rendezte. (Tévéjáték, 185 perc.) (1960)
18. Silberztein, M.: NooJ manual. Forrás: http://www.nooj4nlp.net (2003)

A magyar Wikipédia automatikus bejárása és elemzése

Simkó Marcell[1], Góth Júlia[2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter u. 50/a, Magyarország
simko.marcell@hallgato.ppke.hu
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter u. 50/a, Magyarország
goth.julia@itk.ppke.hu

Kivonat: A mai tudományos élet és hétköznapok elengedhetetlen segédeszköze a Wikipédia. Az ingyenes adatbázis hatalmas mennyiségű információt tesz elérhetővé bárki számára. Az információ azonban nem csak az egyes szócikkek szövegében van, hanem a cikkek összekapcsoltságában is, melyet az oldalakon található hiperhivatkozások adnak. Ebben a cikkben feltérképezzük a magyar Wikipédia gráfszerkezetét, majd elemzéseket és vizualizációkat végzünk. Megállapítjuk, hogy a Wikipédia leírására használható a „kis világ" jelenség. A szócikkek szöveges tartalmát is felhasználva megvizsgáljuk a szavak hatványtörvény szerinti eloszlását, és ezt összevetjük a Zipf törvénnyel.

1. Bevezetés

A tényszerű információk visszakeresésének, lekérdezésének egyik legfontosabb és legtöbbet használt eszköze a mai világban kétségkívül a Wikipédia. Emiatt vált érdekessé a Wikipédia hálózatként való szemlélete és a benne rejlő struktúra feltérképezése. A több száz nyelven létező, sok millió lapot tartalmazó weblapon jelen lévő adatmennyiség hatalmas méretének köszönhetően lehetségessé válik olyan kutatások elvégzése, melyek a felhasználók által összegyűjtött tudásban rejlő struktúrát térképezik fel. A nagy mennyiségű természetes nyelvű szöveg elemzésével pedig kvantitatív mérőszámok alapján összehasonlítást tudunk végezni különböző nyelvek (magyar és angol) között.

Az adathalmaz kiválasztásakor figyelembe vettük, hogy nagy, de még kezelhető mennyiségű szöveges adatra van szükség, melyben az információ több, jól elkülönülő témára van osztva. Ilyen szempontból a Wikipédia ideális, hiszen az egyes lapok eleve egy tematikus bontást jelentenek, a hiperszövegben lévő linkek pedig a lapok között felállítanak egy gráf szerű struktúrát. A webhely ingyenes és szabadon hozzáférhető volta, valamint uniform stílusa pedig könnyűvé teszi annak automatikus vizsgálatát. A többi nyelvhez képest a magyar nyelvű Wikipédia közepes méretűnek mondható, melynek köszönhetően a feldolgozás viszonylag gyorsan elvégezhető, viszont mégis rendelkezésünkre áll elég adat megbízható statisztikák elkészítéséhez.

A magyar Wikipédiát mint hálózatot két megközelítésből is vizsgáljuk: a Wikipédia oldalakon lévő linkstruktúra alapján, valamint a Wikipédia oldalak szócikkeinek szö-

veges tartalma alapján. Egy speciálisan erre a célra írt keresőrobot a Wikipédia kezdőlapjától indulva bejárta az összes magyar lapot, és elmentette a linkek által kifeszített gráfot. Mivel csak a valódi tartalommal rendelkező lapok érdekeltek minket, a meta jellegű, illetve az adott lapon belülre mutató linkeket (pl. vitalap, szerkesztési lap stb.) figyelmen kívül hagytuk, a Wikipédián kívülre mutató hivatkozásokkal együtt. A végeredményként kapott gráf több mint 300 ezer csúccsal és 2 millió éllel rendelkezik, mely kifejezetten nehézzé teszi a gráf vizualizációját. Különböző statisztikai jellemzők alapján kiválasztunk részgráfokat, és szemléletesen ábrázoljuk őket.

Megvizsgáljuk, melyek a számok szerint legfontosabbnak tűnő szócikkek, olyan jellemzők alapján, mint pl. legtöbbet hivatkozott lap. A teljes gráfon kvantitatív leírókat (klaszterezési együttható/klikkesedés, átlagos legrövidebb távolság) számolunk ki. Ezen leírók segítségével megállapítjuk, hogy mivel a Wikipédia gráfra nagy klikkesedés, és kis legrövidebb távolság jellemző, leírására használhatjuk a „kis világ" jelenséget, melyet már sok, valós életben előforduló gráfon megfigyeltek, az élet legkülönfélébb területein (fehérjehálózatok, telefonhívások, baráti hálózatok stb.).

A hiperlinkek által leírt gráf struktúra mellett vizsgálatokat végeztünk a szócikkek szövegén is. A szöveg szavakra bontása, majd szótövesítés elvégzése után megszámoljuk a szógyakoriságokat, és lemérjük a hatványtörvény szerinti eloszlás paramétereit. Ezt összevetjük a Zipf törvénnyel. Magyar nyelvű szöveges korpuszokon már végeztek vizsgálatokat korábban is, ami egy adott szűkebb terület szöveges dokumentumait vizsgálta, míg az általunk vizsgált Wikipédia oldalak sokkal heterogénebb halmazt képeznek. Az általunk kapott eredményeket összehasonlítjuk korábbi kutatásokkal mind a magyar, mind az angol nyelv tekintetében.

2. Irodalmi áttekintés

A Wikipédia adatbányászati felhasználása, gráf szerkezetének elemzése nem újdonság. Bizonyos kutatások azt a célt szolgálják, hogy új linkek automatikus beszúrásával javítsák az enciklopédiát. Ilyen például [6], mely a szócikkek különböző nyelvű verziói közötti kapcsolatokat vizsgálja, algoritmust készítve a hiányzó linkek automatikus pótlására. Más kutatások [5] a vitalapokon zajló kommunikáció fa szerkezetét vizsgálják a felhasználók közötti interakciók mintázatainak azonosításához, majd ezeket a mintázatokat összehasonlítják a különböző témájú szócikkek esetén. A szerkesztők közötti kapcsolatokat tanulmányozza [1] is, azonban vitalapok helyett az határozza meg a hálózatot, hogy kik szerkesztenek együtt egy lapot. Korrelációt fedeznek fel bizonyos gráfindikátorok és a szócikkek minősége között.

A szócikkek gráfját, valamint a kategóriák gráfját vizsgálja [7]. A kategóriagráf további elemzése során kiderül, hogy az skálafüggetlen, és kis világ tulajdonsággal rendelkezik – ezt a tulajdonságot mi a magyar szócikkek hálózatán mérjük. A cikk ezután felhasználja a kategóriagráfot szemantikus hasonlósági vizsgálatokhoz, természetes nyelvfeldolgozás céljából. A Wikipédia gráf topológiájával foglalkozik [2]. A gráf növekedését szociális hálókhoz hasonlítja, és alkalmazza rá „a gazdag még gazdagabb lesz" szabályt.

