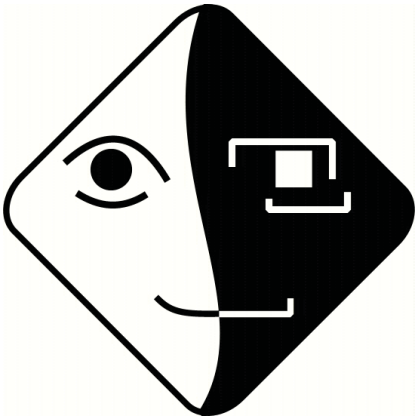
VII. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2010

Szerkesztette:

Tanács Attila
Vincze Veronika

Szeged, 2010. december 2-3.
http://www.inf.u-szeged.hu/mszny2010

ISBN: 978-963-306-075-9

Szerkesztette: Tanács Attila és Vincze Veronika
{tanacs, vinczev}@inf.u-szeged.hu

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

Nyomtatta: Planet Corp. Szolgáltató Kft.
6771 Szeged, Makai út 4.

Szeged, 2010. november

A konferenciakötet megjelenését az NKTH a TECH_08_A2/2-2008-0092 (MASZEKER) azonosítójú projekt keretében támogatta.

NKTH
Nemzeti Kutatási és Technológiai Hivatal

Előszó

2010. december 2-3-én hetedik alkalommal rendezzük meg a Magyar Számítógépes Nyelvészeti Konferenciát. Örömömre szolgál, hogy a rendezvény – az előző évek hagyományaihoz hasonlóan – fokozott érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferencia fő célja továbbra is a nyelv- és beszédtechnológia területén végzett legújabb, illetve folyamatban levő kutatási eredményeinek ismertetése és megvitatása, továbbá az esemény lehetőséget biztosít különféle hallgatói projektek, illetve ipari alkalmazások bemutatására is.

Idén a konferenciafelhívásra szép számban beérkezett tudományos előadások közül a programbizottság 46-ot fogadott el, így 32 előadás és 14 poszter-, illetve laptopos bemutató gazdagítja a konferencia programját.

Nagy örömet jelent számomra az is, hogy az idei konferencián – külön szekció keretében – kiemelt figyelmet szentelünk a szemantikus keresés terén elért eredményeknek. A számítógépes nyelvészet egyik legintenzívebben kutatott területéhez kapcsolódik a MASZEKER projekt, melynek keretében a Nemzeti Kutatási és Technológiai Hivatal is támogatja a rendezvényt. A projekt részleteiről több előadásból, poszterből és laptopos bemutatóból is informálódhat az érdeklődő közönség.

Az eddigi alkalmakhoz hasonlóan 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 díjat az MTA Számítástechnikai és Automatizálási Kutatóintézete ajánlotta fel az idei évben.

Szeretnék köszönetet mondani a programbizottságnak: Vámos Tibor programbizottsági elnöknek, valamint Alberti Gábor, Gordos Géza, László János, Prószéky Gábor és Váradi Tamás programbizottsági tagoknak. Szeretném továbbá megköszönni a rendezőbizottság (Alexin Zoltán, Almási Attila, Vincze Veronika) és a kötetszerkesztők (Tanács Attila, Vincze Veronika) munkáját is.

Csirik János, a rendezőbizottság elnöke

Szeged, 2010. november

Tartalomjegyzék

I. Információkinyerés

II. Párhuzamos korpuszok

III. Szemantika

IV. (Szemantikus) keresés

V. Beszédtechnológia

VI. Morfológia, korpusz

VII. Gépi tanulás

VIII. Poszterek, laptopos bemutatók

I. Információkinyerés

Panaszlevelek szerkezetének gépi felismerése

Bártházi Eszter[1,2⋆], Héder Mihály[3,4]

[1] MTA SZTAKI Géppel Támogatott Megértés Kutatócsoport, barthazi@sztaki.hu
[2] SZTE BTK Nyelvtudományi Doktori Iskola, Elméleti Nyelvészet Program
[3] MTA SZTAKI Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[4] Budapesti Műszaki és Gazdaságtudományi Egyetem Filozófia és Tudománytörténet Tanszék

1. Bevezetés

Kutatásunk célja egy olyan rendszer fejlesztése, amely egy adott intézmény ügyfélszolgálatára beérkezett panaszlevelek megbízható tartalmi kivonatolására képes, majd ennek továbbfejlesztéseképpen dialógus formájában képes segítségére lenni a panaszos ügyfeleknek. Természetesen nem célunk az ügyfél–ügyintéző kapcsolatot pusztán ember–gép kommunikációra korlátozni, ennek megvalósíthatósága amúgy is kétséges, a cél az, hogy az ügyintézők dolgát segítendő a gép olyan mértékű segítséget nyújtson az ügyfélnek, amelyben még biztonsággal kompetensnek tekinthető.

A hivatalos levelezésnek megvan a maga formátuma, vannak elvárások a szerkezetére, tartalmára, valamint a szókincsére vonatkozóan. A korpuszt alkotó, az Igazságügyi és Rendészeti Minisztérium ügyfélszolgálatára érkezett panaszlevelek[5] azonban olyan szabadon alkotott dokumentumok, amelyek írásakor a levélírók az esetek nagy részében nem követték a megalkotásukra vonatkozó szokásos „előírásokat". Ennek megfelelően a levélírók leveleikben nem kizárólag a megoldásra váró problémájukra szorítkoznak, a leveleket hétköznapi nyelvhasználat, a szakkifejezések rendszertelen és pontatlan használata, valamint zavaros megfogalmazás jellemzi. A zavaros megfogalmazás, valamint a többletinformációk a gép számára inkonzisztens információt jelentenek, amely jelentősen megnehezíti, ha nem ellehetetleníti a levelek megfelelő tartalmi kivonatolását. A humán nyelvhasználó számára ez nem jelent problémát, hiszen korábbi tapasztalataira hagyatkozva, valamint a diskurzuskontextus feldolgozásával képes megfelelő kontextust építeni az információk értelmezéséhez. Azt szeretnénk, ha a gép is képes lenne rendszerezni ezeket az inkonzisztens információkat azáltal, hogy kontextust, és egyúttal egyfajta interpretációt tudna rendelni hozzájuk. Ennek érdekében a panaszleveleket kisebb részekre, úgynevezett szerkezeti egységekre osztottuk, úgymint Bemutatkozás, Probléma, Lezárás stb. A szerkezeti egységek

⋆ Ezúton szeretném kifejezni hálás köszönetemet Németh T. Enikőnek és Vámos Tibornak a tanulmányhoz fűzött értékes megjegyzéseikért és támogatásukért.
[5] A korpuszért külön köszönet illeti dr. Vörös Editet, az Igazságügyi és Rendészeti Minisztérium Társadalmi Kapcsolatok Osztályának vezetőjét.

a bennük előforduló nyelvi információk kontextusaként szolgálnak, a tartalmi kivonatolás pontosságát pedig azáltal növelik, hogy jelölik, hogy milyen típusú információt hol érdemes keresni a levelekben, például a levélírót azonosító adatokat a Bemutatkozásban stb. Jelen tanulmányban azt szeretnénk bemutatni, hogy a gép számára milyen felszíni jellemzők állnak rendelkezésre, amelyek alapján a szerkezeti egységeket felismerheti.

A tanulmánnyal ugyanakkor arra is szeretnénk felhívni a figyelmet, hogy az emberek által szabadon alkotott, szerkezeti megkötésektől mentes dokumentumok kivonatolása esetén a szózsákmodell puszta alkalmazása feltehetőleg nem mindig vezet megfelelő eredményre.

A tanulmány felépítése a következő. A Bevezetést követően a 2. pontban bemutatjuk a szerkezeti egységeket, ezt követően pedig a 3. pontban azokat a jellemzőket, amelyeket a szerkezeti egységek azonosítása során figyelembe vettünk, továbbá az azonosítás sikerességének mértékéről számolunk be, valamint arról, hogy milyen elképzelésünk van még az eredmények további javítására, a hatékonyság növelésére. A 4. pontban összefoglaljuk a tanulmány eredményeit, végül a tanulmányt a Hivatkozások listája zárja.

2. A szerkezeti egységek

A levelek szerkesztésére vonatkozóan megfigyelhető, hogy az esetek nagy részében bár a levélíró feltehetőleg a legjobb tudása szerint fogalmazza meg panaszát, gyakran keveredik a hétköznapi és az egyes szakterületekre jellemző nyelvhasználat. A szakterminusok használata gyakran pontatlan, jelentése van, hogy eltér az adott szakterületen belül használatos jelentéstől. A levelek megfogalmazása zavaros, rendezetlen, és a levelek nagy hányadban tartalmaznak olyan információkat, amelyeknek a további érdemi ügyintézésben nincs szerepük. A levelek szerkezete sem egységes, sok esetben nem felel meg a hivatalos levelekkel szemben támasztott általános szerkezeti elvárásoknak. Ennek feltehetően a levélírók iskolázatlansága és alacsony társadalmi státusza az oka, bár erre vonatkozóan nem állnak adatok a rendelkezésre, mivel a korpuszt a kutatócsoport anonimizálva kapta.

A problémát az 1. ábra szemlélteti, amely egy panaszlevelet reprezentál. A panaszlevélben megfigyelhető, hogy a levél írója azt kéri, hogy értesítsék a Franciaországban élő nagybátyját megromlott egészségi állapotáról, ugyanakkor a levélben számos olyan téma is felmerül, amely a további intézkedések szempontjából irreleváns, mint ebben az esetben az arra vonatkozó információ, hogy az édesapja hogy bánt vele gyermekkorában, valamint, hogy az egyik fia a levélírás idején épp börtönbüntetését tölti.

Annak ellenére, hogy a levelek szerkezete nem egységes, mégis megfigyelhetőek olyan szerkezeti részek, amelyek a korpuszban következetesen visszatérnek, és amelyekből az egyes levelek felépülnek. A korpusz tanulmányozása alapján az alábbi tizenegy szerkezeti egységet szükséges megkülönböztetni:

1. **Megszólítás**: a levélíró azt fejezi ki valamilyen módon, hogy kinek szánja levelét, azaz kitől vár megoldást a problémájára, pl. Tisztelt [személynév/ti-

s 29. levél .
s Tisztelt Minisztérium A nevem Person édesanyám neve Name Apám Name volt .
s Az édesanyám itt halt meg Magyarországon 28 éves volt .
s Sajnos én genetikailag örököltem egy súlyos betegséget erős fájdalmaim vannak izom ízületi gyulladás mivel gyermekkorom óta depresszióban szenvedek ez súlyosbítja a betegséget .
s Az immunrendszerem elhagyott 2 éve klimaxolok ez felgyorsította a folyamatot .
s 67%-os rokkant vagyok egy középiskolás fiammal élek önkormányzati bérlakásban Gyulán .
s A másik fiam Name Gyulai börtönben van ez is nagyon elszomorítja a lelkemet .
s Nem sajnáltatni szeretném magamat csak az lenne a kérésem , hogy szóljanak a Francia Nagykövetségen ha tudnak segítsenek nekem .
s A bácsikámat Name értesítsék , hogy beteg vagyok .
s Én mostoha gyermeke vagyok ennek az országnak nem én akartam ide jönni apán hozott ide minket és nem engedte , hogy édesanyámmal visszamenjünk .
s És édesanyámnak emiatt megszakadott a szíve .
s A férjem 2000-ben rákban meghalt .
s Én mindig becsületesen éltem és sokat dolgoztam Angol Női szabó szakmámban és még betanított lakatosként is dolgoztam a Gyulai Mezőgépnél 10 évig .
s Kérem önöket legyenek szívesek megmondani nekik és sajnos nem beszélem a Francia nyelvet apám nem engedte a rossz gyermekkorom miatt nem is lett volna rá módom .
s Apám vadállat módra nevelt bennünket .
s Nekem az volt a legnagyobb bűnöm , édesanyámra nagyon hasonlítok .

1. ábra. Egy panaszlevél, amely az IM ügyfélszolgálatára érkezett

tulus/szervezet/stb.] (a szögletes zárójelben lévő kategóriák absztrakt címkéket jelölnek, amelyek a levelekben természetesen a konkrét beszédhelyzetnek megfelelő nyelvi elemekkel vannak kitöltve);

2. **Bemutatkozás**: a levélíró (ideális esetben) megadja mindazokat az adatokat, amelyek a kizárólagos azonosításához szükségesek, pl. Alulírott [személynév]. . . ;
3. **Cél**: a levélíró még a panasza ismertetése előtt röviden, néhány szóban, vagy egy mondatban kifejezi, hogy milyen területen vár segítséget, pl. Tárgy: nyugdíjügy;
4. **Probléma**: a levélíró azt a problémáját részletezi, amelynek kapcsán megoldást vár az IM részéről;
5. **Javaslat**: a Probléma alternatívája, erre abban az esetben találunk példát, ha a levélíró nem egy adott probléma megoldását várja a minisztériumtól, hanem ő maga tesz egy javaslatot valamivel kapcsolatban;

6. **Elismerés**: a levélíró elismerését fejezi ki a levél címzettjének eddigi tevékenységével, eredményeivel kapcsolatban, pl. Engedje meg, hogy gratuláljak...;
7. **Egyéb körülmények**: a levélíró a problémájához szorosan nem, vagy egyáltalán nem kapcsolódó egyéb problémáját, életkörülményeit, egészségügyi állapotát stb. ecseteli;
8. **Elvárás**: a levélíró azt fogalmazza meg, hogy milyen viselkedést, intézkedést vár el az ügyintéző részéről, pl. Kérem, hogy a fentiek alapján...;
9. **Köszönet**: a levélíró megköszöni az eddigi intézkedést, türelmet, illetve előre is megköszöni a további intézkedéseket, pl. Előre is köszönöm, hogy válaszlevelével megtisztel;
10. **Lezárás**: a levélíró egy adott formulával befejezi a levelét, pl. Minden jót!
11. **Csatolmányra hivatkozás**: a levélíró a levél egy mellékletére hivatkozik, pl. Mellékelten megküldöm kérelmemet.

A nyelvészeti pragmatikai, valamint a számítógépes pragmatikai kutatásokban egyre inkább az az uralkodó nézet, hogy a nyelvi és a kontextuális információk együttes figyelembevétele szükséges nem csak a megnyilatkozás-, de a szójelentés megalkotásában is (l. [1,2,3,4,5,6]). A panaszlevelek esetében ilyen kontextuális információ, hogy panaszlevélről van szó, amellyel egy magyar állampolgár az Igazságügyi Minisztériumhoz fordult, de a kellően pontos információkinyeréshez a fent említett, a panaszlevelekre jellemző tulajdonságok miatt ez még nem elegendő. A szerkezeti egységek kisebb, úgymond „minikontextusát" adják a bennük előforduló nyelvi kifejezéseknek.

A kontextus korábbi, statikus értelmezésével szemben annak dinamikus jellegét a pragmatikában először [1] fogalmazta meg. Eszerint a kontextust nem vehetjük előre adottnak, azt az értelmezés során kell felépíteni. A kontextus megnyilatkozásról megnyilatkozásra változik, valamint az egyes kifejezések is egymás kontextusaként szolgálnak az értelmezés során. Vannak továbbá olyan nyelvi mutatók, amelyek segítik a befogadót a megfelelő kontextus felépítésében, és ezáltal a megfelelő interpretáció megalkotásában. A szövegkörnyezet is segíti a kontextus felépítését, hiszen bizonyos kifejezések (együttes) jelenléte leszűkítheti az adott megnyilatkozás értelmezési lehetőségeit. Feltételezésünk szerint a szerkezeti egységek, azaz a „minikontextusok" felismerését ennek megfelelően bizonyos nyelvi kifejezések, azok nyelvtani tulajdonságai, valamint a szerkezetre vonatkozó heurisztikák segíthetik (l. 3. pont).

A szerkezeti egységek tematikus kontextusként szolgálnak a bennük előforduló nyelvi információ értelmezéséhez, azaz arra vonatkozóan szolgáltatnak háttértudást, hogy az adott nyelvi információ az adott kontextusban kire/mire vonatkozik, kiről/miről szól ([7]: 481). A szerkezeti egységek megmutatják, hogy a különböző, a levélíró panaszára, valamint elvárásaira vonatkozó lényegi információk a levél mely részében találhatók, egyszersmind lehetővé teszik a nem lényegi részek elhagyását.

A fent felsorolt szerkezeti egységek természetesen nem minden levélben fordulnak elő teljes repertoárjukban, és a sorrendjük is igen nagy változatosságot mutat az egyes levelekben, ugyanakkor a korpuszban újra vissza-visszatérnek.

A következő pontban a szerkezeti egységek azonosítására vonatkozó, általunk alkalmazott módszereket, valamint ezek eredményességét ismertetjük.

3. A szerkezeti egységek azonosítása: eredmények

A korpusz létrehozása a következőképpen történt: A teljes korpusz 888 panaszlevelet tartalmaz, amelynek 20%-án végeztünk kézi annotálást. Az így annotált 198 levélben jelöltük a szerkezeti egységeket, és absztrakt címkével láttuk el a következő entitásokat: SZERVEZET, TITULUS, SZEMÉLY, JOGHIVATKOZÁS és BETEGSÉG, valamint a helységneveket a rendszer automatikusan felcímkézte. Az annotálással összesen 1384 szerkezeti egységet kaptunk. Minden annotált szerkezeti egységet egy rövid dokumentumként kezelve tehát lett 1384 dokumentumunk, amelyek a szerkezeti egységeknek megfelelő tizenegy kategóriába lettek besorolva. A teljes korpuszt annotáltuk a Szegedi Tudományegyetem által fejlesztett magyarlánc [8] szoftverrel, amely a lemmákat és a POS-tageket adta meg. Egy további előfeldolgozási lépésként kiszűrtük a háromnál kevesebbszer előforduló szavakat és a szokásos stopszavakat is, majd az előfordulási értékeket tf-idf módszer szerint normáltuk. Végeredményként 2750 lemmát kaptunk.

A korpusznak többféle változatát készítettük el. Az első változat kizárólag lemmákat tartalmaz, a második szintén lemmákat tartalmaz, azzal a kiegészítéssel, hogy az igék lemmái el vannak látva igeidő (jelen vs. múlt), valamint igemód (kijelentő vs. felszólító) címkékkel. A harmadik változat szintúgy lemmákat tartalmaz, kivéve, hogy azok helyett a kifejezések helyett, amelyeket az annotálás során absztrakt címkével láttunk el, azok absztrakt címkéje szerepel. A következő kategorizáló megoldásokat alkalmaztuk:

- Döntési fák: itt a J48 és a REPTree algoritmusok szerint C4.5 [9] típusú döntési fákat hoztunk létre.
- Naive Bayes [10] kategorizáló
- SMO [11] kategorizáló

A tanításhoz általában véletlenszerűen kiválasztottuk az adatok 2/3-át, majd a maradék 1/3-dal teszteltünk, néhol azonban tízfordulós keresztvalidációt is végeztünk. A tanításokat a weka [12] keretrendszerrel végeztük.

3.1. Lexikai jellemzők

A kiindulási kísérletünk a korpusz tanítása volt lemmák szerint, minden egyéb információ hozzáadása nélkül, néhány standard módszerrel. A jellemző lemmák tekintetében a J48-as fa teljesített a legjobban, a Köszönet szerkezeti egység a kizárólagos lemmatizálással hozta a legjobb eredményt, 86%-ot,[6] a többi dimenzió csak rontott ezen. A többi szerkezeti egység tekintetében a lemmatizálás azonban önmagában kiemelkedő eredményt nem hozott.

[6] A tanulmányban szereplő százalékos értékek az F-mértékre vonatkoznak.

3.2. Lexikai és szemantikai jellemzők

A következő kísérletünk az volt, hogy a korpuszunkban az absztrakt címkével ellátott szavakat lecseréltük az absztrakt címkének a nevére, tehát például az OTP-t SZERVEZET-re cseréltük minden előfordulás esetén. Ezt a változatot tehát lemmák, illetve egyes lemmák helyett absztrakt címkék alkották. Ez természetesen némileg csökkentette a teljes korpusz lemmáinak számát. Ez a kísérlet már hozott némi javulást, ezzel értük el a legjobb eredményeket az Elvárás, a Lezárás, valamint a Cél tekintetében. Az Elvárás és a Lezárás felismerését az SMO algoritmus javította, míg a Cél felismerését a Naive Bayes, l. 1. és 2. táblázat[7].

1. táblázat. Lexikai és szemantikai jellemzőkre vonatkozó kísérletek eredményei Naive Bayes algoritmussal

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.839	0.011	0.839	0.839	0.839	0.97	Bemutatkozas
0.571	**0.007**	**0.727**	**0.571**	**0.64**	**0.931**	**Cel**
0.789	0.007	0.833	0.789	0.811	0.983	Csatolmany
0.47	0.079	0.492	0.47	0.481	0.775	Egyebkorulmenyek
0.556	0.015	0.417	0.556	0.476	0.894	Elismeres
0.794	0.078	0.726	0.794	0.759	0.891	Elvaras
0	0	0	0	0	0.94	Javaslat
0.72	0.02	0.667	0.72	0.692	0.977	Koszonet
0.632	0.051	0.343	0.632	0.444	0.95	Lezaras
0.919	0.008	0.958	0.919	0.938	0.985	Megszolitas
0.621	0.07	0.742	0.621	0.676	0.914	Problema
0.705	0.05	0.719	0.705	0.708	0.913	Weighted Avg.