A magyar nyelvre specifikusan végeztek kutatást Dominich és társai [3]. Különböző szépirodalmi és webes korpuszokat felhasználva összehasonlítja a magyar szavak gyakoriságának eloszlását a Zipf törvénnyel. Megállapítja továbbá, hogy a magyar nyelv is „kis világ", azonban az általa használt gráf alapja szavak együtt előfordulása volt, mely lényegesen különbözik ennek a cikknek a témájától.

3. A Wikipédia mint gráf

A vizsgálatunk tárgyát képező adathalmaz a Wikipédia gráfszerkezete, mely csúcsok és élek halmazát jelenti. Elméleti szempontból fontos elkülöníteni két feladatot, illetve két gráftípust. Az első feladat az adatok begyűjtése, mely a keresőmotorok botjaihoz (crawler) hasonlóan lapról lapra ugrálva, hiperhivatkozások segítségével történik. A gráf csúcsai az egyes lapok, az irányított élek pedig az egyik lapról a másikra mutató linkek. Mivel konkrétan a Wikipédia képzi az adatgyűjtés tárgyát, a lapok szócikkeket jelentenek, a hivatkozások pedig kizárólag a Wikipédián belülre mutathatnak, egyik szócikkről a másikra.

A másik feladat az adatok elemzése. Ebben a cikkben a vizsgálódás tárgyát képző gráf megegyezik a bejárási gráffal, azonban későbbi kutatásokban célszerű lenne ennél absztraktabb gráfokat vizsgálni, ahol pl. a gráf csúcsai témaköröket jelentenének [7], az összekötő élek pedig valamilyen jelentésbeli kapcsolatokat. Az ilyen vizsgálatok túlmutatnak ennek a cikknek a hatáskörén.

3.1. Adatgyűjtés

Mielőtt a kutatás méréseit el lehetne végezni, először természetesen adatok gyűjtésére van szükség. Ebben a fázisban egy kifejezetten erre a célra írt crawlert használunk, mely a keresőmotorokéhoz hasonlóan jár lapról lapra, azonban velük ellentétben nem törődik azok szöveges tartalmával, egyetlen célja csupán a hivatkozások kigyűjtése.

A crawler vázlatos működése a következő: egy tetszőleges lap (pl. a Kezdőlap) címét berakjuk egy sorba. Amíg ez a sor nem üres, lekérdezzük a következő címhez tartozó HTML kódot, és vesszük a benne található linkeket. Eldobjuk a már látott, vagy haszontalan linkeket (lásd lentebb), majd a maradékot berakjuk a sorba. Nincs szükség prioritás felállítására a soron belül, mert az egész gráfot bejárjuk.

Ha megvizsgáljuk a szócikkekben található hivatkozásokat, rögtön feltűnik, hogy nagy részük számunkra haszontalan, ugyanis:

a) a (magyar) Wikipédián kívülre mutatnak. Ilyenek a hivatkozásjegyzék elemei, a más nyelvű wikire mutató linkek stb.. A helyes linkek szerencsére könnyen felismerhetők, mert úgy kezdődnek, hogy *„/wiki/"*.
b) a Wikipédián belülre mutatnak, de nem szócikkre. Ezek olyan meta jellegű lapokra hivatkoznak, mint pl. vitalap, szerkesztési lap, kategória lapok stb.. Ezen linkek jól meghatározott formátummal rendelkeznek, pl. *„Vita:<link>"*.

c) szócikken belül konkrét fejezetre mutatnak. Az ilyen link nem dobandó el teljesen, de a fejezet információt ki kell vágni az URL-ből. Ezek a linkek *„<szócikk>#<fejezet>”* alakúak.

Az algoritmus sebességének a szűk keresztmetszete a szerverrel való kommunikáció. A Wikipédia szerverek, miután főleg szöveges adatot szolgáltatnak a felhasználóknak, erősen korlátozzák az egy kliensre jutó sávszélességet. A teljes magyar wiki bejárásához kb. 100 órára volt szükség, összesen 19 GB adatot letöltve. A folyamat végeredménye egy 317 ezer csúcsú, 23 millió élű gráf.

3.2. A gráf elemzése

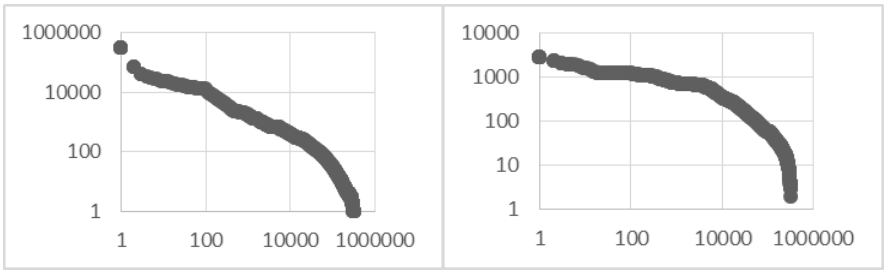
A gráf vizsgálatának a legegyszerűbb és legszemléletesebb módja természetesen a gráf vizualizációja lenne. Sajnos azonban ez a gráf egy nagyságrenddel nagyobb annál, mint amit egy hétköznapi PC-n ábrázolni lehet, ezért érdekessé válik a gráfból releváns részgráfok kiválasztása. Ehhez szükség van valamilyen statisztikai jellemzőre, mellyel kvantitatív módon lehet mérni egy szócikk fontosságát. Ennek kiválasztása azonban közel sem triviális feladat, mint elsőre gondolnánk.

A legegyszerűbb jellemző egyszerűen az adott szócikkre mutató hivatkozások száma, avagy a gráfban a csúcs befoka. Bár az adatgyűjtés során a meta jellegű lapokat már figyelmen kívül hagytuk, a szócikkek között mégis nagyon sok olyan van, melyre nem azért hivatkozik sok szócikk, mert fontosak, hanem egyéb okok miatt. A *Kezdőlap*ra például, – mely formáját tekintve szócikk, de tartalma miatt egészen speciális – minden szócikk hivatkozik. A második leghivatkozottabb lap a *Wikimédia Commons*, a harmadik a *Földrajzi koordináta-rendszer.* Nyilvánvaló, hogy ezek nem a legfontosabb cikkek, csupán azért hivatkozik rájuk sok cikk, mert bizonyos kontextusban mindig relevánsak. (Médiaállományok, illetve földrajzi helyek.) Hasonlóan sokat hivatkozott szócikk csoportok: országok, évszámok, biológiai rendszertannal kapcsolatos cikkek. Az első 10 cikket az 1. táblázatban láthatjuk. Kétféleképpen állíthatjuk fel a sorrendet: hivatkozó egyedi szócikkek száma („Befok”), illetve az összes hivatkozások száma („Befok (többszörös)”). Megnézhetjük természetesen a szócikken található linkeket is, mely a gráfban a kifokszámnak felel meg. A befokhoz hasonlóan itt is pár szócikkfajta dominálja az eredményt, így ez a mérőszám sem jól használható eszköz a lapok fontosságának méréséhez. A legjellemzőbb típusok itt a listák, táblázatok, illetve a sporttal és vasúttal kapcsolatos szócikkek. Ugyanezeket a típusokat látjuk, ha a szócikkek hosszát vesszük alapul. (Nem meglepő módon erős korreláció van egy cikk hossza és a benne lévő linkek száma között.)

Az 1. ábra bemutatja a hivatkozások számának eloszlását az összes szócikken. Csökkenő sorban a 10. szócikktől a 10000.-ig szép hatványtörvény szerinti csökkenést láthatunk, a 10000. cikk után viszont exponenciálisnál is gyorsabb lecsengés jellemző. A hatványtörvény paramétere α=-0.64, illetve α=-0.19 a befok, illetve a kifok esetén. A többszörös hivatkozásszámlálás esetén a grafikonok hasonlóan néznek ki.

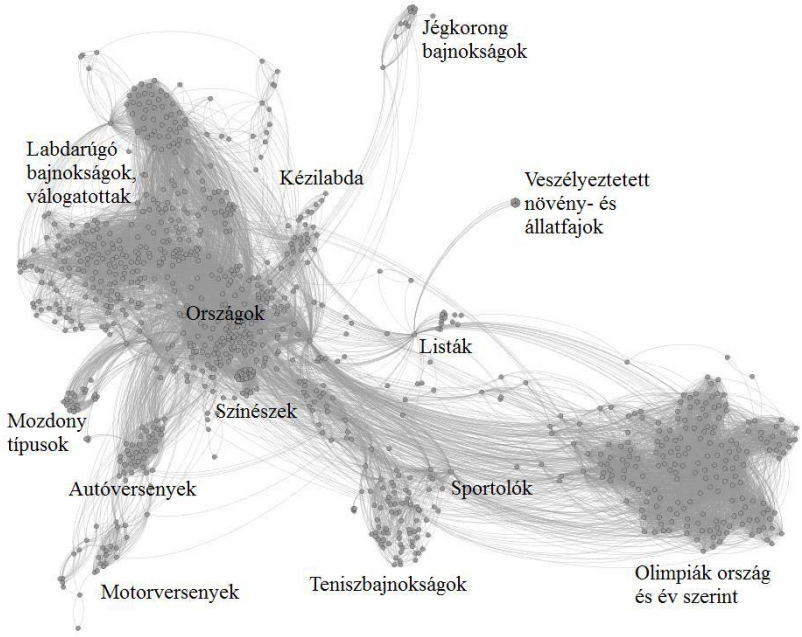
1. táblázat: Szócikkek egyszerű statisztikái. Zárójelben a hivatkozások száma.

	Befok	**Befok (többszörös)**
1.	"Kezdőlap" (316565)	"Kezdőlap" (633239)
2.	"Wikimédia Commons" (74098)	"Wikimédia Commons" (85606)
3.	"Földrajzi koordináta-rendszer" (41227)	"Budapest" (65042)
4.	"Magyarország" (33818)	"Amerikai Egyesült Államok" (59743)
5.	"Amerikai Egyesült Államok" (30819)	"Rend (rendszertan)" (59634)
6.	"Időzóna" (29207)	"Család (rendszertan)" (54555)
7.	"Egyezményes koordinált világidő" (27705)	"Magyarország" (50758)
8.	"Budapest" (27093)	"Osztály (rendszertan)" (44756)
9.	"Rendszertan (biológia)" (24351)	"Törzs (rendszertan)" (42027)
10.	"Ország (rendszertan)" (24186)	"Földrajzi koordináta-rendszer" (41292)
	Kifok	**Kifok (többszörös)**
1.	"Listák listája" (2904)	"Magyar névnapok betűrendben" (9341)
2.	"Magyar névnapok betűrendben" (2467)	"Romániai magyarok listája" (4915)
3.	"Labdarúgó-játékvezetők listája" (2191)	"2014-es labdarúgó-világbajnokság (keretek)" (4452)
4.	"Vegyületek összegképlete" (2043)	"Külföldi festők listája" (3638)
5.	"Vegyületek összegképlet-táblázata" (1955)	"Festőművészek listája" (3626)
6.	"Romániai magyarok listája" (1819)	"2010-es labdarúgó-világbajnokság (keretek)" (3608)
7.	"Katolikus szentek és boldogok listája naptár szerint" (1741)	"Római pápák listája" (3502)
8.	"Labdarúgócsapatok listája" (1656)	"2006-os labdarúgó-világbajnokság (keretek)" (3349)
9.	"A madarak nemeinek listája" (1608)	"Olimpiai érmesek listája atlétikában (férfiak)" (3281)
10.	"2014-es labdarúgó-világbajnokság (keretek)" (1411)	"Katolikus szentek és boldogok listája naptár szerint" (3242)



1. ábra: A befele (balra), illetve kifele (jobbra) irányuló hivatkozások számának eloszlása. Log-log grafikon.

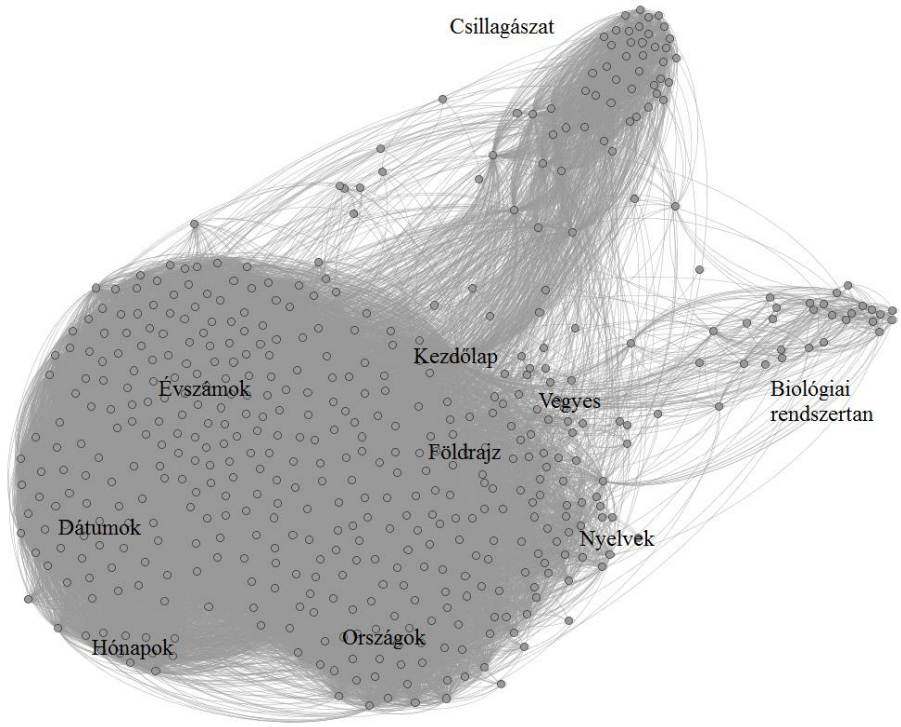
Ezek után ezeket a jellemzőket felhasználhatjuk arra, hogy részgráfokat vizualizáljunk. Bár, amint láthattuk, önmagukban ezek a számok nem tükrözik hűen egy-egy szócikk fontosságát, későbbi kutatások megalapozásaként mégis hasznosak lehetnek.