3.3. Lexikai, szemantikai és grammatikai jellemzők

A lemmák és az absztrakt címkék mellett vizsgáltuk még, hogy az igeidő (jelen vs. múlt) milyen szerepet játszik a szerkezeti egységek felismerésében. Ennek előzménye az volt, hogy a korpusz tanulmányozása során feltűnt, hogy azt a problémáját, amelyre segítséget vár, a levélíró múlt időben ismerteti, míg az Egyéb körülményekben előforduló szánalmas életkörülményekre való hivatkozás gyakran jelen időben történik. Ennek oka feltehetően az, hogy a megoldásra váró probléma általában egy múltban történt eseménynek vagy események sorozatának a közvetlen következménye, és a levélíró ezt az eseményt vagy események sorozatát részletezi a problémája ismertetésekor, míg a levélíró szánalomra méltó életkörülményei esetében inkább azt hangsúlyozza, hogy azok a jelenben is fennállnak, mintegy tetézik a bajt. Ennek megfelelően a magyarlanc által igéknek

[7] Az aláhúzás a 70% fölötti eredményeket jelöli, a vastag betűs kiemelés pedig az adott szerkezeti egység azonosításában elért legjobb eredményt az összes szempontot és eszközt figyelembe véve.

2. táblázat. Lexikai és szemantikai jellemzőkre vonatkozó kísérletek eredményei SMO algoritmussal

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.839	0.02	0.743	0.839	0.788	0.977	Bemutatkozas
0.5	0.018	0.467	0.5	0.483	0.814	Cel
0.789	0.009	0.789	0.789	0.789	0.929	Csatolmany
0.576	0.104	0.475	0.576	0.521	0.841	Egyebkorulmenyek
0.556	0.015	0.417	0.556	0.476	0.804	Elismeres
0.814	**0.064**	**0.767**	**0.814**	**0.79**	**0.938**	**Elvaras**
0	0	0	0	0	0.471	Javaslat
0.76	0.013	0.76	0.76	0.76	0.963	Koszonet
0.684	**0.015**	**0.65**	**0.684**	**0.667**	**0.975**	**Lezaras**
0.919	0.023	0.883	0.919	0.901	0.982	Megszolitas
0.586	0.048	0.8	0.586	0.677	0.884	Problema
0.718	0.047	0.729	0.718	0.718	0.916	Weighted Avg.

ítélt tokenek mellé felvettük egy külön dimenzióként, hogy azok múlt vagy jelen idejűek. Az eredmények azonban nem feleltek meg a várakozásoknak, egyik algoritmussal sem hoztak jelentős javulást a puszta lemmatizáláshoz képest. További paraméter volt a szerkezeti egységek azonosításában az igemód, hiszen az elvárásait a levélíró az esetek döntő többségében expliciten és felszólító módban fogalmazza meg, míg a problémáját jellemzően kijelentő módban. Az absztrakt címkék és az igemód a Naive Bayes algoritmussal az Elismerés felismerését javította, ám ebben az esetben is igen gyenge eredményt értünk csak el, 52%-os pontosságot. Az Elismerés szerkezeti egység felismerésének tekintetében azonban ez volt a legjobb eredményünk.

Mind az igeidőnél, mind az igemódnál elmondható, hogy bár a döntési fákban előkelően közel kerültek a gyökérhez, tehát szignifikánsak, a globális pontosságon mégsem tudtak érdemben javítani.

3.4. Lexikai és szerkezeti jellemzők

Ebben az esetben a lemmatizálást kiegészítettük a szerkezeti jellemzők tesztelésével is, úgymint a szerkezeti egységek levélen belüli abszolút elhelyezkedése, valamint a szerkezeti egységek egymáshoz viszonyított elhelyezkedése. Ehhez a következő dimenziókkal egészítettük ki a szerkezeti egységeket:

- CU_START_Q1..4 -> A levél melyik negyedében kezdődik a példány
- CU_START_D1..10 -> Melyik tizedben kezdődik a példány
- CU_LENGHT_Q1..4 -> A 100 százalék hány negyedét teszi ki a szerkezeti egység hossza
- CU_LENGHT_Q1..10 -> A 100 százalék hány tizedét teszi ki a szerkezeti egység hossza
- CU_PRECEDING_Megszolitas, Bemutatkozas, stb: a példány előtt ilyen szerkezeti egységek találhatók

– CU_FOLLOWING_Megszolitas, Bemutatkozas, stb: a példány után ilyen szerkezeti egységek találhatók

Itt több lépésben végeztük kísérleteinket. Vizsgáltuk egyrészt (1) csak a lemmákat és a szerkezeti egység abszolút elhelyezkedését, (2) a lemmákat, a szerkezeti egység abszolút elhelyezkedését és hosszát, valamint (3) a lemmákat és a szerkezeti egység abszolút és relatív elhelyezkedését. Ezek a kísérletek már hoztak jelentős javulást a felismerésben. Az (1) esetben a Naive Bayes algoritmus hozta a Bemutatkozás és a Megszólítás szerkezeti egységek tekintetében a legjobb eredményeket, rendre 89%-os és 96%-os felismerési pontosságot, l. 3. táblázat. Az SMO algoritmus pedig a Csatolmány, valamint az Elvárás szerkezeti egységek felismerési pontosságát maximalizálta, a Csatolmány esetében 89%-ra, míg az Elvárás esetében 79%-ra.

3. táblázat. Lexikai és szerkezeti jellemzőkre vonatkozó kísérletek eredményei Naive Bayes algoritmussal

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.867	**0.005**	**0.929**	**0.867**	**0.897**	**0.996**	**Bemutatkozas**
0.722	0.04	0.419	0.722	0.531	0.965	Cel
0.947	0.02	0.667	0.947	0.783	0.991	Csatolmany
0.491	0.072	0.483	0.491	0.487	0.827	Egyebkorulmenyek
0.182	0.009	0.333	0.182	0.235	0.945	Elismeres
0.684	0.053	0.765	0.684	0.722	0.913	Elvaras
0	0	0	0	0	0.081	Javaslat
0.846	0.018	0.733	0.846	0.786	0.992	Koszonet
0.625	0.033	0.4	0.625	0.488	0.982	Lezaras
0.938	**0.003**	**0.987**	**0.938**	**0.962**	**0.994**	**Megszolitas**
0.686	0.068	0.771	0.686	0.726	0.915	Problema
0.722	0.042	0.737	0.722	0.724	0.933	Weighted Avg.

A (2) kísérlet az egyeshez hasonlóan javította a Megszólítás szerkezeti egység felismerését (szintén 96%) a Naive Bayes algoritmus esetében, a Bemutatkozás tekintetében azonban gyengébben szerepelt.

A (3) esetben az Egyéb körülmények és a Probléma szerkezeti egységek tekintetében értünk el maximum pontosságot. Az Egyéb körülmények esetében ez 55%-os pontosságot jelent, míg a Probléma esetében jobb az eredmény, 78%.

3.5. Összevont kategóriák

Érdekesnek találtuk megvizsgálni azt is, hogy a tanítás szempontjából melyek azok a szerkezeti egységek amelyeket könnyen azonosít a gép és melyek azok, amelyeket nehezen, és hogyan lehetne összevonni az egyes szerkezeti egységeket, ha nagy pontosságú, de kisebb felbontású kategorizálásra lenne szükség. A konfúziós mátrix megvizsgálásával az alábbi három csoport esik közel egymáshoz:

Bemutatkozás, Megszólítás, Cél, Csatolmány (BeMegCelCsat) Probléma, Javaslat, Elvárás, Elismerés, Egyéb körülmények (ProElisEgyebElvarJav) Köszönet, Lezárás (KoLe). Ezen csoportokra is lefuttattuk a kategorizáló eljárásokat a lexikai és a szemantikai jellemzők (azaz a lemmák és az absztrakt címkék) figyelembevételével, és nem meglepő módon 90% körüli eredményeket kaptunk. A legjobb eredményt az SMO algoritmussal értük el, minden csoport esetén 85% fölötti pontossággal, l. 4. táblázat.

4. táblázat. Az összevont kategóriák tesztelésének eredményei SMO algoritmussal

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.887	**0.056**	**0.881**	**0.887**	**0.884**	**0.925**	**BeMegCelCsat**
0.826	**0.012**	**0.884**	**0.826**	**0.854**	**0.948**	**KoLe**
0.942	**0.092**	**0.935**	**0.942**	**0.938**	**0.93**	**ProElisEgyebElvarJav**
0.913	0.073	0.913	0.913	0.913	0.93	Weighted Avg.

3.6. Az eredmények összegzése

Az egyes szerkezeti egységek felismerésében tehát 90% fölötti pontosságot csak a Megszólítás esetében sikerült elérni, amelynek a megformálása a legkonvencionálisabb módon történik, így ez az eredmény nem is meglepő. A konvencionális formák a mentális lexikonban egy egységként tárolódnak, előre megkomponáltak, így rutinszerűen működtethetők, különösebb kreativitást nem igényelnek a beszélőtől ([13]: 23). 80% fölötti pontosságot értünk el a Bemutatkozás, a Köszönet, valamint a Csatolmány szerkezeti egységek felismerésében, amely szerkezeti egységek (vagy legalábbis egy részük) megformálása szintén erősen konvencionális. 70% fölötti a pontossága a Problémának és az Elvárásnak, 60% fölötti a Célnak és a Lezárásnak és a leggyengébb, azaz 50% fölötti a pontossága az Egyéb körülmények és az Elismerés szerkezeti egységeknek. Könnyen belátható, hogy ezeknek a megformálása már sokkal szabadabb, az egy Lezárást kivéve, amelynek a definíciója feltehetően pontosításra szorul. A Javaslat szerkezeti egységre nem volt példa a tesztadatok között.

Az egyes szerkezeti egységekre lebontva az eredményeket a következőképpen összegezhetjük. A Bemutatkozás szerkezeti egység felismeréséhez a legnagyobb mértékben a lemmatizálás és a szerkezeti egységek abszolút elhelyezkedése járult hozzá, 14%-os javulást hozva az egyedüli lemmatizáláshoz képest. A Cél felismeréséhez az absztrakt címkék megoszlása járult hozzá a legnagyobb mértékben, a lemmatizáláshoz képest szintén 14%-os javulást eredményezett. Az Egyéb körülmények szerkezeti egység felismerését a legnagyobb mértékben a szerkezeti egységek abszolút és relatív elhelyezkedése, valamint hosszúsága és a lemmatizálás javította 11%-kal. Az Elismerés felismerésében a legjobb eredményt az absztrakt címkék és az igemód vizsgálata hozta, a puszta lemmák figyelembevételéhez képeset 10%-os javulást eredményezve, az Elvárás felismerését pedig

az absztrakt címkék vizsgálata, valamint a szerkezeti egységek abszolút és relatív elhelyezkedése a lemmatizálással együtt azonos mértékben javította a puszta lemmatizáláshoz képest, 9%-os javulást hozva. A Köszönet szerkezeti egység felismerésében önmagában a lemmatizálással kaptuk a legjobb eredményt, a Lezárás azonosításához a legnagyobb mértékben az absztrakt címkék és a lemmatizálás járult hozzá, az egyedüli lemmatizáláshoz képest 15%-kal javítva a pontosságot. A Megszólítás szerkezeti egység felismerésében a legjobb eredményt a lemmatizálással és a szerkezeti egységek abszolút elhelyezkedésével értük el, a puszta lemmatizáláshoz képest 13%-kal jobb eredményt értünk el. A Probléma szerkezeti egység felismerésében a legjobb eredményt a lemmatizálás, a szerkezeti egységek abszolút és relatív elhelyezkedése, valamint a hosszúsága hozta, 15%-os javulást a puszta lemmatizáláshoz képest. Végül a Csatolmány felismerésében a legjobb eredményeket az absztrakt címkék, valamint a lemmatizálás és a szerkezeti egységek abszolút és relatív elhelyezkedése hozta.

A szerkezeti egységek egyenkénti vizsgálati eredményeit jelentősen javították, amikor ezeket a konfúziós mátrixból kiolvasható szisztematikus tévesztések szerint három nagy csoportba vontuk össze. Ezek azonosításában a lemmákat és az absztrakt címkéket vettük figyelembe. Az azonosításban a legjobb eredményt az SMO algoritmussal értük el. Ebben az esetben viszont nem tudtuk elkülöníteni egymástól a Probléma és az Egyéb körülmények szerkezeti egységeket, amit viszont szerettünk volna.

Meglátásunk szerint növelné a találati pontosság hatékonyságát, ha a szerkezeti egységek felismerését két lépésben végeznénk. Első lépésben az összevont kategóriák felismerése történne, majd második lépésben ezeket a szerkezetiegység-csoportokat bontanánk tovább az egyes önálló szerkezeti egységekre. Külön érdemesnek tartanám megvizsgálni a következőket: ige-főnév eloszlást a KoLe csoportban, valamint igeidőt és igemódot vizsgálni a ProElisEgyebElvarJav csoportban, ezzel javítani a Probléma és Egyéb körülmények egymástól szétválasztását, valamint az Elvárás leválasztását. Emellett szükségesnek látszik kézi annotálással jelölni az egyes szerkezeti egységekre jellemző kifejezéseket, konvencionalizálódott szókapcsolatokat, ezzel segítve a szerkezeti egységek pontosabb gépi felismerését.

4. Összefoglalás

A tanulmány elsődleges célja annak bemutatása volt, hogy milyen módon nyerhetők ki a tematikus kontextusra vonatkozó információk a panaszlevelekből felszíni jellemzők alapján. A panaszleveleket tehát szerkezeti egységekre bontottuk, amelyek a bennük előforduló nyelvi információ tematikus kontextusaként szolgáltak. Ezután azt teszteltük, hogy ezeknek a szerkezeti egységeknek a gépi felismerését milyen felszíni jellemzők segítik elő. A szokásos lexikai tulajdonságok mellett vizsgáltuk még a kifejezések szemantikai és grammatikai tulajdonságait is, valamint a szerkezeti egységek egymáshoz képesti és abszolút elhelyezkedését. A jellemző lemmák vizsgálatához képest az extra dimenziók vizsgálata átlag 11%-os

javulását eredményezett, amely az esetek kb. felében szignifikánsnak tekinthető. A legjobb eredményt az összevont szerkezetiegység-csoportokkal értük el.

Célunk volt még, hogy felhívjuk a figyelmet arra, hogy szabadon alkotott szövegek esetében érdemes lehet a szerkezeti jellemzők figyelembevétele, tekintve, hogy a szövegen belüli tematikus kontextus felismerése pontosabb tartalmi kivonatolást tesz lehetővé.

Hivatkozások

1. Sperber, D., Wilson, D.: Relevance: Communication and Congition. Blackwell, Oxford (1986/1995)
2. Rott, H. In: Words in Context: Fregean Elucidations. Volume 23. (2000) 621–641
3. Bunt, H., Black, B. In: The ABC of computational linguistics. Benjamins, Amsterdam (2000) 1–46
4. Bibok, K., Németh T., E. In: Lexikai és kontextuális információk interakciója a megnyilatkozásjelentés megalkotása során. Blackwell, Szeged (2002) 335–368
5. Carston, R. In: Relevance Theory and the Saying/Implicating Distinction. MIT Press, Cambridge MA (2004) 633–656
6. Németh T., E., Bibok, K.: Interaction between Grammar and Pragmatics: The Case of Implicit Arguments, Implicit Predicates and Co-composition in Hungarian. Journal of Pragmatics **4** (2010) 501–524
7. Tátrai, S.: A kontextus fogalmáról. **128**(4) (2004) 479–494
8. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: Magyar Számítógépes Nyelvészeti Konferencia. (2009)
9. Quinlan, J.R.: C4.5: Programs for machine learning (1992)
10. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann (1995) 338–345
11. Platt, J.C.: Fast training of support vector machines using sequential minimum optimization. In: Smola (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press (1998) 185–208
12. Holmes, G., Donkin, A., Witten, I.: Weka: A machine learning workbench. In: Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia (1994)
13. Szili, K.: A kérés pragmatikája a magyar nyelvben. **126**(1) (2002) 12–30

OpinHu: online szövegek többnyelvű véleményelemzése

Miháltz Márton[1]

[1] GeoX Kft.
1034 Budapest, Bécsi út 126-128.
mmihaltz@gmail.com

Kivonat: Az OpinHu rendszer célja internetes hírportálokon, blogokon, közösségi oldalakon megjelent szövegek tartalomelemzése. A begyűjtött szövegek automatikus véleményelemzését, témaosztályozását, névelem-felismerését és az ehhez kapcsolódó statisztikákat több nyelven (ezek jelenleg: angol, magyar, német, arab, kínai) is képes elvégezni. A cikkben részletesen bemutatjuk a véleményelemzés általunk alkalmazott modelljét, valamint a felhasznált, mély nyelvi elemzésre támaszkodó, szabályalapú algoritmust. Ismertetjük a rendszer teljesítményének kiértékelésével kapcsolatos kísérleteinket is, melyeket humán annotátorokkal létrehozott szabványos adathalmazokon végeztünk el (SemEval-2007, JRC korpusz).

1 Bevezetés

Napjainkban az online írott sajtóban, de még inkább a felhasználók által generált tartalmakban (blogoszféra, közösségi hálózatok stb.) nap mint nap világszerte megjelenő szövegmennyiség új lehetőségeket teremt a számítógépes tartalomelemzés, ezen belül is az automatikus véleményelemzés (sentiment analysis, opinion mining) alkalmazása számára. A véleményelemzés célja a szövegekben megjelenő „érzelmek, értékelések, álláspontok (vélemények, hiedelmek, gondolatok, érzések, ítéletek, spekulációk) pozitív vagy negatív kifejezéseinek" [12] feltárása, amely hatékonyan felhasználható cégek, brandek, politikusok, hírességek stb. online jelenlétének és megítélésének monitorozására.

A cikkben szeretnénk bemutatni a *GeoX Kft.* és a *Zetema Ltd*[1]. kooperációjában fejlesztett *OpinHu* internetes tartalomelemző rendszer nyelvtechnológiai hátterét. A rendszer célja naponta akár több száz online forrásból több tízezer dokumentum (hírek, blog- és fórumbejegyzések, *Facebook* és *Twitter* üzenetek stb.) automatikus letöltése és feldolgozása, amely többek között a szövegek automatikus véleményelemzését, témaosztályozását, összegzését, névelemek, kulcsszavak és együtt előforduló szavak kivonatolását jelenti, több különböző nyelven (ezek jelenleg: angol, magyar, német, kínai, arab.)

A dolgozat további felépítése a következő: a következő részben röviden áttekintjük az automatikus véleményelemzés irodalmának számunkra legrelevánsabb eredménye-

1 http://zetema.co.uk

it. A 3. részben részletesen bemutatjuk a véleményelemzésben alkalmazott modellt, valamint az ezt megvalósító szabályalapú véleményelemző modult. A 4. részben bemutatjuk két kísérlet eredményeit, melyeket a véleményelemzés teljesítményének kiértékelésére és ismert rendszerek teljesítményével való összevetésére végeztünk, végül az 5. részben összefoglaljuk eredményeinket.

2 Irodalom

Pang et al. [6] felügyelt gépi tanulást alkalmazó szövegosztályozó módszereket alkalmazott filmkritikák polaritásának elemzésére. Naiv Bayes (NB), Maximum Entropy (ME) és Support Vector Machine (SVM) algoritmusokkal kísérleteztek az IMDB weboldal filmkritikáinak felhasználásával, unigramok, bigramok és melléknevek, valamint a mondatbeli pozíciók mint jegyek alkalmazásával, a negációk figyelembevételével. A legjobb eredményt SVM algoritmussal és csupán unigramok felhasználásával érték el (82.9% pontosság), 69%-os baseline érték mellett (manuálisan kiválasztott pozitív-negatív indikátorszavak számlálása.)

Pang és Lee [7] tovább tudta javítani ezt az eredményt kétszintű elemzés alkalmazásával. A filmkritikák szövegében egy osztályozó először a szubjektív mondatokat különítette el az objektív mondatoktól, ezután az előbbiekre alkalmaztak egy pozitív-negatív osztályozót. A szubjektivitásosztályozót a rottentomatoes.com oldal filmkritikáiból származó értékelő (szubjektív), ill. cselekményt bemutató (objektív) szövegrészleteken tanították, NB algoritmussal (92% pontosság). A hierarchikus véleményosztályozó pontossága 86.4%-ot ért el.