2. ábra: Az 1000 leghosszabb szócikk.

A 2. ábrán az 1000 leghosszabb szócikket láthatjuk. A gráf megjelenítése a Gephi nevű open-source programmal történt. A programnak semmilyen információja nincs az egyes csúcsokról, pusztán a gráfstruktúrát ismeri. Ennek ellenére a képen jól elkülönülő klaszterekre bomlanak az egyes témakörök. A sportösszefoglalók láthatóan dominálnak, de a lapok szépen elkülönülnek az egyes sportágak szerint. Megfigyelhető, hogy a híres sportolók, a helyszínek, illetve a listák úgynevezett „hub”-ként működnek, nagyon sok kapcsolattal kritikus összekötő láncszemeket alkotva. A 3. ábra az 500 leghivatkozottabb szócikket szemlélteti. Az itt hangsúlyos kategóriák az évszá-

mok és dátumok, de az országok itt is megjelennek. Jól elkülönülve képviseli magát a csillagászat és a biológiai rendszertan.



3. ábra: Az 500 leghivatkozottabb szócikk.

Az eddig tárgyalt felületes leíróknál tovább menve megmértük a gráf klaszterezési (klikkesedési) együtthatóját, valamint a csúcsok közötti átlagos legrövidebb távolságot. Gráfelméletben klikknek nevezzük egy csúcs szomszédságát, ha a szomszédok mind össze vannak egymással kötve. A gyakorlatban a tökéletes nagy klikkek természetesen ritkák, ezért egy aránnyal fejezzük ki, hogy egy adott csúcs szomszédjai mennyire „klikkesednek”. A teljes gráf klaszterezési együtthatója az egyes csúcsok klikkesedéseinek az átlaga. Ezt a mérőszámot összevetjük azzal, amit egy olyan gráfon mérünk, amelynek ugyanennyi csúcsa és éle van, de az élek illeszkedése véletlen. Hasonlóan járunk el az átlagos legrövidebb távolsággal kapcsolatban is. Az eredményeket a 2. táblázat mutatja be. Jól látható, hogy míg a legrövidebb távolság kicsit kisebb, a klaszterezési együttható majdnem 4 nagyságrenddel nagyobb. Ez a két tulajdonság együtt az úgynevezett „kis világ” jelenségre utal. A kis világ számtalan valós életben előforduló gráfban tapasztalható, mint például telefonhívások, szociális hálózatok, fehérje láncok stb.. Fontos megjegyezni, hogy a Wikipédia gráf irányított, ezt mindkét mérőszámnál figyelembe kell venni. Ha a linkek irányítottságát figyelmen kívül hagyjuk, a legrövidebb távolság lecsökkenne 2 alá, a „Kezdőlap” és a „Wikimédia Commons” lapnak köszönhetően. Megjegyzendő, hogy míg [3] is megál-

lapította a Wikipédia kis világ tulajdonságát, az általuk vizsgált gráf egészen más felépítésű.

2. táblázat: Klaszterezési együttható és átlagos legrövidebb távolság.

	Wikipédia	Véletlen gráf
Klaszterezési együttható	23.8%	0.003%
Átlagos legrövidebb távolság	3.92	4.67

4. A Wikipédián belüli szógyakoriságok elemzése

Az előzőekben csak a Wikipédia gráfszerkezetét vizsgáltuk. Értékes információ található azonban a szócikkek szövegében is. Ismert, hogy egy korpuszban a szavak gyakoriságának eloszlása, – mint sok más eloszlás is – hatványtörvényt követ. A hatvány kitevője a Zipf törvény szerint α=-1, azaz egy szó gyakorisága fordított arányosságban áll a sorrendben elfoglalt helyével. Lemértük a magyar Wikipédia szógyakoriságait is.

Korábban az angol Wikipédián végzett kutatás [4] szerint az első 10000 vizsgált szóra igaz a Zipf törvény, azonban az eloszlás végét egy erősebb, α=-2 lecsengés jellemzi. A magyar nyelv tekintetében is történtek mérések, azonban nem a Wikipédiát használva. Dominich és társai [3] különböző szépirodalmi korpuszokat felhasználó kutatása szerint a magyar nyelvre nem igaz a Zipf törvény, mivel a leíró paraméter értéke α=-1.21.

A mérés elvégzéséhez először elő kell készíteni a szöveget. A crawlert módosítottuk, hogy ne a linkeket szedje ki a HTML kódból, hanem a szöveges tartalmat. A szöveget ezután szavakra bontottuk, majd a toldalékok levágásával szótövesítettünk. Összeszámoltuk a szavak előfordulásait, majd csökkenő sorrendet állítottunk fel. Az eloszlást egy log-log grafikonon ábrázoltuk, majd a ráillesztett egyenes meredekségét lemérve megkaptuk az α paramétert.

A mérést elvégezve azt tapasztaltuk, hogy a 10000. szó környékén valóban történik egy törés, ahogy azt [4] is állítja. A magyar nyelv esetében azonban ez a törés közel sem tűnik olyan erősnek, mint az angolban. Az általunk mért értékek: a törés előtt α=-1.36, utána α=-1.66. Eredményeink szerint tehát az eloszlás lényegesen meredekebb, mint azt akár a Zipf törvény, akár [3] állítja.

Hivatkozások

1. Brandes, U., Kenis, P., Lerner, J., van Raaij, D.: Network Analysis of Collaboration Structure in Wikipedia. Proc. of the 18th int. conf. on World wide web (2009) 731–740
2. Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., Caldarelli, G.: Preferential attachment in the growth of social networks: the case of Wikipedia. arXiv:physics/0602026v2 [physics.soc-ph] (2006)
3. Dominich, S., Kiezer T.: Hatványtörvény, „kis világ” és magyar nyelv. Alkalmazott nyelvtudomány, Vol. 5. (2005) 5–24

4. Grishchenko V.: Bouillon project. https://web.archive.org/web/20080217050922/http://oc-co.org/?p=79 (2006)
5. Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. Proc. of the Fifth International AAAI Conf. on Weblogs and Social Media (2011)
6. Sorg, P., Cimiano, P.: Enriching the Crosslingual Link Structure of Wikipedia – A Classification-Based Approach. Proc. of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence (2008)
7. Zesch, T., Gurevych, I.: Analysis of the Wikipedia Category Graph for NLP Applications. Proc. of the TextGraphs-2 Workshop (2007) 1–8

Univerzális dependencia és morfológia magyar nyelvre

Vincze Veronika[1,2], Farkas Richárd[1], Simkó Katalin Ilona[1],
Szántó Zsolt[1], Varga Viktor[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{kata.simko, viktor.varga.1991}@gmail.com, {szantozs,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat Ebben a cikkben beszámolunk az univerzális dependencia és morfológia elveinek magyarra történő alkalmazásáról, és bemutatjuk a kihívást jelentő nyelvi jelenségeket és az azokra nyújtott megoldásainkat. A kidolgozott elvek alapján részben automatikus, részben kézi átalakítás segítségével létrehozzuk a Szeged Treebank egy újabb változatát.

Kulcsszavak: szintaxis, dependencia, morfológia

1. Bevezetés

A szófaji egyértelműsítés és a szintaktikai elemzés napjainkban is a számítógépes nyelvészet leginkább kutatott területei közé sorolható. A téma népszerűségét mutatja, hogy az utóbbi években több versenyt is hirdettek, ahol számos nyelv szövegeinek morfológiai, illetve szintaktikai elemzése volt a feladat [1,2]. A különféle nyelvű szövegeken elért eredmények összevetése azonban nehézségekbe ütközik, hiszen a különböző nyelvű adatbázisok eltérő címkekészleteket használnak, illetve más annotációs elvek alapján készültek. Ezen problémák áthidalását célozza az Univerzális Dependencia és Morfológia (UD) című, nemzetközi együttműködésben megvalósuló projekt [3].

Az UD projekt fő célja, hogy egy „univerzális", azaz nyelvfüggetlen szintaktikai és morfológiai reprezentációt dolgozzon ki, mely számítógépes nyelvészeti oldalról elősegíti a többnyelvű morfológiai és szintaktikai elemzők fejlesztését, továbbá elméleti nyelvészeti oldalról megkönnyíti a nyelvtipológiai és kontrasztív nyelvészeti vizsgálatok elvégzését. E cikkben az UD elveinek magyarra való alkalmazását mutatjuk be, különös figyelmet fordítva a speciálisan magyar nyelvi jelenségekre. Ehhez kiindulópontként a Szeged Korpusz és Treebank 2.5-ös verzióját [4] használtuk.

2. Az UD projekt

Az Univerzális Dependencia projekt célja, hogy számos nyelven ugyanazokra az annotációs elvek alapján hozzanak létre morfológiai és szintaktikai korpuszokat,

ugyanazokat az annotációs kódkészleteket használva. Ehhez hasonló egységesítési törekvések már korábban is megfigyelhetők voltak a számítógépes nyelvészetben. Például a Stanfordban kialakított függőségi címkekészletet [5] több nyelv reprezentációjában is hasznosítják. A morfológia terén az MSD kódrendszert közép- és kelet-európai nyelvekre alakították ki, többek között magyarra is [6]. Az Interset kódkészlet egyfajta közvetítő nyelvként szolgál különféle kódkészletek között, a rá épülő konverziós eljárások lehetővé teszik a kódkészletek azonos morfológiai reprezentációra történő átalakítását [7]. Rambow és munkatársai [8] a szófaji egyértelműsítést és szintaktikai elemzést szem előtt tartva megalkottak egy több nyelvre is alkalmazható morfológiai kódkészletet, míg a CoNLL-2007 verseny [9] adatai alapján McDonald és munkatársai [10] 8 fő univerzális szófajt azonosítottak. A későbbiekben Petrov és munkatársai [11] 12 fő univerzális szófajt alkalmaztak 22 nyelvre.

Az univerzális és többnyelvű morfológiai és szintaktikai kódkészletekre való törekvés legújabban az Univerzális Dependencia projektben jelenik meg. A 2015 novemberében publikált 1.2-es verzióban összesen 33 nyelv annotált adatbázisait találhatjuk meg, melyen között az angol, német, francia ugyanúgy szerepel, mint a magyar vagy a koreai.

3. Univerzális morfológia a magyarban

Az alábbiakban röviden bemutatjuk az univerzális morfológiai kódkészlet legfontosabb jellemzőit. A morfológai információt szófaji kód és jegy–érték párok formájában tároljuk. A szófajok és a jegyek halmaza kötött, azaz nincs lehetőség újabbak felvételére, ezzel szemben az értékek között szerepelhet nyelvfüggő érték, amennyiben szükséges. A jegyek között lexikai és inflexiós jegyeket egyaránt találunk: a lexikai jegyek magukra a lemmákra jellemző tulajdonságokat kódolnak, míg az inflexiós jegyek a szóalakot írják le. A jegyek lehetnek hierarchikusak is: a magyarban például a szám jegy többszörösen is megjelenhet a főnéven, így a főnév számát és a főnév birtokosának a számát két külön hierarchikus jeggyel írjuk le.

Az univerzális morfológia magyarosításakor a Szeged Korpusz 2.5-ben is használt morfológiai jellemzők nagy részét automatikusan át tudtuk konvertálni UD formátumra, ugyanakkor néhány megoldandó problémával is szembesültünk. A legnagyobb nehézséget a birtokjelölés jelentette, ugyanis a projekt addigi nyelveiben a birtokos jelölése elsődlegesen determináns segítségével valósult meg, így a magyar birtokos és birtokjel szám- és személyjelölésére külön morfológiai jellemzőket kellett felvennünk. Így a *házaiménak* szó morfológiai elemzése az alábbi lesz, ahol a Number a főnév számát, a Number[psor] és Person[psor] jegyek a birtokos számát és személyét, a Number[psed] pedig a birtok számát jelöli:

NOUN
Case=Dat|Number=Plur|Number[psed]=Sing|Number[psor]=Sing|Person[psor]=1

További sajátosságot jelentett például a determinánsok és a sorszámnevek kezelése. A sorszámnevek a Szeged Korpusz hagyományai szerint számnévként kódolandók, az univerzális morfológiában azonban melléknévként kell jelölni őket.

Ezek átkonvertálása automatikus eszközökkel történt. A mutató névmások kezelése szintén eltérő a Szeged Korpusz eredeti annotációjában és az univerzális morfológiában. Míg az *ez/az* névmások pozíciótól függetlenül névmásként kódolandók az eredeti korpuszban, addig az univerzális morfológiában a névelő előtti használat (*Olvastam azt a könyvet*) determináns, míg az önálló NP-értékű használat (*Olvastam azt*) névmás címkét követel meg. Ezek átkódolása szintén automatikus úton valósult meg.