Wilson et al. [13] bemutatja az *OpinionFinder* rendszert, amely az általunk is alkalmazott érzelmikifejezés-modellben, a miénkhez hasonlóan mély nyelvi elemzésre és kifejezésszintű véleményelemzésre épül (l. 3. rész.) A rendszer a dokumentum nyelvi előfeldolgozása (szegmentálás, szófaji egyértelműsítés, tövesítés, függőségi elemzés, szubjektív lexikális elemek felismerése) után négy lépésben végzi el a véleményelemzést. Elsőként egy Naiv Bayes osztályozó megkülönbözteti a szubjektív és objektív mondatokat [11]. Ezután egy szabályalapú osztályozó ismeri fel a beszédaktusokat és közvetlen szubjektív kifejezéseket (pl. „mondta", „véleménye szerint", „attól tart" stb.) A vélemények forrásának felismerését egy Conditional Random Field (CRF) szekvenciafelismerő modell és egy mintafelismerő algoritmus kombinációja végzi el. Végül a véleménykifejezések (sentiment expressions) felismerését és ezekben a pozitív-negatív polaritások felismerését 2 újabb osztályozó végzi el [12].

Godbole et al. [4] egyszerű szabályalapú megközelítést alkalmaz, saját fejlesztésű érzelmi szótárak felhasználásával. A szótárakat automatikusan, néhány kézzel megadott kiinduló (seed) pozitív-negatív fogalom és WordNet [3] szinonimáik és antonimáik segítségével hozták létre. Részben a cikkben bemutatott rendszerhez hasonlóan (l. 3. rész), Godbole et al. [4] a szövegben a felismert entitásokra vonatkozó érzelmeket az entitással egy mondatban előforduló felismert pozitív-negatív kifejezések számlálásával határozzák meg, a negációs kifejezések figyelembevételével, valamint névmási anafora- és koreferenciafeloldás alkalmazásával.

3 Az OpinHu rendszer

Ellentétben a szövegosztályozó algoritmusokat alkalmazó megközelítésekkel ([6, 7]), véleményelemző rendszerünk a mondatok alatti szinten, kifejezéseken működik, így teljes dokumentumok véleményértékelését a bennük található szubjektív (érzelmi) kifejezések azonosításával és összegzésével lehet elérni, hasonlóan Wilson et al. [13] és Godbole et al. [4] munkájához.

Modellünkben minden érzelmi kifejezésben azonosítható egy forrás (a vélemény képviselője) és egy célpont (akire vagy amire a vélemény irányul), valamint meghatározhatók polaritás- (pozitív, negatív vagy semleges/kiegyenlített) és intenzitásértékek (a polaritástól függetlenül mennyire erős érzelem jelenik meg). A vélemények célpontjait előre meghatározott kulcsszavak halmazával detektáljuk. A szövegekben felismert érzelmi kifejezések polaritását egy speciális érzelmi lexikon elemeinek segítségével, valamint a kontextusban felismert polaritásmódosító elemek (pl. tagadás) figyelembevételével számítjuk ki.

Az ismert érzelmi kifejezésekhez a priori (tehát a kontextusban módosítható) polaritást angol nyelvre a General Inquirer (GI) [8] közismert pszichológiai tartalomelemző szótár használatával társítottunk. Magyar, német, arab és kínai nyelvekre érzelmi lexikonhoz az angol GI szótár pozitív-negatív besorolású tételeinek fordításával és szinonimákkal való bővítésével jutottunk (1. táblázat).

1. táblázat: Az érzelmi lexikonokban található címszavak száma angol, magyar, német, arab és kínai nyelvekre.

Nyelv	Pozitív	Negatív	Összesen
Angol	2 291	4 102	6 393
Magyar	6 034	8 438	14 472
Német	2 242	3 406	5 648
Arab	1 438	1 665	3 103
Kínai	2 812	8 180	10 992

A feldolgozott dokumentumok érzelmi elemzését két szempontból végezzük el:

- Célponthoz kapcsolódó érzelem (target sentiment), melyet csak a kulcsszavakhoz kapcsolódó szubjektív kifejezések alapján számítunk. Az érzelmi kifejezések és a kulcsszavak közötti kapcsolatok azonosítására a rendelkezésre álló nyelvi erőforrások függvényében 2 különböző algoritmus egyikét használjuk.

- Általános érzelem (overall sentiment), melyet a dokumentumban található összes érzelmi kifejezés feldolgozásával számítunk. Célja a szövegben található összes érzelem kvantifikációja, nem csak a meghatározott célponthoz kapcsolódó véleményeké.

A szöveg nyelvétől függően különböző szintű nyelvi feldolgozást tudunk elvégezni. Jelenleg minden, a rendszer által kezelt nyelven először az alábbi előfeldolgozási lépéseket hajtjuk végre (1. szint):

- Szegmentálás (mondatok és szavak)
- Szófaji egyértelműsítés, szótövesítés
- Kulcsszavak, polaritást módosító és érzelmi kifejezések annotációja.

Angol nyelvre ezen felül a következő feldolgozási lépéseket tudjuk végrehajtani (2. szint):

- Névelem-felismerés 29 előre meghatározott kategóriával (Inxight ThingFinder[2]), pl. *ADDRESS, ADDRESS_INTERNET, CITY, COMPANY, COUNTRY, CURRENCY, DATE* stb.
- Függőségi elemzés a mondatok szintaktikai viszonyainak azonosítására (Stanford Parser [5])
- Koreferenciaazonosítás: az ugyanarra az entitásra referáló kifejezések azonosítása (pl. Barack Obama, President Obama, Mr. Obama, he stb.), valamint a felhasználó által megadott kanonikus névalakkal való helyettesítése (OpenNLP[3]).

Az 1. szintű nyelvek (magyar, német, kínai, arab) esetében az érzelmek felismerése a durva, de robusztus szózsák (bag-of-words) algoritmussal működik. Ennek lényege, hogy feltételezzük, hogy ha egy érzelmi kifejezés és egy kulcsszó együtt fordul elő egy mondatban, akkor az érzelem a célpontra irányul [4].

A 2. szinten feldolgozható nyelveken (jelenleg: angol) a szintaktikai elemzés kifinomultabb megközelítést tesz lehetővé, amellyel magasabb pontosság érhető el. A rendszer 16 meghatározott függőségi minta segítségével próbál a mondatokban a felismert érzelmi kifejezések és a kulcsszavak között kapcsolatot találni, ezeket a 2. táblázatban foglaltuk össze.

Angol nyelvű dokumentumokra a véleményelemzésen túl a tartalomelemző rendszer az alábbi elemzési feladatokat képes elvégezni:

- automatikus témaosztályozás (Autonomy Idol[4])
- automatikus tartalomkivonatolás: a szöveg rövid összefoglalása a legrelevánsabb 5 mondat segítségével
- kulcsszó-előfordulási statisztikák: a célpontok online jelenlétének időbeli változásának figyelésére, különböző témakörökben vagy forrásokban (blogok, közösségi oldalak stb.)
- a kulcsszavakhoz tartozó névelemek vagy egyéb szavak (tartalmas szavak, pozitív-negatív kifejezések) kinyerése a célpontokhoz kapcsolódó egyéb fogalmak címkefelhőkben (1. ábra), energiatérképeken stb. történő ábrázolásához.

2 © Inxight Software, Inc.

3 http://opennlp.sourceforge.net/

4 http://www.autonomy.com/

2. táblázat: Függőségi minták az angol szövegek érzelmi elemzéséhez (*k*: kulcsszó, *s*: szubjektív (pozitív-negatív) kifejezés).

Függőségi viszony	Magyarázat
nsubj(*k*, *s*)	*k* az *s* aktív ige vagy névszói állítmány alanya
nsubj(*s*, *k*)	*s* a *k* névszói állítmány alanya
nsubjpass(*s*, *k*)	*k* az *s* passzív ige alanya
dobj(*s*, *k*)	*k* az *s* ige tárgya
agent(*s*, *k*)	*k* az *s* passzív ige ágense
amod(*k*, *s*)	*s* melléknév a *k* főnév módosítója
appos(*k*, *s*)	*s* főnév a *k* főnév appozíciós módosítója
appos(*s*, *k*)	*k* főnév az *s* főnév appozíciós módosítója
infmod(*k*, *s*)	*s* infinitivuszi ige a *k* főnév módosítója
nn(*k*, *s*)	*k* és *s* összetett főnevet alkotnak
nn(*s*, *k*)	*k* és *s* összetett főnevet alkotnak
partmod(*k*, *s*)	*s* a *k* igenévi módosítója
poss(*s*, *k*)	*k* az *s* birtokosa
prep_*(*s*, *k*)	*k* az *s* prepozíciós módosítója
rcmod(*k*, *s*)	*s* ige a feje a *k*-t módosító mellékmondatnak
xsubj(*s*, *k*)	*k* a vezérlő alanya annak a mellékmondatnak, amelynek *s* ige a feje

3.1 Alkalmazások

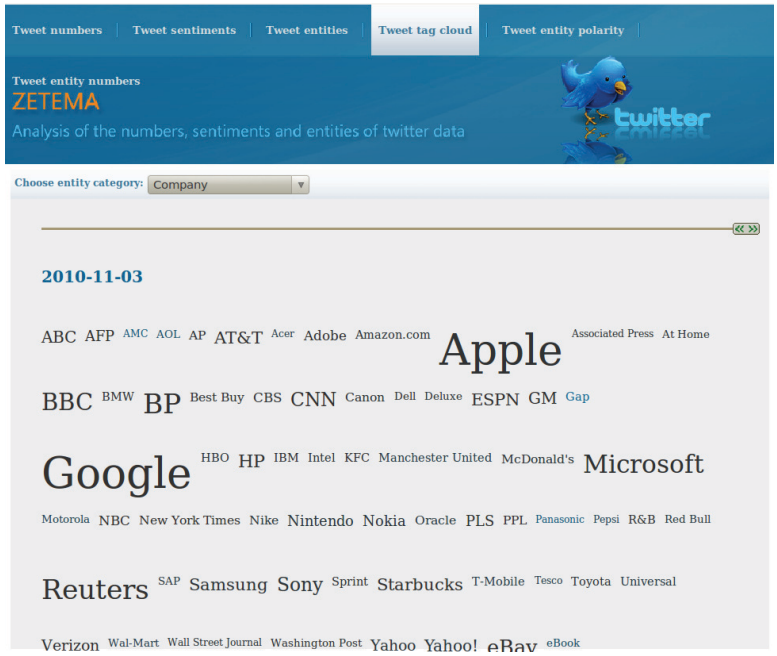
Az OpinHu rendszerhez jelenleg három felhasználói felületet (dashboard) készítettünk el, melyek közül kettő publikusan kipróbálható. Az első dashboard a Twitter közösségi oldalon megjelenő nyilvános üzenetek (tweetek) elemzését mutatja be (Twitter Streaming API, Gardenhose (~5%) minta, napi 8-12 millió tweet[5]). A weboldalon[6] lehetőségünk van időszakokra, illetve dátumokra lebontva megvizsgálni a megjelent üzenetek számát, azok polaritását, az üzenetekben megjelenő főbb entitáskategóriákat, illetve a kapcsolódó fogalmakat (1. ábra).

Második demonstrációs dashboard-unk[7] az USA 2010 novemberi időközi kongresszusi, szenátusi és kormányzói választására készült. Több mint 300 politikai témával foglalkozó (angol nyelvű) blogon, illetve a Facebook Graph API segítségével a Facebook közösségi oldal nyilvános státuszüzeneteiben vizsgáltuk 2010 május óta az összes jelölt megítélését. A felületen nyomon követhetjük ebben az időszakban az adott célpontok említésének, illetve a forrásainkból összegzett megítélésének változását (2. ábra).

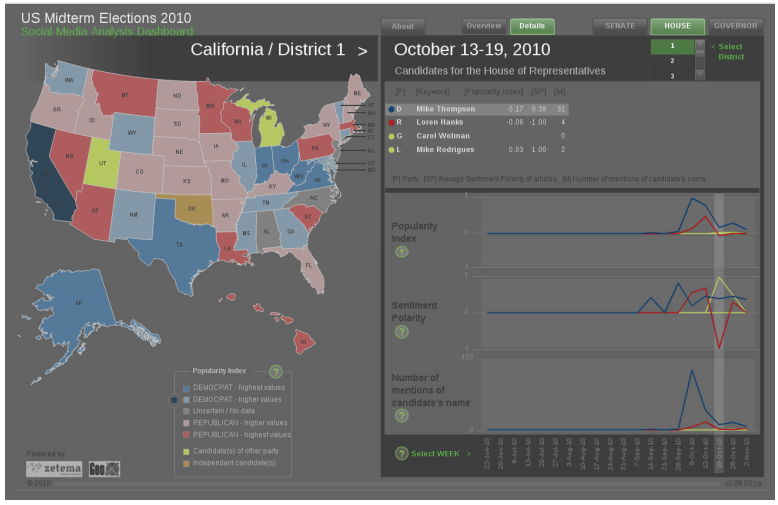
5 http://dev.twitter.com/pages/streaming_api_concepts

6 http://twitter.zetema.co.uk/

7 http://usa.zetema.co.uk/



1. ábra. A Twitter közösségi oldal publikus üzeneteiben megjelenő cégnevek ábrázolása címkefelhőben a Twitter dashboard-on (képernyőkép).



2. ábra. A USA 2010-es időközi választásokra készült véleményelemző rendszer dashboardjának képernyőképe.

4 Kiértékelés

Az angol nyelvű véleményelemző rendszer kiértékelésére eddigi munkánk során két kísérletet végeztünk el. Az első kísérlethez az EC Joint Research Center[8] (JRC) által annotált idézeteket használtuk fel [2], ami lehetővé tette a célpontokra irányuló véleményelemzés kiértékelését. A második vizsgálatban a SemEval-2007[9] 14-es feladatának („Affective Text") [9] standard annotált adathalmazát használtuk fel, így lehetőség nyílt rendszerünk teljesítményének más rendszerekkel való összehasonlításra is.

4.1 JRC korpusz

A korpusz 1590 db angol nyelvű, különböző hírekből származó rövid (1-3 mondatos) idézetet (függő beszéd) tartalmaz. Minden idézethez kézzel azonosítottak egy célpontot (személy vagy intézmény), amely az idézet szövegében szerepel, majd 2 annotátor kézzel megjelölte, hogy az idézet a célpontra nézve pozitív, negatív vagy semleges polaritású. A munka során külön figyelmet szenteltek a pozitív-negatív érzelmek és a jó-rossz hírek fogalmának elkülönítésének [1].

A rendelkezésre álló korpuszon először több adattisztítási lépést kellett elvégeznünk. Az 1590 idézetből csupán 1290 esetében volt egyetértés a 2 annotátor között (ez 81.13%-os egyetértési arányt jelent), így a továbbiakban csak ezekkel foglalkoztunk. Mivel a célpontok nem az idézetek szövegében bejelölve, hanem minden egyes tételhez külön megadva álltak rendelkezésre, kísérletet kellett tennünk ezek azonosítására az idézetek szövegében. A megadott célpontok sajnos nem minden esetben voltak pontosan megtalálhatók a szövegekben, sok esetben valamilyen más névváltozatot, rövidítést stb. használt az eredeti szöveg, így egy egyszerű heurisztikus algoritmussal próbáltunk meg minél több névváltozatot felismerni (nevek tokenalapú részsorozatai, betűszavak generálása, kötőjelek és szóközök variálása stb.). Ezzel a módszerrel végül 1249 db idézetben sikerült az eredeti célpontot megjelölni. Utolsó lépésben azok közül az idézetek közül, amelyek többször is szerepeltek a korpuszban (feltehetőleg más-más hírforrásokból idézve) egyetlen példányt tartottunk csak meg, így végül 1136 db, célponttal és polaritással annotált idézetet tudtunk felhasználni a kiértékeléshez.

Kíváncsiak voltunk az OpinHu rendszerben alkalmazott mindhárom véleményelemző algoritmus teljesítményére: általános érzelem szózsák algoritmussal (AZ), célpontra irányuló érzelem szózsák algoritmussal (CZ), célpontra irányuló érzelem függőségi elemzéssel (CF). Mivel a semleges polaritású idézetek aránya igen magas (66.81%) volt, az algoritmusok teljesítményét kétféle módon is kiértékeltük. Első lépésben egyszerű pontosságot (accuracy) mértünk a pozitív-negatív-semleges osztályozáshoz képest. Semlegesnek a [-0.1, 0.1] intervallumba eső polaritást feltételeztük. A 3. táblázatban láthatók ennek a vizsgálatnak az eredményei.

8 http://langtech.jrc.ec.europa.eu/JRC_Resources.html

9 http://nlp.cs.swarthmore.edu/semeval/

3. táblázat: A három algoritmus egyszerű pontossága (accuracy) a pozitív-negatív-semleges osztályozáshoz képest a JRC korpuszon.

Algoritmus	Pontosság
Baseline (mindig semleges)	66,81%
AZ	39,88%
CZ	44,01%
CF	64,88%

A második vizsgálatban elkülönítettük a semleges polaritású cikkeket, és csak a pozitív-negatív besorolású tételeket vizsgáltuk (377 idézet). Ezeken az adatokon pontosság (precision) és fedés (recall) értékeket számítottunk. Pontosság alatt a rendszer által (a manuális annotációhoz képest) helyesen megadott polaritású idézetek arányát értjük azokban az esetekben, ahol a rendszer nem semleges ([-0.1, 0.1] intervallumba eső) polaritást adott vissza. Fedés alatt a rendszer által eltalált esetek arányát értjük az összes 377 idézethez képest. Az eredmények a 4. táblázatban láthatók.

4. táblázat: A három algoritmus pontossága (precision), fedése (recall), valamint az F-mérték a pozitív-negatív osztályozáshoz képest a JRC korpuszon.

Algoritmus	Precision	Recall	F1
AZ	71,01%	57,83%	63,74%
CZ	71,10%	54,11%	61,45%
CF	52,17%	6,40%	11,35%

A 3. táblázatból látható, hogy amennyiben használjuk a semleges kategóriát, a függőségi elemzést használó algoritmus (CF) jobb, mint a szózsák algoritmus, továbbá a célpontra irányuló érzelemfelismerés (CZ) jobban közelít a gold standardhoz, mint az általános érzelemfelismerés algoritmusa (AZ). Ugyanakkor fontos észrevenni, hogy egyik algoritmus sem tudta meghaladni a relatív magas baseline értéket (66.81% semleges polaritású idézetek aránya 81.13%-os humán egyetértési ráta, tehát lehetséges felső határ mellett).

Csak a pozitív-negatív polaritású idézeteket használva azonban megfordul a kép (4. táblázat). A szózsák algoritmus jobban teljesít, mint a függőségi elemzést használó algoritmus, továbbá a célpontra irányuló, szózsák algoritmust használó módszer teljesítménye (F-mérték) rosszabb, mint az általános érzelemfelismerő, szózsák algoritmust használó módszeré.

4.2 SemEval-2007 korpusz

A 2007-es SemEval verseny 14-es feladata számára 1000 db angol nyelvű címet (híroldalak, újságok) láttak el a 6 alapérzelem, valamint a pozitív-negatív dimenzió mentén kézi annotációval. Utóbbit egy [-100..100] intervallumban értelmezett pontértékkel adták meg, ahol 0 semleges érzelmet, -100 erősen negatív, 100 pedig erősen pozitív érzelmet jelent. A munkát 6 annotátor végezte, közöttük az egyetértés a Pearson egyetértési mértékkel számítva 78.01% volt [9]. A verseny számára meghatároztak egy ún. durva felbontású kiértékelő halmazt is, melyben a [-100..100] intervallumba

eső értékeket leképezték a {-1, 0, 1} halmazra, a (-50..50) semleges intervallum használatával. A versenyben részt vevő rendszerek teljesítményének értékelésére – hasonlóan ahhoz, ahogy mi a JRC korpusszal tettük – meghatározták a pontosságot (accuracy) a pozitív-negatív-semleges osztályozás, valamint a pontosságot és a fedést (precision és recall) csak a pozitív és a negatív besorolású tételek esetében is (410 cím).

Mivel ebben az esetben nem volt annotált célpont, így csak az általános érzelmet felismerő, szózsák algoritmust alkalmazó módszer teljesítményét tudtuk értékelni. Az 5. táblázatban látható a SemEval-2007-ben résztvevő rendszerek és a mi algoritmusunk teljesítményének összevetése 3 kategória (pozitív-negatív-semleges) használatával (accuracy), illetve 2 kategória (pozitív-negatív) használatával (precision, recall, F-measure). A baseline algoritmus az első esetben a leggyakoribb, semleges osztály konstans hozzárendelését jelentette.

5. táblázat: A SemEval-2007 résztvevői és a cikkben bemutatott rendszer teljesítményének összevetése a SemEval-2007 „Affective Text" feladat adathalmazán durva felbontású (coarse-grained) kiértékeléssel.

Rendszer	Accuracy	Precision	Recall	F1
CLaC	55.10%	61.42%	9.20%	16.00%
UPAR7	55.00%	57.54%	8.78%	15.24%
SWAT	53.20%	45.71%	3.42%	6.36%
CLaC-NB	31.20%	31.18%	**66.38%**	42.43%
SICS	29.00%	28.41%	60.17%	38.60%
OpinHu	**55.20%**	**90.25%**	51.95%	**65.94%**
Baseline	59.00%	n.a.	n.a.	n.a.