Szintén magyar sajátosságnak számított az ige tárgyi egyeztetése, azaz a külön igei paradigma használata határozott és határozatlan tárgy mellett. Így a Definiteness jegy a magyarban az igére is alkalmazandó lett, továbbá a második személyű tárgy által megkövetelt *-lAk* morféma miatt új értéket is fel kellett venni a meglévő *határozott* és *határozatlan* érték mellé.

Az univerzális morfológia nem tartalmaz külön igekötő szófajt, az igei partikulák és igekötők szófaji besorolását az eredeti szófaj szerint végzi. Ennek megfelelően a magyarban is összeállítottunk egy táblázatot, melyben feltüntettük az igekötőként kezelt nyelvi elemek eredeti szófaját (pl. az *el* igekötőt határozószóként vettük fel, az *agyon*-t pedig főnévként). Ezt követően a táblázat alapján automatikusan átcímkéztük az igekötőket.

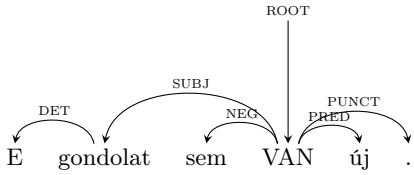
4. Univerzális dependencia a magyarban

Az univerzális dependencia címkekészletének magyarosításakor szintén automatikusan tudtuk konvertálni a függőségi viszonyok többségét, időnként a szintén rendelkezésre álló konstituens-, illetve koreferenciaannotációt is figyelembe véve. Néhány problémás jelenséget azonban részletesebben is bemutatunk.

Klasszikusan a függőségi elemzésekben a mondatok feje a főmondat ragozott igéje, viszont a kopulát és névszói predikátumot tartalmazó mondatok, és főként a fonológiailag üres kopulát tartalmazók esetén problémába ütközik ez a leírás. Mel'čuk [12] szerint azokban az esetekben, amikor a kopula csak bizonyos szám, személy, idő esetén üres (mint a magyarban), feltételeznünk kell egy zéró igealakot. A szerkezet feje ez a zéró igealak, ehhez kapcsolódik a névszói predikátum. Ezt az úgynevezett funkciószó fej elemzést követi a Szeged Dependencia Treebank is, melyet az 1. ábrán láthatunk.

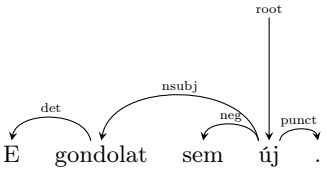
A tartalmas fej elemzés abban tér el a funkciószó fej változattól, hogy a funkciószavak helyett a tartalmas szavakat preferálja fejként. Az elmélet szerint a mondat vázát a benne szereplő tartalmas szavak és közöttük lévő kapcsolatok fejezik ki. A funkciószavak ehhez a vázhoz kapcsolódnak. A funkciószó fej elemzéstől a kopulás és adpozíciós szerkezetek kezelésében tér el. Mindkét esetben a tartalmas szó (névszói predikátum vagy az adpozíció névszói vonzata) lesz a fej a funkciószó (kopula vagy adpozíció) helyett.

Az univerzális elvek szerint a tartalmas fej elemzést kell követni. Ezen okokból automatikusan átalakítottuk a névszói és névszói-igei predikátumot tartalmazó mondatokat: mindegyik esetben a predikatív címkét viselő szót tüntettük fel fejként (vö. 2. ábra), és amennyiben szerepelt mellette kopula, az `cop` (kopula) címkével kapcsolódott a fejhez. Hasonlóképpen a névutós szerkezetekben



1. ábra: Virtuális kopula a Szeged Dependencia Treebankben.

a főnevet szerepeltettük fejként, a névutó pedig ehhez kapcsolódik `case` (eset) címkével.

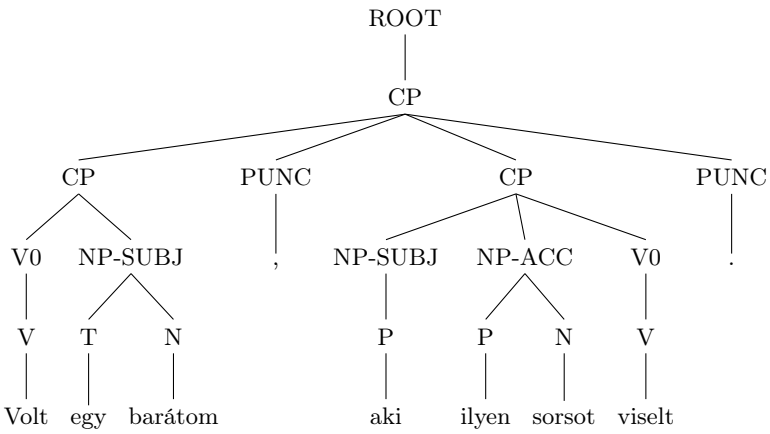


2. ábra: Tartalmas szó mint fej.

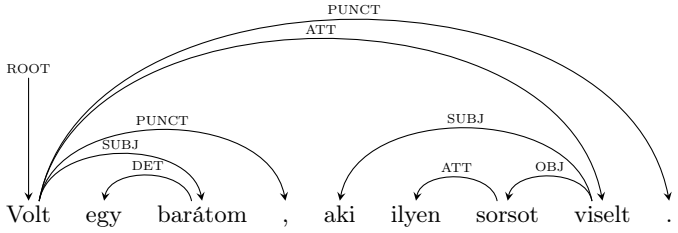
Sajátos problémát jelentett az alárendelő mellékmondatok kezelése, ahol is az univerzális elvek – a Szeged Treebanktől eltérően – megkülönböztetik az alárendelő mondatok több fajtáját is, így például az alanyi, tárgyi és határozói alárendelést, illetve a vonatkozó mellékmondatokat. Ezzel szemben a Szeged Dependencia Treebank egységesen alárendelő mellékmondatként jelölte ezeket, az altípusokat meg nem különböztetve. A Szeged Treebank konstituens változatában ezek egy része megkülönböztető jelöléssel rendelkezett, így azokat át tudtuk onnan emelni, más esetekben pedig nyelvészeti szabályok segítségével valósult meg az átalakítás, azonban az esetek egy részében az automatikus konverziót kézzel kellett javítanunk. Az alábbi példán látjuk, hogy a vonatkozó mellékmondat ATT címkét visel az eredeti dependencia treebankben (4. ábra), és a főmondat igéjéhez van kötve, a konstituens treebankben pedig nincs jelölve a CP szerepe (3. ábra).

A koreferenciaviszonyokból azonban láthatjuk (5. ábra), hogy az *aki* névmás voltaképpen az *egy barátom* szókapcsolatra vonatkozik, így vonatkozó mellékmondatról van szó, melyet a *barátom* szóhoz kell csatlakoztatni. Ezek után a szükséges átalakítások segítségével átkonvertálhatjuk a mondatot az univerzális dependencia elveinek megfelelően (6. ábra).

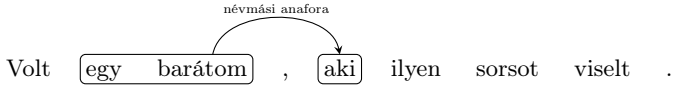
A többtagú tulajdonnevek esetében az univerzális elvek szerint az első tagot kellene fejnek jelölni. A magyarban azonban morfoszintaktikai okok miatt az utolsó tagot tekintettük fejnek, hiszen az ragozódik (*Kovács Jánosnak* vs. **Kovácsnak János*). Így ebben az esetben eltértünk az univerzális elvektől, és az utolsó tagot jelöltük fejként, míg a többi treebank esetén az első elem szerepel a szerkezet fejeként.



3. ábra: Konstituenselemzés a Szeged Treebankben.



4. ábra: Függőségi elemzés a Szeged Dependencia Treebankben.

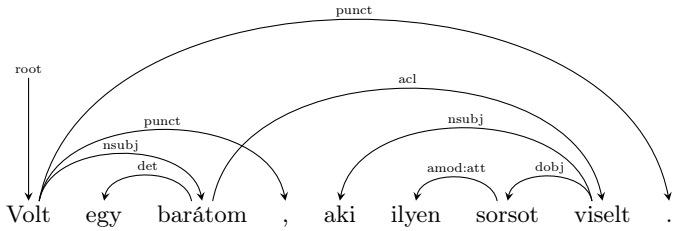


5. ábra: Koreferenciaviszonyok a SzegedKorefben.

A dativus kezelése szintén problematikusnak bizonyult. A magyarban a *-nAk* ragos főnevek számos különböző szerepet tölthetnek be a mondatban, például:

- részeshatározó: *Laci adott a padtársának egy almát.*
- birtokos: *Laci elvette a padtársának a könyvét.*
- dativus ethicus: *Nekem nehogy eladd az autódat!*
- experiens: *Nekem nagyon tetszett az előadás.*
- szemantikai alany: *Lacinak bocsánatot kellett kérnie a padtársától.*

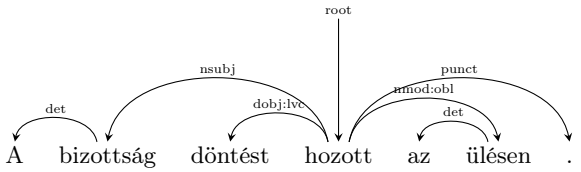
Míg morfológiai szinten a fenti alakok teljesen egybeesnek, addig szintaktikai és szemantikai szinten különféle szerepeket jelölnek. Így amellett döntöttünk, hogy míg a fenti példák morfológiai annotációját egyformán jelöljük, addig



6. ábra: Függőségi elemzés az Univerzális Dependencia Treebankben.

szintaktikai szinten elkülönítjük őket. A részeshatározót `iobj` (indirect object) címkével látjuk el, a birtokost `nmod:att` címkével (főnévi módosító), az egyéb előfordulásokat pedig `nmod:obl` címkével (főnévi vonzat) láttuk el. Természetesen ezen átalakítások kézi annotációt igényeltek, hiszen pusztán morfológiára és szintaxisra hagyatkozva ezeknek az eseteknek a nagy részében nem tudtuk volna egyszerűen és egyértelműen megvalósítani az automatikus konverziót (vö. *Nekem nehogy eladd az autódat!* és *Nehogy eladd nekem az autódat!*, ahol az első mondatban dativus ethicust találunk, a másodikban pedig részeshatározót, a két esetet egyedül a szórend különbözteti meg).

Az úgynevezett félig kompozicionális szerkezetek kezelése az 1.2-es verzióban egyelőre nem egységes a különböző nyelvek között: az UD treebankek vagy nem jelölik a szerkezeteket, vagy ha pedig jelölik, akkor ezt vagy a szokványos vonzatjelöléstől eltérő szerkezettel és/vagy speciális címkézéssel teszik [13]. A magyar az utóbbi csoportba tartozik, azaz speciális címkékkel látja el a félig kompozicionális szerkezetek tagjait. Például a 7. ábrán a `dobj:lvc` címke köti össze a szerkezet főnévi és igei tagját, azaz a *döntést* és a *hoz* szavakat, jelölve ezáltal, hogy szintaktikai értelemben ige–tárgy kapcsolatról van szó, azonban szemantikai értelemben sajátos a két összetevő viszonya. Az UD projekt legújabb egységesítési törekvései szerint ezt az elemzést terjesztjük ki a későbbiekben a többi UD treebankre is.



7. ábra: Félig kompozicionális szerkezet az Univerzális Dependencia Treebankben.

5. Szeged Univerzális Treebank

A fenti elveknek megfelelően elkészítettük a Szeged Treebank univerzális morfológiára konvertált változatát. Emellett elkészült a Népszava-alkorpusz univerzális dependenciára konvertált változata is, és folyamatosan dolgozunk a további alkorpuszok dependenciakonverzióján is: az automatikus konverzió már lezajlott, a kézi ellenőrzést igénylő annotációs lépések pedig folyamatban vannak. Az elkészült adatbázisok elérhetők az UD projekt honlapján[1].

Köszönetnyilvánítás

Szeretnénk megköszönni az Univerzális Dependencia projekt tagjainak, különösen Joakim Nivrének, Chris Manningnek és Daniel Zemannak a magyar UD treebank annotálási elveinek kialakításában nyújtott önzetlen segítségüket.

Hivatkozások

1. Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A.: Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA, ACL (2013) 146–182
2. Seddah, D., Kübler, S., Tsarfaty, R.: Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, Dublin, Ireland, Dublin City University (2014) 103–109
3. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
4. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
5. de Marneffe, M.C., Manning, C.D.: Stanford dependencies manual. Technical report, Stanford University (2008)
6. Erjavec, T.: MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation **46**(1) (2012) 131–142
7. Zeman, D.: Reusable tagset conversion using tagset drivers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) http://www.lrec-conf.org/proceedings/lrec2008/.

[1] http://universaldependencies.github.io/docs/

8. Rambow, O., Dorr, B., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K.J., Mitamura, T., Reeder, Florence, Siddharthan, A.: Parallel syntactic annotation of multiple languages. In: Proceedings of LREC. (2006)
9. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. (2007) 915–932
10. McDonald, R., Nivre, J.: Characterizing the errors of data-driven dependency parsing models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 122–131
11. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of LREC. (2012)
12. Polguère, A., Mel'čuk, I.A., eds.: Dependency in Linguistic Description. Studies in language companion series. Amsterdam Philadelphia, Pa. J. Benjamins (2009)
13. Nivre, J., Vincze, V.: Light verb constructions in universal dependencies (2015) Poszter. PARSEME 5th General Meeting.