A rendszerünkben használt szózsák algoritmus mind pontosság (accuracy), mind F-mérték tekintetében jobban teljesít a SemEval-2007 versenyben legjobban teljesítő rendszerekhez képest. Az accuracy érték tekintetében a különbség nem szignifikáns (0.10%), a precision érték viszont kimagaslóan felülmúlja a legjobb rendszerét (28.83% eltérés), így az F-mérték is szignifikánsabban magasabb (23.51% különbség).

5 Összegzés

Bemutattuk az OpinHu tartalomelemző rendszer véleményelemző komponensét, amely a nyelvi erőforrások függvényében bag-of-words algoritmust, illetve függőségi viszonyokon alapuló mintakeresést alkalmaz. Az angol nyelven működő rendszer teljesítménye a SemEval-2007 kiértékelő adathalmazon szignifikánsan meghaladta a korábbi rendszerek teljesítményét.

Bibliográfia

1. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010). Valletta, Malta (2010) 2216–2220
2. Balahur-Dobrescu, A., Steinberger, R.: Rethinking sentiment analysis in the news: from theory to practice and back. In: Workshop on Opinion Mining and Sentiment Analysis (WOMSA), held at the 2009 CAEPIA-TTIA 13th Conference of the Spanish Association for Artificial Intelligence. Sevilla, Spain (2009) 1–12
3. Fellbaum, C. (szerk.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
4. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale Sentiment Analysis for News and Blogs. In: Proceedings of ICWSM-2007. Boulder, Colorado, USA (2007)
5. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP-2002 (2002) 79–86
7. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL-2004 (2004)
8. Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge, MA (1966)
9. Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07). Association for Computational Linguistics, Morristown, NJ, USA (2007) 70–74
10. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the ACL (2002)
11. Wiebe, J., Riloff, E.: Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In: LNCS Computational Linguistics and Intelligent Text Processing (2005) 486–497
12. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of HLT/EMNLP 2005 (2005)
13. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: OpinionFinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP 2005 Interactive Demonstrations (2005)

Videókhoz kapcsolódó kiegészítő információk többnyelvű keresése a Wikipédia segítségével

Gyarmati Ágnes[1], Gareth J.F. Jones[1]

[1] Dublin City University, Centre for Digital Video Processing,
Dublin 9, agyarmati@computing.dcu.ie

Kivonat: (ld. a törzsben)[1]

1 Rövid kivonat

Az egyre nagyobb mennyiségben szabadon elérhető digitális tartalmak (hang- és videófelvételek) akkor válnak igazán értékessé, ha a felhasználók könnyen megtalálhatják a számukra érdekes tartalmakat, részleteket, azaz hatékonyan lehet köztük és bennük keresni.

Az 1990-es évek közepén a TREC kutatói fórum egyik céljaként tűzte ki a hangzó szövegben való keresés hatékonyságának fejlesztését. Ehhez amerikai rádiós híradásokat használtak, és csupán pár év elteltével a próbálkozásokat sikeresnek, a problémát gyakorlatilag megoldottnak tekintették [1].

Később azonban felmerült az igény, hogy a viszonylag zárt szókincsű, gondozott beszéddel felolvasott szöveget rögzítő jó minőségű stúdiófelvételeken túl az élőbeszéddel is érdemben foglalkozzanak az információ-visszakeresés területén. Erre például a CLEF fórum beszéd-visszakereső modulja vállalkozott [3].

A VideoCLEF 2009 kiírása más szemszögből nyújtott vizsgálati lehetőséget élőbeszéd és keresés kapcsolatához [2]. Az úgynevezett „Összekapcsoló feladat" (Linking Task) egy kulturális dokumentumműsor egyes rövid, csupán pár másodperces részleteihez keres tartalmilag kapcsolódó oldalakat a Wikipédiában. A feladat nehézségét a híranyagoknál tapasztaltaknál spontánabb beszéd és változatosabb tartalom mellett egy nyelvi csavar adja: a dokumentumsorozat holland nyelvű, míg a linkelésnél az angol Wikipédiából kellett keresni a legmegfelelőbb oldalakat – ezáltal szimulálva egy haladó nyelvtanulót, aki már rendelkezik kellő nyelvismerettel, hogy idegen nyelven is érdemes legyen tévét, videót, műsorokat néznie, de még szükségét érezheti, hogy a számára érdekes vagy magyarázó kiegészítő információkhoz anyanyelvén, vagy legalábbis egy általa magasabb szinten beszélt (értett) harmadik nyelven szeretne hozzájutni (esetünkben angolul).

Alapvetően két különböző megközelítés kínálkozik. Az egyik a holland Wikipédiát veszi alapul, ott végzi a keresést az eredeti holland szöveg felhasználásával, majd a

[1] Ez a kutatás a Science Foundation Ireland (SFI) által támogatott Improving Indexing for Search of Spontaneous Conversational Speech (IISSCoS) projekt keretében zajlik.

Wikipédia saját, nyelvek közötti linkjeit követve jut el a relevánsnak tartott angol oldalakig. A másik előbb gépi fordítással angol nyelvű szöveget gyárt a műsorok holland átirataiból, majd közvetlenül az angol Wikipédiában keresi a megfelelő oldalakat. Az előadás e két módszert kívánja tárgyalni, bemutatni előnyeiket, hátrányaikat, különféle változataikat, melyek pl. az adatok használatában, keresőkifejezések generálásában is különbözhetnek. Eddigi eredményeink azt mutatják, hogy a keresés szempontjából nem hatékony az átiratot automatikusan lefordítani, hanem inkább a forrásnyelven érdemes a keresést végezni.

A felvázolt módszerek bármely nyelvre, nyelvpárra alkalmazhatók, így a magyarra is, akár tárgy-, akár célnyelvként, feltéve, hogy létezik az adott (a videóban használt nyelvre) automatikus beszédfelismerő program, és hogy a kívánt nyelvű Wikipédia kellően gazdag.

Bibliográfia

1. John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The TREC spoken document retrieval track: A success story. In Text Retrieval Conference (TREC) 8, 16–19, (2000).
2. M. Larson, E. Newman and G. J. F. Jones, Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment, In Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation, Korfu, Görögország, (2009). Megjelenés alatt.
3. D. W. Oard, J. Wang, G. J. F. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang and I. Shafran, Overview of the CLEF-2006 Cross-Language Speech Retrieval Track, In Proceedings of the CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation, Alicante, Spanyolország, Springer (2006)

DBPedia magyar nyelvű szövegek elemzéséhez

Németh Bottyán[1], Vándor Tamás[2]

[1]BME, TMIT,
Budapest, Magyar tudósok körútja 2., e-mail:bottyan@gmail.com
[2]Webra International Kft.,
Budapest, Francia út 33. IV/1, e-mail:tamas.vandor@webra.hu

Kivonat A publikációban egy olyan szövegannotáló rendszer kerül bemutatásra, ami a Wikipédiát is felhasználja az általa ismert fogalmak körének bővítésére. Ehhez szükséges volt a Wikipédia formális ábrázolása, amire eddig az egyik legsikeresebb kísérlet a DBPedia projekt. A DBPedia magyar változatának elkészítése után ezt a tudásbázist használtuk fel szövegek szemantikus annotálására, több más nyelvészeti eszközzel és doménspecifikus ontológiákkal kiegészítve. Így egy komplex rendszer jött létre, ami képes magyar nyelvű szövegek elemzésére és a benne található szavak szemantikus annotálására. A Wikipédiára épülő tudásbázisnak köszönhetően nagy lefedettséget, míg a formális ábrázolás miatt megfelelő pontosságot sikerült elérni. [1]

Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, ontológia, DBPedia, Wikipédia, névelem-felismerés

1. Bevezetés

Az ismertetésre kerülő magyar nyelvű szemantikus szövegannotáló rendszer elkészítése során több már korábban meglévő eszközt integráltunk egy egységes keretrendszerbe, kiegészítve saját fejlesztésű modulokkal. A munka során elért egyik legjelentősebb eredmény, hogy létrejött egy szabadon hozzáférhető formális tudásbázis a Wikipédia alapján, a DBPedia magyar változata. Ez a tudásbázis alkotja a rendszer fogalmi adatbázisának magját.

A cikkben bemutatásra kerülő rendszer egy átfogóbb projekt részét képezi. A hosszabb távú cél egy intelligens ügyfélszolgálati megoldás, amely képes egy szűkebb tárgyterületen gyakran előforduló problémákat automatikusan megválaszolni. Az ehhez vezető úton első lépésként egy olyan átfogó tudásbázis félautomatikus építését tűztük ki célul, amely kiegészítve kisebb doménspecifikus ontológiákkal az ügyfélszolgálatra érkező levelekben található fogalmakra nagyarányú lefedettséget biztosít. A következőkben az itt felhasználandó ontológia felépítéséről és első alkalmazásáról fogunk beszámolni egy magyar nyelvű szemantikus szöveganntáló rendszer keretében.

[1] A cikkben tárgyalt rendszer a "KMOP-2007-1.1.1 Intelligens Multi-Modális Tudásközpont" projekt keretében jött létre.

2. Kapcsolódó munkák

Mivel még nincsen igazán hatékony tanuló algoritmus ontológiák automatikus építésére, az ontológiák főleg kézzel készülnek (Cyc, SUMO). Egy ontológia elkészítése költséges folyamat, mint egy komoly lexikoné is. Utóbbi probléma megoldására született a Wikipédia, ami a nagyméretű közösség erejét használja fel a világ legnagyobb lexikonjának elkészítéséhez és karbantartásához. A Wikipédia az emberi tudás egy hatalmas és folyamatosan bővülő tárháza, kézenfekvő hát az ötlet, hogy a Wikipédia alapján építsünk formalizált tudásbázist, ontológiát. A nehézség, hogy a Wikipédiában található információ nagy része informális, természetes nyelvű szöveg. Ennek ellenére több ígéretes kísérlet történt az itt fellelhető tudás hasznosítására. Suchanek [6] a Wikipédia cikkeire illesztett felszíni minták alapján próbált bővíteni egy már meglévő ontológiát új tényekkel. A hamis találatok kiszűrése érdekében a talált tényeket összevetette az ontológiában már megtalálhatóakkal, és így szűrte ki az inkonzisztens találatokat. Mások [10] a Wikipédián található címszavak kategorizálásával foglalkoztak a hozzájuk tartozó cikkek tartalma alapján. Cucerzan [9] pedig a Wikipédia korpuszát használta fel névelemek egyértelműsítésére, összevetve a szövegben talált névelemek kontextusát az elemhez tartozó Wikipédia-cikk tartalmával. A DBPedia projekt a Wikipédiában található címszavakat rendezi egy ontológiába kategorizálva azokat, és az egyes cikkekből a címszavakhoz tartozó fontosabb tulajdonságokat is automatikusan meghatározzák. A tudásbázist a Wikipédián kívül más tudásbázisokban található adatokkal is bővítik a LinkedData szabvány segítségével. A DBPedia projekt ma is aktív és folyamatosan fejlődik. A nagy fogalmi lefedettség mellett a projekt köré szerveződő közösség aktivitása volt az indok, hogy a DBPedia adatbázisát választottuk az általános témájú ontológiánk alapjául.

Az ontológia építése mellett készítettünk egy többlépcsős szövegelemző rendszert, amely képes magyar nyelvű szövegekben az ontológiában tárolt fogalmak beazonosítására és így a szöveg szemantikus címkézésére. A szövegek automatikus címkézésének problémájával már többen foglalkoztak. A legtöbb megközelítés, mint a GATE [4], a PANKOW [3] vagy az ONTEA [7] mintafelismerésen alapszik. Ezen algoritmusok nehezen alkalmazhatóak kiterjedt ontológiákra, mivel a szabályrendszer létrehozása emberi erőforrást igényel. Ráadásul általában csak egyszerűbb összefüggéseket lehet jól leírni szabályok segítségével. A bonyolultabb vagy kevésbé gyakori megfogalmazások gyakran kimaradnak a szabályok közül, ezért a szabályalapú megközelítések találati aránya legtöbbször alacsony. Ezektől eltérően statisztikai módszereket és felügyelt tanulást kombináló szövegannotáló rendszer a SemTag [5]. A SemTag az ontológia elemeit a szövegkörnyezet alapján próbálja egyértelműsíteni nagy szövegkorpuszon készített statisztikák segítségével. A statisztikák készítésénél kihasználják a fogalmi hierarchiában rejlő információkat. Az algoritmus tovább pontosítható, ha kézzel címkézett tanítópéldákat is megadunk. A projekt során igen nagy mennyiségű szöveget annotáltak, viszont a felhasznált ontológia mérete kicsi volt, ezért a szöveg lefedettsége elég alacsony maradt, weboldalanként átlagosan alig több mint másfél entitást címkéztek fel.

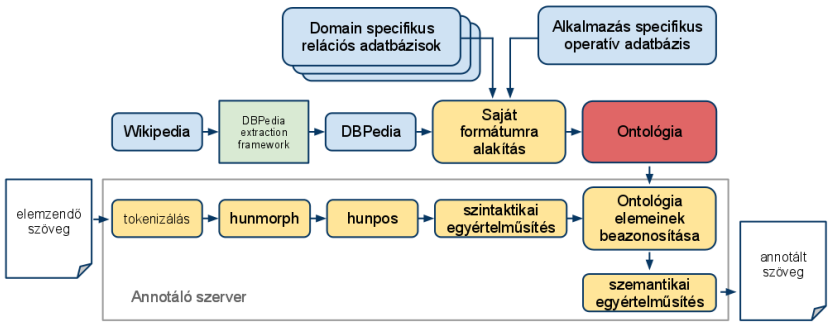
Mivel célunk volt, hogy nagyméretű ontológia alapján annotáljuk a szövegeket, a szabályalapú módszereket elvetettük, mert nagyon sok munka lenne az ontológiában előforduló összes osztályhoz szabályokat felvenni. Továbbá problémás lenne a szabályok karbantartása is, ami elengedhetetlen, ha egy olyan élő és folyamatosan változó tudásbázist alkalmazunk, ami a Wikipédián alapul. Így az egyértelműsítésnél olyan módszereket kerestünk, amik az ontológiában rejlő strukturális információk kihasználásával automatikusan oldják meg a feladatot. A projekt jelenlegi fázisában nem volt célunk a teljes egyértelműsítés, megelégedtünk azzal, hogy a lehetséges opciók számát ember számára gyorsan felfogható méretűre csökkentsük, kiválogatva néhány releváns opciót.

3. Az elemzőrendszer

Az annotálandó szövegek egy többlépcsős feldolgozási folyamaton mennek keresztül, aminek része a tokenizálás, morfológiai elemzés, POS-elemzés, névelem-felismerés és az utolsó fázisban az ontológia fogalmainak beazonosítása. A szintaktikai elemzésre szabadon felhasználható szoftvereket alkalmaztunk, mint a hunmorph [8], hunpos, valamint az OpenNLP tool maximum entrópián alapuló eszközeire [1] épülő saját elemző szoftvereket. Az elemzési lánc egyes komponensei közötti kommunikációra egy belső XML-reprezentációt alkalmazunk, ami tartalmazza az eredeti szöveget is, és az egyes modulok ezt egészítik ki egymásra épülő plusz információkkal. Az XML formátuma úgy lett kialakítva, hogy a bemenetre egyszerű szöveg vagy HTML is érkezhet és az elemzőrendszer megőrzi a kapott szöveg formázását. A külső modulokhoz készítettünk egy-egy csomagoló komponenst, ami elvégzi a transzformációt a belső XML formátum és a külső modul saját formátuma között. Az elemzőrendszer futtatható batch módban, ha egy meglévő szövegbázist kell feldolgozni és szolgáltatásként is, ahogyan használható HTML-oldalak valós idejű annotálására.

A rendszer felépítését és a felhasznált modulok sorrendjét az 1. ábra szemlélteti. A beérkezett szöveget először mondatokra, illetve szavakra bontjuk. Ezt a lépést az OpenNLP keretrendszer segítségével végeztük. Az OpenNLP alapvetően angol nyelvre készült, de csekély módosításokkal és a maximum entrópia nyelvi modellek újratanításával magyar nyelvre is alkalmazhatónak bizonyult. A szavakat ezek után szintaktikailag elemeztettük a hunmorph morfológiai elemzővel. A szövegen a hunpos POS-taggert is lefuttattuk. Mivel a hunmorph több lehetséges elemzést is előállít egy szóhoz, ki kellett választanunk az aktuális szövegkörnyezetnek legmegfelelőbb verziót. Ezt szintén egy maximum entrópia modellel oldottuk meg, ahol a kontextuális jellemzők előállításához formai jegyeket és a POS-tagger kimenetét használtuk. A jellemzők között szerepelt az aktuális szó mondaton belüli pozíciója, az aktuális, előző és következő szó végződése, illetve POS-tagje, a szó hossza és hogy kis- vagy nagybetűvel kezdődik-e. A hunmorph alternatívák leírására az elemzésben szereplő toldalékok nyelvtani jelöléseit és azok darabszámát használtuk, valamint a szótő és a teljes szó hosszának különbségét. Így elfogadható pontossággal sikerült a hunmorph által adott alternatívák közül a kontextusnak megfelelőt kiválasztani.

A szöveg annotálásakor a formai jegyek alapján jól felismerhető névelemek azonosítására, mint e-mail cím vagy telefonszám, használunk mintaillesztéses módszert is, de többnyire az ontológia alapján próbáljuk beazonosítani az egyes entitásokat. A felhasznált ontológia saját forrásokból, külső adatbázisokból is táplálkozik, és jelen esetben az ontológiába nemcsak a fogalmi osztálymodellt értjük bele és annak relációit, hanem a konkrét példányokat is. A példányok tárolása azért fontos, hogy az elemzőrendszer alkalmazásakor a világra vonatkozó tudás segítségével pontosabban meghatározhassuk a szöveg értelmét, kiegészíthessük automatikusan további információkkal. Így következtetéseket vonhatunk le, és végül a megfogalmazott kérdésekre sokkal pontosabb választ adhatunk. Ha a tudásbázisban megtalálható egy fogalom, pl. Budapest, amiről tudjuk, hogy egy város, és Magyarország fővárosa, akkor rögtön jóval több információ áll rendelkezésünkre, mintha csak a formai jegyek és a szövegkörnyezet alapján határoztuk volna meg, hogy egy városról van szó. Alapvető elképzelésünk az volt, hogy létrehozunk egy széles tárgyterületet lefedő általános ontológiát, és ezt egészítjük ki konkrét feladatokhoz kötődő speciális ontológiákkal. A különböző fogalmi rendszerek és adatbázisok összefogására létrehoztunk egy saját adatbázis-architektúrát, ami az OWL-szabvánnyal kompatibilis módon képes a fogalmak és példányok tárolására. Annak érdekében, hogy a tudásbázis alkalmas legyen valós idejű nyelvi elemzésre, ahol gyors válaszidőkre van szükség a nagyszámú lekérdezés miatt, az adatbázisréteg elé egy erre a célra kifejlesztett cache réteget is elhelyeztünk.



1. ábra. Az elemzőrendszer architektúrája.

Mivel igen sok entitást tartalmazó ontológiát használunk, fokozott problémát jelentett az annotálás során az egyértelműsítés, ugyanis egy hétköznapibb szóra akár több ezer találatot is adhat a keresés, amik közül általában egy helyes megoldás van az adott szövegkörnyezetben. Ugyan a projekt jelenlegi fázisában nem volt cél a teljes egyértelműsítés, de az mindenképpen szükséges, hogy ne szülessenek százával értelmetlen, a felhasználót felesleges adatokkal ellepő találatok. Ezt elkerülendő meghatároztuk a maximálisan megengedett többértelműségi szintet,

vagyis egy határértéket, aminél nem adhatunk vissza több találatot egy szóra. Hogy a találatokat szűrni tudjuk, relevanciaértékeket kellett rendelni hozzájuk. Ezt úgy állítjuk elő, hogy a talált entitásokra megszámoljuk a közvetlen és közvetett hivatkozásokat, és a legtöbbet hivatkozott elemek közül építjük a találati listát. Az ontológia fogalmainak keresésénél figyelembe vesszük az osztályhierarchiát és a fogalmak közötti összerendeléseket is. Tehát egy intézményre hivatkozásnak tekintjük azt is, ha a címe vagy az igazgatója neve szerepel a szövegben. Az egyértelműsítési lépéseket részletesebben a következő lista mutatja meg:

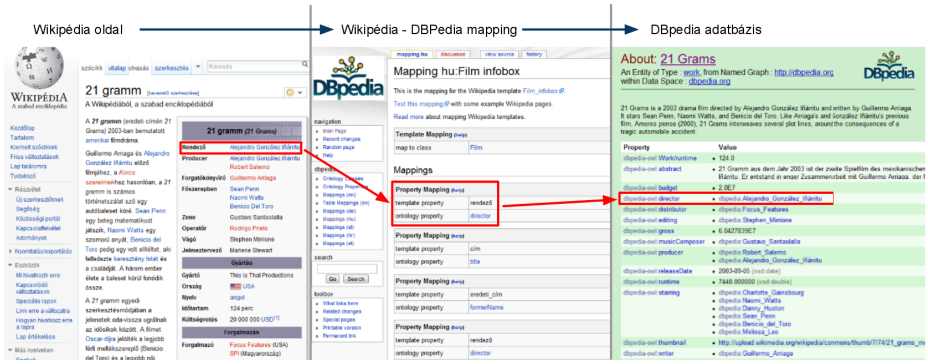
1. Az ontológiában szereplő összes entitás összes tulajdonságára megnézzük, hogy szerepel-e a szövegben. És azokat az objektumokat, amiknek valamelyik tulajdonságát megtaláltuk, felvesszük a találati listára. Ráadásul az illeszkedést nemcsak a szavak eredeti alakjára vizsgáljuk, hanem a szótövekre és többszavas kifejezésekre is.
2. Bizonyos mélységig (a kísérletekben 3-ra állítottuk ezt a limitet) felvesszük azokat az objektumokat is a találati listára, amikhez legalább az egyik talált objektum valamilyen tulajdonság mentén kapcsolódik. Pl. ha szerepel a szövegben egy utca neve, akkor az a város is, ahol az utca található, bekerül a találati listába.
3. Sorrendezzük a találati listát a referenciák alapján. Megszámoljuk, hogy melyik entitáshoz hány generáló szó volt. Egy közvetett úton felvett entitás pontszáma az őt generáló entitások pontszámának összege lesz. Ezek után meghatározzuk azt a pontszámot, amihez már kevesebb, minimum ezt a pontszámot elérő entitás tartozik, mint a maximálisan megengedett többértelműségi szint. Azokat az entitásokat, amik nem érik el a kívánt ponthatárt, eltávolítjuk a találati listából. Ha nincs ilyen pontszám, akkor elhagyjuk azokat az entitásokat, amik nem közvetlenül kapcsolódnak szavakhoz, hanem egy másik entitás generálta őket, és nincs több pontjuk, mint az őket generáló entitások közül a legtöbb ponttal rendelkező.

A fenti módszert még kiegészítettük egy kézi szűréssel, amit a doménspecifikus ontológiáknál alkalmaztunk, hogy a felszíni formákban nem megjelenő, belső, adminisztratív jellemzők, mint pl. adatbázis-azonosító, ne adjanak hamis találatokat. Itt mindössze arról van szó, hogy egyes tulajdonságoknál meg lehet adni, hogy ne keressünk rá illeszkedő szavakat a szövegben.

4. Az ontológia

Már említésre került, hogy az ontológia általános részét a Wikipédia, illetve a DBPedia alapján építettük. A DBPedia a Wikipédia legtöbb oldalán megtalálható "infobox"-ok félig formalizált tartalmát használja fel az ontológia építéséhez. Így összesen 2,6 millió egyedi entitást gyűjtöttek össze a Wikipédiáról. Ehhez társulnak a különböző doménspecifikus ontológiák, amelyek a LinkedData szabvány alapján csatlakoznak a DBPedia ontológiájához. Így az egész hálózat már 4,7 milliárd információdarabkát tartalmaz [2]. A struktúra magját képező

fogalmi hierarchia angol nyelvű, de összesen hat nyelven van összerendelés az ontológia és a Wikipédia-cikkek között, melyek egyike a magyar.



2. ábra. Wikipédia - DBPedia összerendelés

A DBPedia létrehozásához az adta az ötletet, hogy a legtöbb Wikipédia-oldalon a cikkek tartalmaznak úgynevezett "infobox"-okat. Ezek általában az oldal jobb felső oldalán megjelenő dobozok, amik táblázatszerűen foglalják össze az adott cikkben található fontosabb információkat. Az infoboxok kinézetét sablonok határozzák meg, amik témáról témára különbözőek. Így a Wikipédia-cikkeket az infobox sablonok alapján akkor is tudjuk kategorizálni, ha a Wikipédia kategóriarendszere nem megfelelő minőségű (ami sajnos igaz, különösen a magyar nyelvű Wikipédiára). Az infoboxok nem csak a kategorizálásban segítenek, hanem a sablonok tartalmazzák az adott kategóriára leginkább jellemző tulajdonságokat. A DBPedia projekt során ezeket a már-már formalizált adatokat gyűjtik össze egy RDF-en alapuló ontológiába. Az átalakítás során egyrészt automatikusan is kigyűjtik az egyes cikkekhez tartozó tulajdonságokat az infoboxokból, de van egy kézzel szerkesztett összerendelés is, ami pontosan megmondja, hogy az infoboxokban tárolt tulajdonságok minek felelnek meg az ontológiában. Erre azért volt szükség, mert az automatikus kinyerés zajos adatokat produkál. Az infoboxok nem eléggé egységes formátumúak, a Wikipédián az egyes infobox sablonok külön-külön jöttek létre, mindegyiket más-más ember szerkeszti, és így nincs igazán egységes elnevezési konvenció sem. Például semmi nem garantálja, hogy egy labdarúgónál ugyanolyan címkével jelölik a születési dátumát, mint egy festőnél. Az összerendelés karbantartására és bővítésére létrehoztak egy regisztráció után bárki által szerkeszthető szintén wiki alapon működő oldalt, ahol az infoboxok és az ontológia fogalmai közötti összerendelést meghatározó összes szabály megtalálható (http://mappings.dbpedia.org/). A projekt során mi is a kézi összerendelést bővítettük és az ez alapján létrejött ontológiát használtuk fel. A kézi összerendelésnek megvan az az előnye is, hogy a különböző nyelveken megtalálható tulajdonságok a közös fogalmi struktúrának köszönhetően egy az egyben megfeleltethetők egymásnak.

A magyar nyelvű infobox sablonok és az ontológia fogalmai közötti összerendelés elkészítésén kívül (2) kisebb módosítások az elemző program kódjában is szükségesek voltak, hogy a magyar Wikipédia-cikkek is helyesen kerüljenek bele a DBPedia ontológiájába. A magyar verzió eddig a leggyakrabban használt infobox sablonokhoz tartalmaz összerendelést. Az összerendelés a wiki elven működő oldalnak köszönhetően folyamatosan bővíthető és finomítható. Az annotáló rendszerben ezt az ontológiát kiegészítettük saját doménspecifikus adatokkal is, amelyek tartalmazzák a magyar település-adatbázist a Központi Statisztikai Hivatal településjegyzékét a települések statisztikai adataival, Magyarországon egységes kódrendszerével bővítve, valamint a Magyar Posta irányítószám-jegyzékét és a közoktatási, felsőoktatási intézmények publikusan elérhető címjegyzékét. A publikusan elérhető, hivatalos adatbázisok importálása folyamatosan történik azok kódrendszerével együtt. A kódrendszernek köszönhetően, más gépi feloldozású adatbázisokhoz is kapcsolható a rendszer.

5. Értékelés

Az eredményeket szubjektíven értékeltük több példaelemzés kézi átvizsgálásával. Jelenleg az elemzéshez felhasznált ontológia a DBPediából importált entitásokból, a magyar települések és közoktatási intézmények adatbázisából áll. Így 300 osztályt, közel 250 000 entitást és hozzájuk kapcsolódóan 510 000 tulajdonságértéket tartalmaz. A tesztek során ügyfélszolgálati leveleket és rövidebb, általánosabb témájú híreket elemeztünk. Általában a szövegekre sok illeszkedő példát talál a rendszer, és inkább a találatok szűrése okoz problémát. Sokszor magas a hamis találatok száma, de úgy gondoljuk, hogy az egyértelműsítő algoritmus fejlesztésével ezen sokat tudunk javítani. A rendszer teljesítményét a következő egyszerű példamondattal szemléltetnénk: "A Szilágyi Erzsébet Gimnázium tanulói gyakran hallgatnak Rockzenét, például GunsNRosest." A találatokat pedig a következő táblázat mutatja:

Illeszkedő szó	Tulajdonságok	Talált példányok
A	Settlement.VehicleCode	Augsburg
A	Athlete.Nationality.VehicleCode	Gregor Baumgartner, Julia Schruff
A	Person.BirthPlace.VehicleCode	Alexander Grimm
A	Country.VehicleCode	Austria
Szilágyi	last_name	SZILÁGYI, SZIL
Szilágyi	first_name	SZILÁGYI
Szilágyi	Settlement.name	Szil
Szilágyi	address.settlement.name	Dózsa György u. 1. Szil
Erzsébet	Settlement.name	ERZSÉBET
Erzsébet	first_name	ERZSÉBET

Illeszkedő szó	Tulajdonságok	Talált példányok
Erzsébet	address.settlement.name	Erzsébet Általános Iskola
Gimnázium	School.CampusType	Móra Ferenc Gimnázium Szerb Antal Gimnázium Eötvös József Gimnázium
hallgatnak	last_name	HALLGAT
Rockzenét	MusicGenre.Foaf:name	Rockmusic
GunsNRosest	Band.Foaf:name	Guns N Roses
Szilágyi Erzsébet	address.postal_name	1016 Mészáros u. 5-7. Budapest Szilágyi Erzsébet Gimnázium
Szilágyi Erzsébet	edu_institute.name edu_institute.short_name	Budapest Szilágyi Erzsébet Gimnázium

6. Összegzés

A munka során több különböző eszközt olvasztottunk egy egységes rendszerbe, kiegészítve saját fejlesztésű komponensekkel. Emellett egy magyar nyelvű ontológia építésével is foglalkoztunk, és ennek keretében elkészítettük a DBPedia magyar változatát. Az eddigiekben vázolt rendszer mélységében és funkcionálisan is fejlesztés alatt áll. A meglévő részek pontosságának fejlesztésénél a legfőbb pont az egyértelműsítés javítása a szintaktikai elemzés és főleg az ontológia elemeinek felismerése terén. Az egyértelműsítésnek kiterjedt irodalma van, és sok lehetséges megoldást vázoltak már különböző kutatók. Ezek közül olyan módszereket szándékozunk alkalmazni, melyek nem igényelnek emberi beavatkozást, és így kezelni tudjuk a nagyméretű tudásbázist. Tervezzük a meglévő ontológia finomítását és a rendszer kiegészítését úgy, hogy ne csak statikus entitásokat, hanem folyamatokat és azok állapotát is képes legyen kezelni. Fontos az is, hogy relációs operatív adatbázisokkal is összekapcsolható legyen a rendszer. Ebben az esetben már nem egy konstans, hanem egy állandóan változó tudásbázissal van dolgunk, aminek hatékony kezelése további kihívásokat rejt magában.

Hivatkozások

1. Adwait Ratnaparkhi: Maximum entropy models for natural language ambiguity resolution PhD thesis, University of Pennsylvania, 1998
2. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: DBpedia – A Crystallization Point for the Web of Data. in Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154–165, 2009.
3. Cimiano P., Ladwig G., Staab S.: Gimme' the Context: Context-Driven Automatic Semantic Annotation With C-Pankow. in proc. of the 14th International Conference onWorldWideWeb, New York, NY, USA. ACM Press, 2005, ISBN 1-59593-046-9, pp. 332–341.

4. Cunningham H., Maynard D., Bontcheva K., Tablan V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. in proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02), Philadelphia, 2002.
5. Dill S., Eiron N. et al.: A Case for Automated Large-Scale Semantic Annotation. Journal of Web Semantics, 2003.
6. Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum: SOFIE: A Self-Organizing Framework for Information Extraction 18th International World Wide Web conference (WWW 2009), Madrid
7. Laclavík Michal, Seleng Martin, Ciglan Marek, Hluchy Ladislav: ONTEA: Platform for pattern based automated semantic annotation Computing and Informatics, Vol. 28, 2009, 555–579, V 2009-Sep-16
8. Trón Viktor, László Németh, Péter Halácsy, András Kornai, György Gyepesi, and Dániel Varga: Hunmorph: open source word analysis in proc. of ACL., 2005
9. Silviu Cucerzan: Large-Scale Named Entity Disambiguation Based on Wikipedia Data in proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716, Prague, June 2007.
10. Zeno Gantner,Lars Schmidt-Thieme: Automatic Content-based Categorization of Wikipedia Articles in proc. of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 32–37, Suntec, Singapore, 7 August 2009.

Kontextualizált névelem-felismerés és relációkinyerés kórházi zárójelentésekben

Solt Illés[1,2], Szidarovszky P. Ferenc[1], Tikk Domonkos[1,2]

[1] Budapesti Műszaki és Gazdaságtud. Egyetem, Távközlési és Médiainf. Tanszék
H-1117 Budapest, Magyar Tud. krt. 2, e-mail: {solt,szidarovszky,tikk}@tmit.bme.hu
[2] Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics
D-10099 Berlin, Unter den Linden 6, e-mail: {solt,tikk}@informatik.hu-berlin.de

Kivonat Cikkünkben a kórházi zárójelentések szövegbányászati feldolgozásával foglalkozó i2b2 szervezet 2010-es, információkinyeréssel kapcsolatos feladatára (Fourth i2b2/VA Shared-Task) készített megoldásunkat ismertetjük. Az első, névelem-felismerési feladatban három entitástípus szövegbeli előfordulásait, pontosabban egy szűk bennfoglaló nyelvtani egységet kellett megjelölni. A második, állításosztályozási feladatban ezen entitások említésének jellegét (kijelentő, tagadó, spekulatív stb.) kellett osztályozni. Végül a harmadik, relációkinyerési feladatban az egy mondatban szereplő entitások között fennálló kapcsolat meglétét és pozitív esetben a típusát kellett megállapítani. Megoldásainkban kontextusra épülő, a rendelkezésünkre bocsátott tanítóadaton betanított – részben szabályalapú, részben felügyelt gépi tanuláson alapuló – módszereket alkalmaztunk. Munkánkban elemezzük az egyes eljárások hatékonyságát és megvizsgálunk néhány lehetséges továbbfejlesztési irányt.

1. Bevezetés

Az orvosbiológia a szövegbányászat egyik vezető alkalmazási területe, mivel számos feladattípus esetén sikerült hatékony eljárásokat fejleszteni, amelyek képesek például kutatóbiológusokat, klinikai és kutatóorvosokat, betegbiztosítási szakértőket a mindennapi munkájukban támogatni (l. pl. a [3] áttekintő tanulmányt). A már széles körben elterjedt információ-visszakeresési megoldások mellett manapság már egyre jellemzőbb az alkalmazási területtől jobban függő, és gyakran lényegesen bonyolultabb információkinyerő módszerek gyakorlati alkalmazása, vagy legalábbis ezek kísérleti bevezetése [7].

Három alapvető információkinyerési feladattípus a névelem-felismerés (named entity recognition, NER) [11], állításosztályozás (assertion classification) [8,21] és a relációkinyerés (relation extraction, RE) [5,8,9]. Névelem-felismerésnél az adott feladat szempontjából releváns névelem-, vagy más szóval entitástípusok előfordulásait kell egy szövegben azonosítani; orvosbiológiai alkalmazásokban ezek többnyire fehérjék, gének, mutációk, betegségek, gyógyszerek, szimptómák, tünetek stb. nevei. Állításosztályozásnál a feladat az entitás- és/vagy reláció-előfordulások szemantikai értékének meghatározása, pl. az állítás, tagadás és

spekuláció megkülönböztetése. Relációkinyerésnél az első lépésben azonosított vagy esetleg már eleve adott, entitás-előfordulások közötti kapcsolatok meglétét és típusát kell meghatározni; jellemző feladatok közé tartoznak a fehérje–fehérje kölcsönhatások (protein-protein interaction, PPI), betegség–kezelés, ill. gén–betegség összefüggések kinyerése.

A kórházi zárójelentések számos értékes információt tartalmaznak, amelyek segítségére lehetnek az orvoskutatóknak gyógykezeléseknek a páciensekre gyakorolt hatásának tanulmányozásában, a betegségek és gyógyszerezésük hatásvizsgálatában, az elvégzett vizsgálatok és betegségek felderítési arányának feltérképezésében stb. Ahhoz azonban, hogy a szöveges zárójelentésekből — melyek gyakran nem felelnek meg a nyelvtan és a helyesírás szabályainak, valamint számos szaknyelvi rövidítést is tartalmaznak — további kutatási célra felhasználható, statisztikailag szignifikáns mennyiségű céladathoz jussunk, ki kell nyerni a szövegből a releváns információkat és azok kapcsolatát. Az i2b2 (Informatics for Integrating Biology & the Bedside)[3] szakcsoport 2010-ben immár negyedik alkalommal[4] rendezett nemzetközi megmérettetést; minden évben más-más klinikai szövegekre vonatkozó aktuális szövegbányászati problémát helyezve a verseny fókuszába. A versenyen hagyományosan jól szerepelnek a magyar csapatok: 2006-ban 2. helyezést értek el Szarvas György és kollégái [18], 2008-ban pedig a mi csapatunk végzett az első helyen [17]. Jelen munkánkban kutatócsoportunknak a 2010-es kiírás feladataira adott megoldásait és az utólagos elemzések tanulságait ismertetjük. Fontosnak tartjuk megjegyezni, hogy bár a versenyen angol nyelvű szövegeken kellett dolgozni, a megoldásaink nagy része átvihető magyar nyelvű szövegfeldolgozásra is a megfelelő nyelvi eszközök magyar verziójának behelyettesítésével, így az általunk javasolt megoldások a hazai szakemberek részéről is érdeklődésre tarthatnak számot.

2. Feladatok

2.1. Névelem-felismerés

A 2010. évi verseny három egymásra épülő feladatból állt [1]. Az első feladat alapvetően névelem-felismerés volt, ahol három entitástípus felismerése volt a cél:

1. általános egészségügyi probléma (*medical problem*): ide tartoznak betegségek, tünetek, szimptómák, sérülések, abnormalitások stb.;
2. gyógykezelés (*treatment*): ide tartoznak a gyógyszerek, biológiai anyagok, gyógyszeradagolók, gyógyászati segédeszközök stb.;
3. vizsgálat (*test*): vizsgálati kezelések, testnedveken végzett laborvizsgálatok, életjelfunkciók mérési eredményei.

A feladatban a szokásos, kizárólag az entitásra koncentráló névelem-annotációs feladatokon túlmutatóan az entitás előfordulásának értelmezését támogatandó

[3] `https://www.i2b2.org/`

[4] `https://www.i2b2.org/NLP/Relations/`

az entitást fejként tartalmazó főnévi vagy melléknévi csoportokat kellett annotálni. Emellett az ún. prepozíciós szabály szerint bővíteni kellett az annotációkat a főnevet követő első prepozíciós szerkezettel, ha az nem tartalmazott amúgy is annotálandó entitást. Így a „pain in chest" szerkezetet egyben kellett felismerni, a „removal of mass"-t viszont két entitásként (l. még [1]). Az annotálás módja az orvosbiológiai korpuszokon gyakran alkalmazott tokenszintű annotáció. Szemben a karakterszintű annotációval, ez a jelölés emberi annotációnál kevesebb időráfordítással állítható elő, ellenben kevésbé pontos jelölést tesz lehetővé, mint amilyet a klinikai szövegek szintaktikai gyakorlata indokolna (pl. egybeírások, torlódó szavak, szokatlan rövidítések).

2.2. Állításosztályozás

A második feladatban az egészségügyi probléma entitástípus előfordulásait kellett a következő 6 szemantikai osztály valamelyikébe sorolni: megfigyelhető (állítás), nem figyelhető meg (tagadás), lehetséges, feltételezett, feltételhez kötött, vagy mással (nem a páciensszel) kapcsolatos.

2.3. Relációkinyerés

A harmadik feladatban azt kellett meghatározni, hogy milyen kapcsolat van – ha van egyáltalán – az egy mondatban szereplő entitások között. Itt 8 reláció szerint kellett entitáspárokat vizsgálni. A relációk egyet kivéve (*problem indicates problem*, PIP) szimmetrikusak voltak, az irányított relációnál természetesen a megfelelő irányt is meg kellett határozni. A reláció egyértelműen meghatározta, hogy milyen entitástípusok között állhat fent.

2.4. Kiértékelés, lebonyolítás

Minden feladatra három megoldást lehetett csapatonként beküldeni. Az első feladat esetében elsődlegesen a pontos illeszkedés mikro F-mérték, pontosság, felidézés hármast alapján rangsorolták a versenyzőket. Ugyanerre a három mértékre átfedő illeszkedés szerint is kiértékelték a megoldásokat, vagyis ekkor már helyesnek számított az a predikció, ahol tokenszinten van átfedés a helyes és a predikált entitások között. A két osztályozási feladatnál az összes osztályra vetített mikro F-mérték, pontosság, felidézés szerint értékelték a beküldéseket.