VIII. Laptopos bemutatók

Lórum ipse: magyar vakszöveg-generátor

Nagy Viktor[1], Takács Dávid[1]

[1] Prezi
{viktor.nagy,david.takacs}@prezi.com

Lorem ipsum

A kiadványszerkesztési és weboldaltervezési munkamenetben fontos szerepe van egy olyan előnézeti képnek, amikor az oldal már rendelkezik a szedési terv grafikai jellemzőivel, de a valódi tartalom helyett töltelékszöveg jelenik meg. Ilyenkor a grafikai tervező könnyebben koncentrálhat a grafikai elemekre és a szöveg grafikai jellemzőire (betűcsalád, betűméret, sortávolság stb.). Helykitöltőként hagyományosan a *lorem ipsum*-nak nevezett szöveget és annak változatait használják, amely *Cicero: De finibus bonorum et malorum* című művének töredékeiből származik szavak felcserelésével, hozzátoldásával, halandzsaszavak bevezetésével.

A generált halandzsaszöveg fontos tulajdonsága, hogy nem csak 'ránézésre' hasonlít az imitált nyelvre, hanem statisztikai jellemzői is hasonlóak ahhoz, továbbá tartalmazza ugyanazokat a betűkapcsolatokat, amelyek az adott nyelv helyesírásában vagy tipográfiájában speciálisan kezelendők, például ligatúrákat alkothatnak.

Nem tudunk olyan *lorem ipsum*-generátorról, amely a magyar nyelvre adaptálva a fenti tulajdonságoknak megfelelő szöveget generálna, ugyanakkor a feladat számítógépes nyelvészeti eszközök alkalmazását igényli és nem triviális, ezért döntöttünk megvalósításáról.

Az alkalmazás

A demón bemutatni szánt alkalmazás képes arra, hogy a beadott paramétereknek megfelelően nagy mennyiségű szöveget generáljon elfogadható idő alatt. A nem paraméterezett döntéseket pszeudorandom módon hozza meg, azaz az algoritmus determinisztikus, egy már látott szöveg újragenerálható azonos bemeneti paraméterekkel.

A generálás során lehetőség van a szöveg bizonyos tulajdonságainak megszorítására az alapvető elvárásokon túl, mint pl. a hangzógyakoriságok finomhangolása, a szótövek gyakorisági eloszlásának alakjának meghatározása, a preferált mondatszerkezetek korlátozása.

Az alkalmazott NLP-technikák

A halandzsaszövegtől elvárjuk, hogy hangzásában feleljen meg a mai beszélt magyar köznyelv fonológiájának, ugyanakkor – a funkciószavak és néhány, véges számú egyéb szó kivételével – nem létező szótöveket tartalmazzon, de azokat szabályosan toldalékolva. Továbbá, amennyire lehetséges, grammatikus mondatokból álljon.

Ennek megvalósításához tehát szükség van egy szótőgenerátorra, amely figyelembe veszi mai szókincsünk fonológiai jellegzetességeit és lehetséges, de nem létező alakokat állít elő; egy morfológiai generátorra, amely a képzett szótövekhez, azok szándékolt szófaját ismerve ki tud választani egy megfelelő paradigmát, és generálja annak alakjait; és egy mondatgenerátorra, amely grammatikus magyar mondatokat mintaként felhasználva kvázi-grammatikus halandzsamondatokat állít össze.

Halandzsa szótőtár

A halandzsa szótöveket ngram-modellel generáljuk. A modellt az elemzett Magyar Webkorpusz lexikonján építettük, szófajonként elkülönítve. Az ngram-model véletlen karaktersorozatokat generál, amelyekből kézzel válogattuk ki a jól hangzó szótöveket.

Morfológiai generáló

A szóalak-generáló feladata az elvárt alak generálása a szótőből. A saját fejlesztésű, szabályalapú generálónk fiktív vagy ismeretlen tövekre működik. Megállapítja a tő fonológiai tulajdonságait, és azok alapján választ egy megfelelő paradigmát és meghatározza az esetleges tőváltozásokat. Ha további információra van szükség a szóalak generálásához, azt pszeudorandom módon választja meg.

Szintaktikai generáló

A szintaktikai modul a Magyar Webkorpuszból kinyert mondatsablonokon alapul. A mondatokat szófaji és morfológiai annotációk sorozataként tárolja el, megőrizve a funkciószavak alakjait. A generálási folyamat véletlenszerűen kiválaszt egy sablont, és a tartalmas szavak helyére elhelyez egy-egy véletlenszerűen kiválasztott halandzsa szót a megfelelő morfológiai alakban előállítva.

IX. Angol nyelvű absztraktok

Van's upon a Time: Copulas in Dependency Grammars

Katalin Ilona Simkó[1], Veronika Vincze[1,2]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
kata.simko@gmail.com
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Most linguistic phenomena do not have one uniform analysis that describes them perfectly, rather there are multiple frameworks offering a number of different approaches each. This paper aims to investigate the different ways the Hungarian copula *van* is described in dependency syntax.

The *van* copula is not only analysed multiple ways in theoretical syntax, but also in computational syntax [1]. The Hungarian verb *van* – as well as many other equivalent verbs in the languages of the world – has an existential as well as a copular use. The existential is used just like any other main verb, expressing being somewhere or in a certain state. The copular *van* makes up the predicate of the sentence together with a nominal predicate, but in Hungarian, the nominal predicate alone is present in the surface structure in present tense, third person sentences, the verb is absent, which is problematic for the syntactic analysis.

There are three different approaches in computational dependency syntax to describe Hungarian *van*. The function head analysis is the original annotation of the Szeged Dependency Treebank [2]; it treats all different types of *van* the same way: it is always the head of the sentence, when it is not present, a virtual node is inserted into the structure manually to take its place. The content head approach distinguishes the existential and the copular uses of *van*; the existential *van* is the head just like all other main verbs, while the copular verb is never the head: the nominal predicate is. The complex label analysis does not use virtual nodes either, but instead it marks the missing verb on the label of the nominal predicate head.

As all three annotations are available on the same text, the Népszava part of Szeged Corpus [3], we could see how each of them perform under the same conditions. We used the Bohnet parser [4] with all three analyses to measure their performance in ULA and LAS, as well as manual error analysis focusing on the errors related to the copular structure.

We found that the content head analysis works best for parsing copular structures, as it does not require manual insertion of virtual nodes, like the function head analysis and does not make the analysis overly complicated, like the complex label approach.

References

1. Simkó, K.I.: Magyar kopulás szerkezetek az elméleti és a számítógépes szintaxisban. Master's thesis, Szegedi Tudományegyetem (2015)
2. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
3. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
4. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97

Névmutató