A verseny lebonyolítása az alábbi módon igazodott a feladatok sorrendjéhez. Egy-egy feladat leadási határideje között 24 óra állt a versenyzők rendelkezésére. A névelem-felismerési feladat teljesítése után közzétették a résztvevők számára a tesztadatokhoz tartozó helyes értékeket (ground truth), amit a második és harmadik feladat megoldásánál így fel lehetett használni; hasonlóan nyilvánosságra hozták az állításosztályozás megoldókulcsát is a feladat teljesítése után. Ez az egymásra épülő kiértékelési módszer lehetővé tette az egyes részfeladatokra beadott megoldások egyenkénti értékelését, hiszen így nem adódtak össze a hibák. Gyakorlati alkalmazásokban ugyanakkor az állításosztályozásnál és a relációkinyerésnél természetesen magasabb hibaértékkel szükséges kalkulálni, amikor a

három részfeladatot egymásra épülve hajtják végre. A hibanövekedés nagyságrendjét például a [7] cikk alapján becsülhetjük meg, ahol a névelem-felismerés és a PPI-kinyerés kombinált folyamatának hatékonyságát vizsgálták.

3. Módszerek

3.1. Lexikon és szintaktikai minták alapján működő névelem-felismerő

Az első feladat egy többcímkés szekvenciaannotálási feladatként fogható fel, ahol a címkék a három entitástípushoz tartoznak. A feladatkiírás logikáját követve elsőként a fejként szereplő entitásokat határoztuk meg (illesztés), majd kiterjesztettük az entitás szövegkörnyezetét az előírt feltételek szerint (kiterjesztés).

Az illesztési feladatban a fejentitások felismeréséhez a tanító halmazból indultunk ki. Egy entitást akkor tekintettünk jelöltnek az adott osztályhoz, ha az átfedő pontosság – az előfordulások, amelyek átfednek az adott osztály annotációival osztva az összes előfordulással – meghaladott egy bizonyos küszöbértéket. F-mértékre optimalizálva 0,6 bizonyult az optimális küszöbértéknek a tanító halmazon. A lexikont hasonló kiértékelési módszerrel bővítettük az i2b2 Obesity Challenge [20] megoldásánál általunk összeállított fogalomlexikon [17], valamint internetes forrásokból származó adatok alapján. A kiterjesztés alapjául tehát az így összeállított névelemlistának a tesztszövegre illeszkedő tokenjei szolgáltak.

A szövegkörnyezet kiterjesztéséhez több angol nyelvi elemző kimenetét elemeztük abból a szempontból, hogy ezek hogyan illeszkednek a meglehetősen körmönfont annotációs irányelvekhez. Összehasonlításunkban a Stanford szintaktikai elemzőt[5], a Charniak–Lease elemzőnek[6] a McClosky-féle biológiai szövegeken tanított modellel [12] való verzióját, az Enju elemzőt[7], valamint a GENIA chunkert[8] vizsgáltuk. Azt tapasztaltuk, hogy a jól formált mondatokon a Stanford szintaktikai kimenete egyezett meg leginkább a versenyen elvárttal, míg a nyelvtanilag helytelen (pl. felsorolás jellegű) mondatoknál a GENIA chunker bizonyult a legjobbnak.

Bizonyos esetekben azonban az elemző kimenete módosítást igényelt, hogy megfeleljen az annotációs irányelveknek. Az egyik tipikus példa erre, hogy az elemzők a főnévi csoportokat jellemzően nem vágják kötőszavak mentén, azaz a „The patient experienced X and Y" mondatban az X entitást tartalmazó főnévi csoportot az elemzők „X and Y"-ként azonosítják, míg a versenyben a legszűkebb bennfoglaló nyelvtani egységet X-szel kell annotálni. Ilyenkor tehát kettévágjuk a kötőszavak mentén a főnévi és melléknévi csoportokat. Szintén eltért az elemzők kimenete az annotációs irányelvektől a bizonyosságot jelentő határozószók esetén („likely" stb.). Bár az elemzők – nyelvtanilag helyesen – ezeket a főnévi

[5] `http://nlp.stanford.edu/software/lex-parser.shtml`
[6] `ftp://ftp.cs.brown.edu/pub/nlparser/reranking-parserAug06.tar.gz`
[7] `http://www-tsujii.is.s.u-tokyo.ac.jp/enju/`
[8] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/postagger/geniatagger-3.0.1.tar.gz`

csoportba vonták be, az állításosztályozásra való tekintettel (l. ott) ez ellenkezett az irányelvekkel.

Az annotációs irányelveket szintaktikai gráfokra vonatkozó illeszkedési szabályokra írtuk át, és ezekkel határoztuk meg az entitások szövegkörnyezetét. A szabályok az illeszkedő fejentitást kibővítették a tartalmazó főnévi, ill. melléknévi csoportra, valamint további bővítést végeztek a prepozíciós szabály szerint. A kiterjesztésnél a Stanford elemző és a GENIA chunker kimenetének unióját vettük, ugyanis ezek akkor is átlapolódhatnak, ha a fejentitásokra diszjunktak. Például az *egészségügyi probléma* entitástípus tartalmazza mind a „reflux" és „disease" szavakat, de a „reflux disease" kifejezést nem. Ekkor a „The patient has reflux disease." mondatban mindkét szót külön entitásként felismerjük, amelyeknek a kiterjesztése egybeesik. Ilyen esetekben — ha az entitások azonos osztályba tartoztak — az egyesített kiterjesztést vettük.

3.2. Állításosztályozás szabályalapú módszerrel

A második feladatot szabályalapú megközelítéssel oldottuk meg, amely az entitás szövegkörnyezete alapján határozta meg az állításosztályt. Az entitást tartalmazó mondatot az előfordulást megelőző, tartalmazó, követő, illetve körülvevő részre bontottuk. A tanító adat alapján azonosítottuk az egyes állításosztályokra utaló kulcsszavakat és hogy melyik szövegrészletre vonatkozik a hatásköre, l. például az 1. táblázatot.

1. táblázat. Példák az állításosztályokra utaló kulcsszavakra

Osztály	**Kulcsszó**		
	megelőző	**követő**	**körülvevő**
Megfigyelhető (állítás)	due to		
Nem figyelhető meg (tagadás)	without		non
Lehetséges	possible		
Feltételezett	if you		any
Feltételhez kötött		when	exertion
Mással kapcsolatos		wife	

Korábbi kontextusalapú állításosztályozási eljárásunkat [17] használtuk fel frázis-, illetve mondatszintű negáció, allergia és családdal kapcsolatos szövegrészletek, illetve hatáskörök bejelölésére, és az adott hatáskörbe eső entitásokat rendre a *nem figyelhető meg*, *feltételhez kötött* és *mással kapcsolatos* osztályba soroltuk. A különböző módszerek által adott eredmények közötti esetleges ellentmondás esetén a nagyobb *a priori* valószínűségű állításosztályt választottuk. Azokat az entitásokat, amelyeket a fenti módszerek egyike sem sorolt be valamelyik osztályba sem, a *megfigyelhető* osztályhoz rendeltük.

3.3. Relációkinyerés felügyelt tanulókkal

A verseny relációkinyeréssel kapcsolatos feladata többcímkés osztályzásnak tekinthető, ahol minden adott feltételnek megfelelő — azaz a megfelelő entitásosztályokba tartozó — entitáspárt valamely relációtípushoz kell hozzárendelni. A feladatot relációtípusonként bináris osztályozók alkalmazásával oldottuk meg, azaz összesen 9 modellt építettünk: egy-egy osztályozót a szimmetrikus relációtípusokra, és irányonként egy osztályozót a PIP-típusra. Végül feloldottuk az esetleges ellentmondásokat, vagyis amikor több osztályozó is a pozitív osztályba sorolt egy entitáspárt.

A relációtípus meghatározza, hogy milyen típusú entitások között állhatnak fenn: 5 relációtípust definiáltak egészségügyi problémák és kezelések között, kettőt egészségügyi problémák és vizsgálatok között, míg az aszimmetrikus PIP reláció – amelyet szétbontottunk PIP ($\curvearrowright$) és PIP ($\curvearrowleft$) irányfüggő altípusokra – két egészségügyi probléma közt állhat fent. Vegyük észre, hogy minden relációtípus legalább egy egészségügyi probléma entitást tartalmaz (bővebb statisztikákkal a 2. táblázat szolgál).

Két különböző megközelítést alkalmaztunk a feladat megoldására. Alapmódszerként egy együttes előfordulás (kollokáció) alapú módszert alkalmaztunk, amelyik az entitások között gyakorta előforduló szavakat és szó n-gramokat azonosítja relációtípusonként, és amely az így felismert minták alapján osztályozza a tesztpéldányokat.

A második módszer gépi tanulási problémaként közelíti meg a feladatot, és szupport vektor gépeket (SVM) alkalmaz különböző magfüggvényekkel (kernel) az osztályozó modellek megtanulásához, természetesen mind a 9 relációtípusra külön modellt építve. A gépi tanulókat 10-szeres keresztvalidálással értékeltük ki, amihez 10 nagyjából azonos méretű részre osztottuk a tanítóhalmazt. Összehasonlításként az n-gram alapú módszert is kiértékeltük ily módon.

Szó n-gram alapú relációkinyerés. A (szó) n-gram alapú módszer a tanító halmazon megfigyelt szósorozatok előfordulási statisztikái alapján működik. Minden osztályra külön modellt készítettünk.

Elsőként tokenizáltuk a mondatokat, és a tokeneket az alábbi 4 tokenosztályba soroltuk:

1. entitás (a tanítóhalmazon definiálva);
2. számok (csak számjegyeket tartalmazó tokenek);
3. egyéb szavak;
4. írásjelek.

Az entitások előfordulását az entitástípus címkéjével (entity blinding), míg a számtokeneket egységesen egyazon címkével helyettesítettük, majd készítettünk egy n-gram szótárat a tanítóhalmazon. Különböző beállításokkal futtattuk a kísérleteket az n-gramok minimális és maximális hosszát, valamint az összesített minimális előfordulás mennyiségét ($\min_{\text{freq}}$)változtatva.

Egy n-gramnak valamely relációtípus szerinti osztályozási pontosságát a (típus szerinti) pozitív és az összes előfordulás arányának alapján határozzuk meg.

2. táblázat. Relációtípusokra vonatkozó statisztikák és keresztvalidációval mért eredmények a tanító adatokon. A pozitív példák a tanító adatokban szereplő relációk, negatív példák a mondatokban található fennmaradó típushelyes entitáspárok. Mikro F-mérték eredmények, a legjobb eredmény típusonként félkövérrel szedve.

Reláció	**Statisztika**				**Módszer**					
	Poz	**Neg**	**P/N**	**P%**	**kBSPS**	**SpT**	**SL**	**PT**	**APG**	**n-gram**
TERP	1 711	1 858	0,92	48%	0,78	0,74	**0,83**	0,74	**0,83**	0,68
TECP	295	3 274	0,09	8%	0,49	0,44	**0,51**	0,41	0,45	0,35
TRAP	1 413	2 874	0,49	33%	0,64	0,64	0,71	0,64	**0,74**	0,52
TRCP	294	3 993	0,07	7%	**0,45**	0,34	0,42	0,35	0,43	0,29
TRIP	107	4 180	0,03	2%	**0,40**	0,23	0,38	0,24	0,28	0,37
TRNAP	106	4 181	0,03	2%	**0,44**	0,27	0,37	0,23	0,37	0,27
TRWP	56	4 231	0,01	1%	**0,19**	0,13	0,02	0,13	0,11	0,11
PIP ($\curvearrowright$)	900	7 688	0,12	10%	0,49	0,00	**0,55**	0,00	–	0,29
PIP ($\curvearrowleft$)	320	8 268	0,04	4%	0,17	0,00	**0,27**	0,00	–	0,12
Összes	**5 202**	**40 547**	**0,13**	**11%**	**0,60**	**0,47**	**0,65**	**0,47**	**0,52**	**0,54**

Egy entitást is tartalmazó, pozitív példamondatban előforduló n-gramot csak akkor tekintünk pozitív példának, ha az entitás része az annotált relációnak.

Osztályozásnál egy mondatot akkor tekintünk pozitívnak, ha a mondatbeli legmagasabb n-gram pontossági érték elér egy előre definiált küszöbértéket; itt természetesen csak az n-gram szótár elemeit tekintjük. A pontosság kiszámításához a fenti 10-szeres keresztvalidációt alkalmaztuk és minden részhalmaz 10%-án állítottuk be a pontossági küszöbértéket.

Az optimális paraméterértékek meghatározásánál az $n = 1, \ldots, 4$, ill. $\min_{\text{freq}} = 4$ értékek eredményezték a legjobb átlagos keresztvalidált F-mértéket. Mondatszintű osztályozásnál futtattunk kísérleteket a maximumon kívül más aggregáló függvénnyel is, de mindegyik hatékonysága elmaradt a maximumétól.

Kernelfüggvény alapú osztályozás SVM-mel. A szupport vektor gépek adott tanítóhalmaz esetén azt a lineáris hipersíkot határozzák meg, amely a legjobban szeparálja a pozitív és negatív tanító adatokat [6]. Ha a két halmaz lineárisan nem szeparálható, akkor kernelfüggvények segítségével a feladat egy nemlineáris, jellemzően magasabb dimenziójú térbe transzformálható, ahol már fennállhat a szeparálhatóság [16]. A kernelfüggvény egy adott párhoz egy hasonlósági értéket rendel, amely a pár közti belső szorzatként egyszerűen számolható, és lehetővé teszi sokdimenziós problématerek használatát, amelyek például a mondatok strukturális jegyeit jobban leíró, bonyolultabb nyelvtani reprezentációk esetén szükségesek lehetnek.

A kernelfüggvényekkel kapcsolatos kísérleteinkben felhasználtuk azt a kernelösszehasonlító keretrendszerünket, amelyet eredetileg fehérjeinterakciók (PPI)

kinyeréséhez fejlesztettünk [19]. Az abban rendelkezésre álló 13 kernelfüggvényből 5-tel folytattunk kísérleteket:

- a *shallow linguistic* (SL) [4] kizárólag felszíni nyelvtani jegyekkel operál (szófajok, tokenek, lemmák stb.);
- a *partial tree* (PT) [13] és a *spectrum tree* (SpT) [10] kernelfüggvények a mondat szintaktikai fáján dolgoznak;
- a *k-band shortest path spectrum* (kBSPS) [14,19], és az *all-paths graph* (APG) [2] kernelek pedig a mondat függőségigráf-reprezentációja alapján definiálnak hasonlósági függvényt.

A kernelek alkalmazása előtt át kellett alakítani a rendelkezésre bocsátott zárójelentés-dokumentumokat és ezek mondataihoz generált nyelvtani elemzéseket a PPI-relációkinyerésnél *de facto* standardként használt XML formátumra [15], hogy a kerneleket alkalmazni lehessen. A kernelek különböző mondatreprezentációt használnak, ezért az összes reprezentációs formátummal gazdagítani kellett a dokumentumokat. Az SL kernelhez a GENIA taggert alkalmaztuk a lemmák meghatározásához. A PT és SpT kernelekhez szükséges szintaktikai fákhoz a Charniak-Lease elemzőt alkalmaztuk a McClosky-féle biológiai modellel, míg a függőségi gráfokat a Stanford konverterrel állítottuk elő a szintaktikai fákból.

Különböző paraméterbeállításokkal futtattunk keresztvalidációs kísérleteket a kernelekkel, hogy megtaláljuk a legjobb beállításokat, amelyekkel azután az egész korpuszon betanítottuk a modellt, és ezt alkalmaztuk a tesztadatokon. A 2. táblázatban összehasonlítjuk a kernelekkel elért eredményeket és az alapmódszerként használt n-gram alapú módszert a keresztvalidációs adatokon. Minden kernelre csak a legjobb beállítással elért eredményt közöljük.

3. táblázat. A három feladatban elért eredmények a tesztadatokon

Feladat	Módszer	TP	FN	FP	R	P	F
Névelem-felismerés	Illesztés + kiterjesztés	24 892	20 117	16 962	0,55	0,59	0,57
Állításosztályozás	Kulcsszó- és kontextusalapú szabályok	15 805	2 745	2 745	0,85	0,85	0,85
Relációkinyerés	SVM SL kernellel [4]	6 301	2 769	3 122	0,69	0,67	0,68
	kBSPS kernellel [14]	447	8 623	3 772	0,05	0,11	0,07
	Szó n-gram alapon	6 040	3 030	23 697	0,67	0,20	0,31
	SL és kBSPS kernelek kombinációja	4 639	4 431	8 636	0,51	0,35	0,42

4. Eredmények

A 3. táblázat összefoglalja a három feladatban elért eredményeinket, amelyeket már a helyes referenciaadatokkal ellátott tesztadatok segítségével számoltunk ki (a szervezők által rendelkezésre bocsátott kiértékelő ugyanis pontatlan volt). A 4. táblázatban részletesen ismertetjük a névelem-felismerési feladat eredményeit, míg az állításosztályozás részeredményei az 5. táblázatban találhatók. A

névelem-felismerésben 0,57 F-mértéket értünk el pontos, és 0,80-at eredményt átfedő illeszkedés esetén. Állításosztályozásnál 0,92, míg relációkinyerésnél 0,68 volt ez az érték.

4. táblázat. Névelem-felismerési eredmények entitástípusonként a tesztadatokon

Illeszkedés	Entitás	TP	FN	FP	R	P	F
Pontos	Egészségügyi probléma	11,4k	7,2k	5,4k	0,61	0,68	0,64
	Gyógykezelés	7,3k	6,3k	5,6k	0,54	0,57	0,55
	Orvosi vizsgálat	6,3k	6,6k	6,1k	0,49	0,51	0,50
	Összes	**24,9k**	**20,1k**	**17,0k**	**0,55**	**0,59**	**0,57**
Átfedő	Egészségügyi probléma	14,2k	4,3k	2,1k	0,77	0,87	0,82
	Gyógykezelés	10,3k	3,3k	2,4k	0,76	0,81	0,79
	Orvosi vizsgálat	10,0k	2,9k	1,8k	0,77	0,84	0,81
	Összes	**36,3k**	**8,7k**	**6,3k**	**0,77**	**0,85**	**0,80**

5. táblázat. Állításosztályozási eredmények a tesztadatokon

Osztály	TP	FP	FN	R	P	F
Megfigyelhető (állítás)	11 754	1 300	1 271	0,90	0,90	0,90
Nem figyelhető meg (tagadás)	2 934	843	675	0,81	0,78	0,79
Lehetséges	596	392	287	0,67	0,60	0,64
Feltételezett	361	55	356	0,50	0,87	0,64
Feltételhez kötött	26	131	145	0,15	0,17	0,16
Mással kapcsolatos	134	24	11	0,92	0,85	0,88
Összes	**15 805**	**2 745**	**2 745**	**0,85**	**0,85**	**0,85**

5. Diszkusszió

A 4. táblázatban a névelem-felismerésnél látható a pontos és az átfedő illeszkedéssel számolt eredmények közötti jelentős különbség (0,23-os eltérés F-mértékben) arra utal, hogy az entitások hatáskörének helyes kiterjesztése gyakran nehéz feladatnak bizonyult. A pontosság rendre magasabb a felidézésnél a teszthalmazon, ami a F-mértékre vonatkozó lexikonoptimalizálásnak az eredménye.

Az állításosztályozási feladatnál az osztályozási hatékonyság a rendelkezésre álló tanító adatok számával korrelál, l. 5. táblázat. Kivételt csak a *mással kapcsolatos* osztály jelent, amely könnyen felismerhető volt egyes rovatfejlécbeli kulcsszavak gyakori előfordulása miatt (pl. „family history”).

A relációkinyerésnél használt módszerek elemzésénél (2. táblázat) — egybehangzóan a [19] munka megállapításaival — azt találtuk, hogy a szintaktikai elemzési fa alapú kernelfüggvények (PT, SpT) kevésbé képesek a relációtípusra jellemző jegyek kiemelésére, mint a felszíni jegyeken vagy függőségi gráfokon alapuló módszerek (SL, kBSPS és APG), ezért a további kísérletekben nem használtuk fel őket. Az APG kernelt implementációjának lassú sebessége miatt voltunk kénytelenek kizárni (10–50-szeres különbség, l. még 24 órás limit a versenyben). Következésképpen a versenyben az SL a kBSPS kernelek, továbbá a kombinációjuk és az alapmódszer által adott eredményeket vettük számításba. A kBSPS nagyon kevés pozitív osztálycímkét predikált, és így alacsony felidézést, valamint F-mértéket ért el. Az eredmények utólagos kiértékelésénél kiderült, hogy ez a kBSPS kernel paraméterbeállításra való érzékenységének köszönhető: a 9 relációtípusból csak 2-nél adott értékelhető eredményt a keresztvalidációval legjobbnak bizonyult beállítással. Ezzel ellentétben az SL, amely az alapbeállítással is kedvező teszteredményeket produkált, lényegesen robosztusabb, ami valószínűleg azzal is összefügg, hogy ez a megoldás a legkevésbé érzékeny az agrammatikus mondatokból adódó elemzési hibákra, hiszen kizárólag felszíni jegyekkel operál.

6. Összefoglalás

Munkánkban ismertettük a 2010 i2b2/VA Shared-Task nemzetközi verseny három lényegesen különböző részfeladatára elkészített megoldásainkat. Névelemfelismerésnél először lexikonalapú illesztést, majd utána szintaktikai elemzés alapú hatókör-kiterjesztést alkalmaztunk; állításosztályozásnál kulcsszavakon alapuló szabályalapú rendszert fejlesztettünk ki; relációkinyerésnél egy kollokációalapú alapmódszert és több kernelfüggvényt használó SVM-et alkalmaztunk megoldásainkban. Ugyan a versenyben elért helyezéseket a szervezők csak 2010 novemberében, a kapcsolódó konferencián hozzák nyilvánosságra, mindhárom feladattípusnál sikerült a szakterület jelenlegi sztenderdjét legalább elérő eredményeket hozni.

Az utólagos elemzések több olyan irányt mutatnak, amellyel az egyes feladatokra kapott eredmények pontossága esetleg javítható, amelyeket tehát a további munkáinkban vizsgálni kívánunk:

- Névelem-felismerésnél a lexikonalapú illesztés, majd kiterjesztés helyett célszerű lenne az elemzési hibákra érzéketlenebb conditional random fields (CRF) tanuló módszer alkalmazása.
- Relációkinyerésnél számos módon javítható a hatékonyság. A kernelfüggvények esetén a hasonlósági értékek feladatspecifikus módosításával elérhető lenne az agrammatikus mondatok hatékonyabb kezelése. A kernelek optimalizálásánál a keresztvalidációs halmazon való optimalizálás helyett alternatív módon a pozitív/negatív osztálycímkék tanító halmazon mért arányának teszthalmazon való közelítésével kiszűrhetők a nyilvánvalóan téves eredmények (l. kBSPS) — feltéve, hogy ez az arány hasonló a teszthalmazon is. Mivel az adott feladatban nagyon eltért az egyes osztályok egyedszáma, ezért

a több tanító adattal bíró osztályok felülsúlyozásával kedvezőbb mikro F-mérték lett volna elérhető. Végül pedig a kernelek kombinálásával, akár relációtípusonként, akár mondattípusonként is lehet a teljesítményt lényegesen javítani.

Köszönetnyilvánítás

Tikk Domonkost az Alexander von Humboldt Alapítvány, Solt Illést a DAAD támogatta.

Hivatkozások

1. 2010 i2b2/VA Challenge Documentation. `https://www.i2b2.org/NLP/Relations/Documentation.php`.
2. A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2, 2008.
3. A.M. Cohen and W.R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.
4. C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 401–408, Trento, Italy, 2006. The Association for Computer Linguistics.
5. T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18(4):644–52, 2008.
6. T. Joachims. *Making large-scale support vector machine learning practical, Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, 1999.
7. R. Kabiljo, A. Clegg, and A. Shepherd. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233, 2009.
8. J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Overview of BioNLP'09 shared task on event extraction. In *BioNLP'09: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 1–9. Association for Computational Linguistics, 2009.
9. M. Krallinger, R. A. Erhardt, and A. Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10(6):439–45, 2005.
10. Tetsuji Kuboyama, Kouichi Hirata, Hisashi Kashima, Kiyoko F. Aoki-Kinoshita, and Hiroshi Yasuda. A spectrum tree kernel. *Information and Media Technologies*, 2(1):292–299, 2007.
11. U. Leser and J. Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–69, 2005.
12. D. McClosky. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. PhD thesis, Department of Computer Science, Brown University, 2009.
13. A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proc. of The 17th European Conf. on Machine Learning*, pages 318–329, Berlin, Germany, 2006.

14. P. Palaga. Extracting relations from biomedical texts using syntactic information. Master's Thesis, Technische Universität Berlin, May 2009.
15. S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9 Suppl 3:S6, 2008.
16. B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in kernel methods: support vector learning*. The MIT Press, 1999.
17. I. Solt, D. Tikk, V. Gál, and Zs. T. Kardkovács. Context-aware rule based classifier for semantic classification of diseases in discharge summaries. *J. Am. Med. Inform. Assoc.*, 16(4):580–4, July/August 2009.
18. Gy. Szarvas, R. Farkas, and R. Busa-Fekete. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc.*, 14:574–80, Sep-Oct 2007.
19. D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, July 2010.
20. Ö. Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–70, 2009.
21. Ö. Uzuner, X. Zhang, and T. Sibanda. Two approaches to assertion classification. In *AMIA Annual Symposium Proceedings*, volume 2008, page 752. American Medical Informatics Association, 2008.

Kulcsszókinyerés magyar nyelvű tudományos publikációkból

Berend Gábor[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
6720 Szeged, Árpád tér 2.
berendg@inf.u-szeged.hu
[2] SZTE-MTA Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos körút 103.
rfarkas@inf.u-szeged.hu

Kivonat: A szöveges dokumentumokból történő kulcsszókinyerés számos alkalmazási területen hasznosítható, a katalogizáló és kivonatoló rendszerektől kezdve egészen az információ-visszakereső módszerekig. Különösen igaz mindez a tudásalapú társadalom sajátosságaiból adódóan a tudományos publikációkra: [1] szerint a 2006-ban megjelenő publikációk száma meghaladta az 1,3 milliót; [8] szerint pedig az elkészült tudományos munkák közel 50%-a olvasatlanul marad. A publikációk számának folyamatos és markáns növekedése miatt feldolgozásuk mára csupán automatikus eszközök segítségével képzelhető el. Jelen cikk az első magyar nyelvű tudományos publikációkra kidolgozott kulcsszókinyerő rendszert mutatja be.

1 Bevezetés

Kulcsszavak alatt az egyes dokumentumok tartalmát tömören és jól reprezentáló fogalmakat értjük, az automatikus kulcsszókinyerés alapfeladata pedig ezen fogalmak egy halmazának számítógépes meghatározása a dokumentum tartalma alapján. Habár a tudományos publikációk hatékony feldolgozhatóságához ezek szerzőik által történő megadása hasznos lenne, legtöbb esetben mégsem találkozhatunk a cikkekhez rendelt kulcsszavakkal. Jelen cikkben bemutatjuk az első automatikus kulcsszókinyerő rendszert, amelyet magyar nyelvű tudományos publikációkra dolgoztunk ki. Tanított modellünk és rendszerünk kipróbálható és elérhető a Creative Commons licenc alatt a www.inf.u-szeged.hu/rgai/kpe url-ről.

A szöveges dokumentumokból történő kulcsszókinyerés felhasználhatósága széles körökre terjed ki – a katalogizáló és kivonatoló rendszerektől kezdve egészen az információ-visszakereső alkalmazásokig. Különösen igaz mindez a tudás alapú társadalom sajátosságaiból adódóan a tudományos publikációkra: Björk és társai [1] szerint a csupán 2006-ban megjelenő publikációk száma meghaladta az 1,3 milliót; Taleb [8] szerint pedig az elkészült tudományos munkák közel 50%-a olvasatlanul marad. A publikációk számának folyamatos és markáns növekedésének fényében feldolgozásuk mára csupán automatikus eszközök segítségével képzelhető el.

Kulcsszókinyerő rendszerünk eredményességének teszteléséhez két eltérő témájú (politika- és neveléstudományi), szerzői kulcsszavakkal ellátott cikkeket tartalmazó újságra támaszkodva építettünk adatbázist.

Az így előálló adatbázis segítségével – a korábbi kulcsszókinyeréssel foglalkozó munkák többségéhez hasonlóan [4, 7, 9] – mi is felügyelt tanulási feladatot fogalmaztunk meg, ahol a cél a szerzői kulcsszavak szövegből történő kinyerése volt. Munkánk során az eddigi – a feladat angol nyelven történő megvalósítására kialakított – jellemzőkészletet bővítettük ki, illetve adaptáltuk a magyar nyelv sajátosságaihoz alkalmazkodva. Az egyes cikkekből kinyert kulcsszavak meghatározását ezen kibővített jellemzőkészlet segítségével, a belőlük lehetségesnek tartott frázisok kulcsszóként való viselkedésének poszteriori valószínűségének (különböző modellek szerint történő) kiszámítását alapul véve hajtottuk végre.

A keresztvalidáció során tapasztalt eredmények alapján kijelenthető, hogy az angol nyelvre szánt megközelítések – a megfelelő átalakítások után – magyar nyelvre is átültethetők, a jellemzőkészlet kiterjesztésével pedig további javulások érhetők el az eredményekben.

2 Kapcsolódó munkák

A kulcsszógenerálást végző tanuló algoritmusokat a bennük használt felügyelet mértékén túl működési alapelvük alapján különböztethetjük meg. A két nagy irányzat a kulcsszóajánlás és a kulcsszókinyerés.

A *kulcsszóajánló* rendszerek működési elve, hogy egy szóban forgó dokumentum címkéinek meghatározásához az adott dokumentumhoz bizonyos szempontok alapján hasonló dokumentum kulcsszavai közül választ. Az ilyen eljárások előnye, hogy a hasonló dokumentumból vett kulcsszó nem feltétlen van jelen a vizsgált dokumentumban, vagyis támogatja az absztrakt (a vizsgált dokumentum szövegében ténylegesen elő nem forduló) kulcsszavak generálását. Ugyanakkor az efféle eljárások hátránya is épp abból ered, hogy az ajánlott kulcsszavak a hasonló dokumentumok kulcsszavai közül kerülnek ki, vagyis csak a dokumentumhalmaz szintjén legalább egy alkalommal kulcsszóként definiált kifejezések kinyerésére képes, a kulcsszavak dinamikájához nem tud alkalmazkodni.

A *kulcsszókinyerő* módszerek az előbbiekkel ellentétben az aktuálisan vizsgált dokumentum szövegéből nyerik ki a kulcszavakat, olyan módon, hogy egy alkalmasan választott stratégia mellett legenerálják a dokumentum összes potenciális kulcsszójelöltjét, majd ezeket egy gépi tanulási modell alapján rangsorolják, a rangsor első elemeit pedig kulcsszóként kezelik. Ebben az esetben már nem áll fenn az a megszorítás, hogy egy tanítóhalmazbeli dokumentum tényleges kulcsszavai között szerepelnie kelljen a kulcsszójelölteknek, tehát az ebbe a csoportba tartozó algoritmusok képesek alkalmazkodni a kulcsszavak időbeli változásához. Ugyanakkor az is elmondható, hogy mivel a kulcsszójelöltek egytől egyig a dokumentum szövege alapján lettek generálva, így olyan kulcsszavak kinyerésére, amelyek a dokumentum szövegében nem voltak leírva, az ilyen eljárások önmagukban nem képesek, ráadásul ha a cél egy teljes dokumentumhalmaz felkulcsszavazása, a szinonimák és szemantikailag hasonló kulcsszavak egységes kezelésének is külön figyelmet kell szentelnünk.

A kulcsszavak meghatározásának egyik leggyakoribb célja a tudományos publikációk kulcsszavakkal való ellátása, az idei évben a SemEval konferencia egy versenyt is hirdetett a témában [5]. Ezen felül számos publikáció jelent már meg a kulcsszavazás témakörén belül speciálisan az angol nyelvű tudományos publikációk kulcsszavazásával foglalkozva. [4] webes keresésekkel igyekezett pontosabb eredményekre jutni, míg [7] a tudományos cikkek strukturáltságát és a bennük szereplő rövidítések fontosságát hangsúlyozta.

3 Módszertan

Az automatikus címkézés magyar nyelvre való adaptálásához felügyelt tanítás mellett végrehajtott *kulcsszókinyerési* módszertant alkalmaztunk. Tanuló algoritmusnak a generatív logisztikus regressziót választottuk a kulcsszójelöltek poszteriori valószínűségének meghatározására. Ebben a fejezetben részletesen bemutatjuk a jellemzőtér építését megelőző előfeldolgozó lépéseket, valamint magát a jellemzőteret.

3.1 Előfeldolgozás

Az előfeldolgozás magában foglalta a „nem hasznos” dokumentumrészek elhagyását, a fejezethatárok meghatározását, valamint a kulcsszójelöltek és velük kapcsolatos statisztikák kigyűjtését.

A PDF formában fellelhető publikációk feldolgozásának első lépése a szöveges tartalmuk kinyerésére irányult, melyhez a szabadon elérhető PDFBox[1] konvertert használtunk. Következő lépésként a szöveg megtisztítását hajtottuk végre: az egyes dokumentumon belül túl gyakran előforduló szövegrészeket – mint amilyenek a fejlécben található szövegek – egyszerűen eltávolítottuk a feldolgozandó szövegek közül, ily módon megtisztítva szövegeinket a fölöslegesen mondatközbe beékelődő szövegektől. Más jellegű megközelítést igényelt a táblázatok tartalmának kezelése, melyek szintén képesek mondatok belsejébe beékelődni a PDF konvertálása során. Ezeket a tartalmakat a sorok (token- és karakter-) hosszára számított statisztikákat és reguláris kifejezéseket alkalmazva lehetett eredményesen eltávolítani a folyó szövegből. A következő feladat az előálló nyers szövegek nyelvi elemzése volt. A nyelvi elemzést a magyarlanccal [10] végeztük el, amely egyúttal a lemmatizálásért is felelős volt.

A szövegre irányuló előfeldolgozó lépések végeztével tehát előálltak a dokumentumok tisztított, szöveges részeinek szófaji kódokkal ellátott verziói. Ezt követően lettek meghatározva a kulcsszójelöltek az egyes dokumentumokra. Kulcsszójelöltként kezeltük azokat az 1 és 4 tokenhossz közötti kifejezéseket, melyek melléknevekből és főnevekből álltak csupán, és se nem kezdődtek stopszóval, se nem végződtek stopszóval vagy melléknévvel.

[1] http://pdfbox.apache.org/

3.2 Jellemzőtér

A jellemzőtér kidolgozása során a hagyományos jellemzők beépítésén túl új jellemzők hozzáadott értékét is megvizsgáltuk. [9] az egyes kulcsszójelölteket azok tf-idf mértékével, valamint a dokumentumbeli első előfordulásuk relatív pozíciójával jellemezte. Cikkük megjelenése óta ezekre tekint a kulcsszókinyeréssel foglalkozó irodalom standard jellemzőkként. Munkájuk egy kiterjesztésében [6] már felhasználták a kulcsszójelöltek tokenszámát is a kifejezések kulcsszóként való előfordulásának leírására. A fenti jellemzők továbbiak mellett a mi rendszerünkben is szerepet kaptak.

[7] angol nyelvű számítástudományi publikációkon megmutatta, hogy a rövidítések feltérképezése segítségünkre lehet a kulcsszavak meghatározására. Ezért mi úgy jártunk el, hogy kigyűjtöttük a publikációk szövegeiből az összes olyan tokent, amely több nagybetűs karaktert tartalmazott, mint kisbetűst, egy-egy több token hosszú kulcsszójelölt esetében pedig döntést hoztunk, hogy az kiterjesztése lehet-e az őt tartalmazó dokumentum valamelyik rövidítésének, oly módon, hogy ugyanazzal a betűvel kezdődnek, majd pedig ugyanabban a sorrendben fordulnak benne elő egyazon rövidítés betűi.

A kulcsszójelöltek dokumentumbeli elhelyezkedését az első előfordulás relatív pozícióján túl a pozíciókban mutatkozó szórás értéke alapján is vizsgáltuk. Miután vettük egy kulcsszójelölt összes dokumentumbeli előfordulását, egyszerűen kiszámítottuk azok értékeiben jelentkező szórás nagyságát, és ezt az értéket adtuk meg a kérdéses kulcsszóaspiránsnak az adott jellemzőre nézve. Egy másik statisztikai módszerrel számított jellemző a PMI (Pointwise Mutual Information) [2] volt, amely a több tokenből álló kifejezések alkotóelemeinek együttes előfordulási gyakoriságának mértékét vizsgálni. A mérőszám csak akkor adott 1 értéket, ha a kifejezés minden egyes tokenje kizárólag egymást követve fordult elő a dokumentumon belül, valószínűvé téve ezáltal, hogy egy többszavas kifejezéssel van dolgunk.

A pozíciókkal kapcsolatos tulajdonságokat a tokenpozíciókon kívül szekcióbeli elhelyezkedésre is alkalmaztuk. Hasonlóan az első előfordulás tokenhossz függvényében történő relatív meghatározásához, kiszámítottuk a szekciók arányában egy kulcsszójelölt relatív pozícióját. Ezen túl, hasonlóan a tf-idf mértékhez, kiszámításra került minden kulcsszóaspiránshoz egy sf-isf (szekciófrekvencia – invertált szekvenciafrekvencia) érték is,

$$sf\text{-}isf(t_i, d_j) = sf(t_i, d_j) * isf(t_i) \quad \textbf{(1)}$$

alapján, amelyből $sf(t_i, d_j)$ azt mutatja meg, hogy a j-edik dokumentum szekcióinak milyen arányában van jelen a t_i kifejezés, $isf(t_i)$ pedig azt határozza meg, hogy a korpuszban lévő összes szekció mekkora részében szerepel t_i kifejezés.

A jellemzők egy másik fontos részhalmaza a kulcsszójelöltek nyelvi elemzéséből származó információkat használta föl. A szófaji kódokért felelős nominális jellemző az adott kulcsszójelölt szófaji kódjának sorozatát tartalmazta, a *magyar pártfejlődés* kifejezés esetében értéke *melléknév+főnév* volt. A szófajok egyszerű nyilvántartásán túl egy másik jellemző a kulcsszójelölt összes előfordulási formáján belül annak az arányát írta le, hogy mekkora valószínűséggel szerepelt a kifejezésben főnév. Ezeken túlmenően jellemzőként tekintettünk arra is, hogy az adott kulcsszóaspiráns hány elemet tartalmazott egy előre definiált magyar nyelvű stopszólistáról. A motiváció

ezen jellemző használata mögött az volt, hogy magas értéke egy kifejezés kulcsszóként való viselkedés ellen szólhat.

Fontosnak éreztük azt a tényt is egy kulcsszójelölttel kapcsolatban, hogy az azt tartalmazó mondatok közül mennyi esetében fordult elő hivatkozás. Az ez alapján a statisztika alapján számított bináris jellemző igaz értéket vett fel, ha a kulcsszóaspiráns dokumentumában legalább egy olyan mondat szerepelt, amely mind a kulcsszójelöltet, mind pedig legalább egy hivatkozást tartalmazott. Egy további, a dokumentum szerkesztéséből nyerhető információ volt számunkra, hogy egy-egy kulcsszójelölt szerepelt-e a publikáció címében, avagy legalább egy fejezet főcímében. A kulcsszójelöltekkel kapcsolatban fölhasználtuk azon információkat is, hogy más tanító adatbázisbeli dokumentumon az adott kifejezés szerepelt-e kulcsszóként, illetve, hogy a magyar nyelvű Wikipédián található-e azonos névvel szócikk.

Végül egy kifejezés kulcsszóként való szereplését egy dokumentumban nagymértékben meghatározza, hogy milyen igék társaságában szerepel együtt, azaz hogy milyen kontextusban szerepel. Éppen ezért a tanítás folyamán legyűjtöttük a tanítóhalmazban szereplő igéket és azok gyakoriságát, majd a tanítódokumentumok méretének függvényében meghatározott küszöbszámnál többször előforduló igékre szorítkozva folytattuk vizsgálódásainkat. Egy kulcsszójelölt esetében megvizsgáltuk, hogy melyek azok a tanítóhalmazon legalább százszor szereplő igék, amelyekkel a saját dokumentumában együtt előfordultak, majd azon igéknek megfeleltethető jellemzőket, ahol ez a mennyiség 0-tól különböző volt, egyenlővé tettük az együttes előfordulásuk számával.

4 Adatbázisok

A felügyelt tanulás végrehajtásához, valamint a kiértékelés könnyebbé tételéhez kulcsszavakkal ellátott magyar nyelvű online folyóiratokat gyűjtöttünk össze. A feladat nehezebbnek bizonyult, mint azt elsőre gondoltuk, hiszen sok publikációs forrás nem tartalmaz kulcsszavakat, mi több, a kulcsszavakat tartalmazók közül több sem volt végül alkalmas a feldolgozásra, azok többnyelvűsége (angol-magyar) vagy a PDF-dokumentumok minőségbeli hiányosságai miatt. Végül méréseinket a Politikatudományi Szemle[2] archívumán, valamint a Magyar Neveléstudományi Konferencia 2009-es konferenciakiadványa[3] alapján hajtottuk végre.

A Politikatudományi Szemle szerkesztősége a 2007-es évfolyammal kezdődően vezette be részlegesen a megjelent cikkek kulcsszavazását, így adathalmazunkat az ez idő alatt megjelent, kulcsszavakkal bíró cikkeik képezték. A PDF-dokumentumok tartalmát sima szöveges fájlba konvertáló eszköz a 86 kulcsszavazott publikáció esetében 6 alkalommal nem járt sikerrel szerkesztési hibáknak köszönhetően, így legvégül 80 dokumentum képezte vizsgálódásunk tárgyát.

A teljes dokumentumhalmazhoz 351 kulcsszó lett hozzárendelve összesen 412 alkalommal, ami átlagosan 5,15 kulcsszót jelent dokumentumonként. A kulcsszavak között 79 absztrakt kulcsszó volt, ami azt jelenti, hogy ennyi esetben volt egy kifeje-

[2] http://www.poltudszemle.hu/archivum,7.html
[3] http://www.nevelestudomany.hu/onk2009/

zés úgy megadva kulcsszónak, hogy a publikáció szövegében nincsen megemlítve maga a kifejezés. Az ilyen esetek az összes kulcsszó 19,17% tették ki tehát, méghozzá úgy, hogy az összes absztrakt kulcsszó egyedi volt abban az értelemben, hogy nem volt olyan kulcsszó, amely egynél többször fordult volna elő absztrakt minőségében (nem absztraktként ettől még többször is előfordulhatott). A 80 dokumentumból összesen 146508 kulcsszójelöltet generált rendszerünk, ami dokumentum szinten átlagosan 1831,35 kulcsszóaspiránst jelent. Generatív modellünk feladata minden dokumentumra ezeknek a jelölteknek a rangsorolása volt.

1. táblázat: A leggyakoribb politikatudományi kulcsszavak listája.

Kulcsszó	Előfordulási gyakoriság	Absztrakt előfordulás
pártok	5	0
politikai kommunikáció	5	0
politikatudomány	5	0
kormányzás	4	0
média	4	0

2. táblázat: Példa absztrakt kulcsszavakra.

Kulcsszó	Előfordulási gyakoriság	Absztrakt előfordulás
nacionalizmus	2	1
választói magatartás	2	1
kormányzat minősége	1	1
komparativisztika	1	1
témakisajátítás	1	1

A neveléstudománnyal foglalkozó konferenciakiadvány a 2009-es Magyar Neveléstudományi Konferenciára elfogadott publikációk absztraktjait tartalmazta, de a szerzők által meghatározott szabad szavas kulcsszavakat nem minden esetben (a témakörök megjelölése általános volt, de azok használata nem lett volna megfelelő számunkra). A kiadványban végül 19 darab kulcsszavazott tartalmi összefoglaló volt található, melyekhez összesen 63 egyedi kulcsszó volt rendelve, ami átlagosan 3,32 szerzői kulcsszót jelent dokumentumonként. A cikkek szerzői által meghatározott kulcsszavak közül 19 volt jelen absztrakt minőségben, a politikatudományosénál magasabb, 25,4%-os absztraktkulcsszó-arányt eredményezve mindezzel. A neveléstudománnyal foglalkozó dokumentumok esetében a rendszerünk által generált kulcsszójelöltek száma 3169 volt, ami átlagosan 166,79 kifejezést jelent dokumentumokként.

5 Kísérletek

A kiértékelés során keresztvalidációt alkalmaztunk mindkét korpusz esetében. A kiértékelések között egyaránt végrehajtottunk szigorú egyezésen alapulót, valamint a kulcsszavak speciális természetére való tekintettel emberi ellenőrzést is. Mindezen

felül az emberi ellenőrzés megbízhatóságának tesztelésére megvizsgáltuk az annotátorok közötti egyetértés szintjét is.

5.1 Annotátorok közötti egyezés

A kézi kiértékelést két nyelvész egymástól függetlenül végezte. Feladatuk az volt, hogy egy dokumentum automatikus kulcsszavazásának bírálatakor döntést hozzanak egyrészről az eredeti szerzői kulcsszavak automatikus kulcsszavakkal való lefedettségéről, valamint az automatikus kulcsszavaknak az adott dokumentum témájába vágóságáról, azaz pontosságukról.

Az annotációk közötti egyezés mértékét a politikatudományi témájú publikációk esetében egy baseline, illetve a végső módszerre is teszteltük, a neveléstudományi cikkek esetében pedig a kiterjesztett jellemzőteret használó modell kiértékelését mértük. Fontos kiemelnünk, hogy ezen mérések során a két független annotátor az automatikus rendszer által kinyert kulcsszavakról hozott döntéseket, így ezen döntések egymáshoz való hasonlóságának vizsgálatán keresztül a számukra kitűzött feladat jól definiáltságát állt módunkban megvizsgálni.

Az annotátorok közötti egyetértés jellemzését az azonos módon megítélt kulcsszavak arányán túl κ-mértékkel [3] is elvégeztük, mely egy olyan statisztikai mutató, amely a tapasztalt egyezési szintet megpróbálja korrigálni a véletlennek köszönhető egyezéssel. A κ-mérték értéke -1 és 1 között mozog, -1-et akkor veszi föl, ha az annotátorok jelölései teljes ellentétben állnak egymással, 1 értéke pedig akkor lesz, ha tökéletes összhang tapasztalható az annotátorok között. Általánosságban, a 0,4 és 0,6 közötti értékű egyezéseket közepes szintűknek, a 0,6 és 0,8 értékek között mozgókat kielégítőnek, a 0,8 és 1,0 közöttieket pedig közel tökéletes egyezésűként szokás jellemezni.

3. táblázat: Annotátorok közötti egyetértés.

	Egyezés	**κ-mérték**
Baseline fedés (politikatudomány)	90,42%	0,7689
Baseline precízió (politikatudomány)	81,68%	0,55411
Végső fedés (politikatudomány)	85,92%	0,7223
Végső precízió (politikatudomány)	81,92%	0,6256
Végső fedés (neveléstudomány)	92,06%	0,8382
Végső precízió (neveléstudomány)	91,10%	0,7206

5.2 Rendszereredmények

A szigorú egyezés alatt az automatikusan kinyert kulcsszavak és az eredeti szerzői kulcsszavak szótöveinek egyezését követeltük meg az elfogadáshoz. Ebben az esetben csak a szerzői kulcsszavakhoz mért pontos egyezés (fedés) mérésére volt lehetőség, a kizárólag az automatikus kulcsszavak között szereplő, a cikk témáját különben jól összegző kifejezések nem kerültek elfogadásra.

Éppen ezért az emberi kiértékelést is szükségesnek éreztük, hiszen könnyen előfordulhatott, hogy az eredeti kulcsszavak jelentéstartalmát lefedő, de attól eltérő formában álló kifejezést nyert ki rendszerünk, ám ilyen esetekben a szigorú egyezésen alapuló kiértékelés hibás kifejezésként kezelte az egyébként szemantikája alapján elfogadható kifejezéseket is. A szigorú, valamint a megengedőbb, emberi kiértékelést figyelembe vevő rendszereredmények a 4., illetve az 5. táblázatban olvashatók.

Baseline módszerünk jellemzőtere a standard módszert követte, azaz az egyes kifejezéseket az azt tartalmazó dokumentum alapján számított tf-idf mértékkel és dokumentumon belüli legkorábbi előfordulásának relatív pozíciójával jellemezte.

4. táblázat: A szigorú kiértékelés eredményei.

Rendszer	Fedés
Baseline (politikatudomány)	13,02%
Teljes jellemzőtér (politikatudomány)	31,07%
Teljes jellemzőtér (neveléstudomány)	30,16%

5. táblázat: Kézi kiértékelés eredményei.

Rendszer	Pontosság	Fedés	F-mérték
Baseline (politikatudomány)	36,79%	33,91%	0,3529
Teljes jellemzőtér (politikatudomány)	49,76%	54,12%	0,5185
Teljes jellemzőtér (neveléstudomány)	24,15%	46,03%	0,3168

6 Diszkusszió

Dolgozatunkban magyar nyelvű tudományos publikációk eltérő területein (politika- és neveléstudomány) működőképes kulcsszókinyerő rendszert mutattunk be, amely eredményei a baseline rendszert jelentősen meghaladták, valamint az idei évben angol nyelvű tudományos publikációkra meghirdetett hasonló témájú verseny végeredményeit összegző cikk [5] eredményeivel összehasonlítható eredményeket produkált a szigorú kiértékelés alkalmazása mellett.

A 4. táblázat eredményeiből kitűnik az is, hogy a szigorú kiértékelés esetében (fedés) hasonló eredmények születtek mind a politikatudományi, mind pedig a neveléstudományi cikkek esetében. A 30% körül mozgó eredmények kapcsán fontos megjegyezni, hogy a két korpuszon belüli absztraktcímke-eloszlásoknak köszönhetően a szigorú egyezéssel elérhető eredmények maximális értékei 80,83%, illetve 74,6% voltak a politikatudományi, illetve a neveléstudományi cikkekre nézve.

Ha az eredményeket az 5. táblázatban látható emberi kiértékelés szintjén vetjük össze, akkor már eltérés tapasztalható – elsősorban pontosság tekintetében – a két korpuszon elért eredményességet illetően. A politikatudományi publikációk esetében tapasztalható jobb eredmények annak tudhatók be, hogy azok esetében a teljes cikk szövege rendelkezésre állt a kulcsszavak meghatározása során, míg a neveléstudományi dokumentumok között kizárólag a teljes publikációk absztraktjait tudtuk használni, jellemzőterünknek pedig a hosszabb dokumentumok (több információ) kedveznek.

Az annotátorok kiértékelései közötti egyezés elemzése is érdekességeket mutat. A 3. táblázatból kiolvasható, hogy a fedéssel kapcsolatos döntéseket minden esetben jóval nagyobb összhang mellett voltak képesek meghatározni, mint az automatikus kulcsszavak helyességére irányuló döntéseket. Az eltérő típusú jelölésekben mutatkozó különbségeken túl megfigyelhető volt még az is, hogy a jobb minőségű automatikus címkézést nagyobb annotátorok közötti egyetértés is kísérte, hiszen ekkor a több jó kulcsszó miatt vélhetően kevesebb kérdéses szituációban kellett dönteniük az annotáció során.

Jövőbeli munkaként fontosnak érezzük kulcsszókinyerő rendszerünk kiterjesztését abban a tekintetben, hogy a dokumentum szövegében le nem írt (absztrakt) kulcsszavak meghatározására is képes legyen, valamint a továbbiakban a korpuszszintű kohézió (kulcsszavak normalizálása) szem előtt tartásával is szeretnénk foglalkozni.

Köszönetnyilvánítás

A cikk szerzői köszönettel tartoznak a kulcsszavak annotációját végző nyelvészeknek, Vincze Veronikának és Almási Attilának. A kutatást – részben – a TEXTREND kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Björk, B-C., Roos, A., Lauri, M.: Scientific journal publishing: yearly volume and open access availability. Information Research, ISSN 1368-1613, Vol. 14, No. 1 (2009)
2. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the ACL. ACL, New Brunswick, NJ. (1989) 76–83
3. Cohen, J. A.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement Vol. 20 No. 1 (1960) 37–46
4. Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., Nevill-Manning, C, G.: Domain-Specific Keyphrase Extraction. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI99) (1999) 668–673
5. Kim, S. N., Medelyan, O., Kan, M-Y., Baldwin, T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010) (2010)
6. Medelyan, O., Witten, I H.: Thesaurus based automatic keyphrase indexing. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries (2006) 296–297
7. Nguyen, T. D., Kan, M-Y.: Keyphrase extraction in scientific publications. In: Proceedings of International Conference on Asian Digital Libraries (ICADL07) (2007) 317–326
8. Taleb, N. N.: The Black Swan, Chapter 7. ISBN: 1400063515 (2007)
9. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C, G.: Kea: Practical Automatic Keyphrase Extraction. In: ACM DL (1999) 254–255
10. Zsibrita J., Nagy I., Farkas R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 394–395

Bibliográfiai hivatkozások automatikus kinyerése[1]

Váradi Tamás[1], Pintér Tibor[1], Mittelholcz Iván[1], Peredy Márta[1]

[1] MTA Nyelvtudományi Intézet
Benczúr utca 33., 1068 Budapest
{varadi, tpinter, mittelholcz, mperedy}@nytud.hu

Kivonat: A Magyarországon megjelentetett társadalomtudományi folyóiratok tanulmányaiból automatikusan kigyűjtött hivatkozások adatbázisba rendezése jelentős segítség a tudomány számára. A heterogén források által többféle struktúrában megjelenített adatok elemzését és azonos formátumba rendezését a szabad felhasználású NooJ szoftver segítségével végeztük. A folyamat valódi kihívása az adathalmaz elemeinek, valamint a hivatkozások típusának automatikus felismerésében rejlik. A külön-külön létrehozott (ugyanakkor egymással kombinálható) NooJ-grammatikák szerepe a hivatkozások egyes elemeinek felismerése és annotálása. Az automatizált folyamat kimeneteként létrejövő XML-elemek még utólagos kézimunkára szorulnak, részint a hivatkozások rossz minősége miatt (hiányos hivatkozások, szabványoktól eltérő hivatkozások), részint a folyamat formalizált volta miatt (bizonyos hivatkozások automatikusan több hivatkozástípusba is besorolódnak). A BibTex-szabványosítás előtt egyértelműsítő algoritmusokat és/vagy kézi erőt kell használni.

1 Háttér

Munkánk célja a magyar társadalomtudományi folyóiratokban történő hivatkozások adatbázisának létrehozása és folyamatos, automatizált bővítése. Az MTA Nyelvtudományi Intézetben az OTKA támogatásával végzett munka jelentőségét az adja, hogy Magyarországon nem létezik ilyen átfogó adatbázis (bár kétségtelen, hogy hasonlók léteznek, ám lefedettségben és a létrehozás módszerében projektünk ez idáig Magyarországon egyedülálló). Egy ilyen jellegű adatbázis nagyon hasznos szerepet tölthetne be a társadalomtudományi folyóiratok színvonalának, valamint a kutatók publikációs tevékenységének megbízható elbírálásában.

Az adatok mennyisége és sokfélesége nélkülözhetetlenné teszi a korszerű számítógépes technológia használatát. A magyarországi hivatkozás-adatbázisok építésének (MTMT adatbázis) gyakorlatában főként a kézzel történő adatbevitel van elterjedve, ami hosszútávon nem járható út, mivel időigényes és a manuális adatbevitel miatt nehézkes (nem beszélve a többletmunkáról, hiszen az adatbázisba egy már írásban megjelent adathalmazt írnak be – manuálisan).

[1] Jelen munka az OTKA PUB-F, 81666 számú pályázat keretében készült.

2 A feladat

A hivatkozások automatikus kigyűjtésének alapvetően két módja van: a mintaalapú szövegkinyerés (alapvetően előre létrehozott karaktersorok alapján történő szegmentálás és adatkinyerés) és a gépi tanulási technikákon, valamint statisztikai módszereken alapuló szövegbányászat [1, 2, 3, 4, 5]. Az OTKA által támogatott projektünkben elsősorban karakteralapú szövegfelismerést alkalmazzuk, melyet a hivatkozások állandó elemeiből összeállított szótárakkal egészítünk ki (egyúttal további szótárak létrehozását is tervezzük).

A munka az alábbi szakaszokra bontható:

i. a társadalomtudományi folyóiratok körének feltérképezése, a bennük található hivatkozások szöveges alakban történő összeállítása,
ii. a szabad szövegű hivatkozások bibliográfiai elemeinek (pl. szerző, cím, kiadó, kiadás helye stb.) automatikus azonosítása és annotálása,
iii. az adatbázis tényleges létrehozása,
iv. az adatbázis online felületen elérhetővé tétele,
v. az adatbázis folyamatos frissítése.

Jelen előadásunkat a ii. kérdésnek szenteljük, amely a nyelvtechnológiai kihívásokat tartalmazza.

Jelenleg egy olyan 199 folyóiratot számoló mintán dolgozunk, amely a Magyarországon kiadott társadalomtudományi folyóiratok átfogó metszetét teszi ki. Az eddig összegyűjtött anyag mintegy 34 ezer tanulmány, ami akár kézi munkával is feldolgozható lenne, ám az adatbázis folyamatos bővítésének igénye elengedhetetlenné teszi a referenciák kinyerésének automatizálását.

A hivatkozások kinyerését két szakaszban végezzük. Az első lépés során a folyóiratszövegekből kivesszük a hivatkozásokat, a második lépés alatt elvégezzük ezek annotációját. Az első lépésben a következő nehézségek merülhetnek fel, amelyek a ráfordított többletmunka miatt jelentős mértékben lassíthatják a feldolgozást (egy nyomós érv a stíluslapok konzekvens betartása mellett): a referenciák nem csak szövegvégi helyzetben fordulnak elő, a tanulmányvégi közlés mellett gyakori a lapalji hivatkozás (technikailag „lábjegyzetelés”), valamint előfordul még szövegközi megjelenítés is (a folyó szövegben zárójelekkel elhatárolva találhatók a hivatkozott publikációra vonatkozó adatok). A hivatkozásokat tehát nem egyszerűen a folyóiratbeli tanulmányvégi pozíciójuk, hanem jellegzetes alkotóelemeik (személynevek, évszámok, tipikus funkciószavak – pl.: in, szerk. – stb.) és szintaktikai felépítésük (az elemek sorrendje, a köztük lévő írásjelek) teszik felismerhetővé.

A második lépés a szövegből kinyert hivatkozások feldolgozása. Ez az első feladatnál jóval bonyolultabb, mivel ekkor már nem szorítkozhatunk csupán az egyes jellemzők felismerésére: itt a teljes hivatkozás pontos elemzésére van szükség (ami technikailag a folyó szövegek XML-annotációját jelenti). Az egyes elemek elemzésekor azonosítani kell egyrészt a hivatkozás típusát (könyv, fejezet, folyóirat stb.), másrészt az adott típushoz tartozó jellegzetes hivatkozáselemeket (pl. név, dátum, cím, kiadás helye, oldalszám stb.). Ez utóbbi nem egyszerű, főleg az egyes elemek közti elválasztóelemek folyamatos variációi, illetve a különféle stíluslapok megléte, valamint az egyéni hivatkozásstílusok használata miatt [5]. Az annotációt első körben

XML-elemekkel, illetve azok konverziója után a BibTex-szabvány címkéivel végezzük (bár némely esetben – pl. a név elem – utóbbinál részletesebbek vagyunk).

3 A megoldás

3.1 A szoftver

A hivatkozások szintaktikai elemzéséhez a szabad felhasználású NooJ szoftvert használjuk, mely egy, a grammatikai elemzés támogatására létrehozott fejlesztőkörnyezet [7]. Munkánk szempontjából a NooJ legfontosabb előnyeit a következőképpen lehet összegezni:

- moduláris felépítése révén alkalmas az egyes elemek lokalizálására,
- az egyes modulok és gráfok könnyedén kombinálhatók és a célnak megfelelően módosíthatók,
- grafikus ábrázolásmódjának köszönhetően átláthatóbb a más szövegkinyerésre és szövegfeldolgozásra alkalmas programoknál,
- gyors,
- kombinálható a NooJ-ban íródott magyar szintaktikai elemzővel,
- az elemzésbe szótárak is bevonhatók (ez fontos lehet például a magyar személy- és keresztnevek, a kiadók esetében vagy például az egyes hivatkozáselemek identifikálásában).

A hivatkozások annotálása közben külön-külön NooJ-grammatikák ismerik fel a jellegzetes hivatkozáselemeket (pl. név, dátum, cím, kiadás helye, oldalszám stb.), majd ezek megfelelő szintaktikai kombinációi illeszkednek a hivatkozások különféle típusaira. Az elemzési szempontokat tartalmazó algoritmusokból, ún. lokális grammatikákból állnak, amelyek a NooJ grafikus felületén szerkeszthetők (l. az 1. és 2. ábrát). A gráfok – elemzésünkben – elsősorban karaktersorokat meghatározó mátrixokat, illetve szótárakat tartalmaznak, de alkalmasak morfológiai és szintaktikai információk tárolására, illetve visszakérésére is. Az elemzés (részleges vagy teljes) találatairól a program konkordancialistát készít, amelyben az egyes találatokat a gráfokban kódolt XML-elemekkel lát el.

3.2 Az eljárás

Az annotálás automatizálása attól válik izgalmas feladattá, hogy az egy típushoz tartozó hivatkozások is roppant sokfélék lehetnek. Eltérő lehet a szövegelemek sorrendje, illetve az ezeket elválasztó írásjelek használata.

A fő kérdés tehát az, hogy melyek azok a legfőbb vonások, amely alapján az emberi intelligencia képes típusokba sorolni és annotálni egy-egy hivatkozást. A felismerőmodulok létrehozásakor abból a feltevésből indulunk ki, hogy a feladatot pusztán felszíni formai mintákra támaszkodva, a jelentésre való bárminemű hivatkozás nélkül kell megoldanunk. A puszta karaktersorozatok felismeréséhez a tesztek

folyamán nem használtunk szótárakat, ugyanakkor az egyes NooJ-modulok összeállításakor bizonyos elemeket előre bekódoltunk (pl. a hivatkozások egyes elemeinek felismerésében segítő *eds., and, &, In:, in:, in* stb. formánsokat). A bibliográfiai tételek egyes elemei (dátum, névkifejezés, bizonyos határoló központozó jegyek) viszonylag jól felismerhetőek, bár a határoló elemek inkonzekvens használata, illetve azok lehetséges kombinációi okoznak némi fejfájást. Mindemellett az is lényeges, hogy a hivatkozáson belül a cím felismerése csakis annak teljes körű szemantikai értelmezése révén lenne elvégezhető, ez pedig meghaladja jelenlegi tudásunkat (bár a többi elem felismerése nagyban segíti a cím felismerését is).

3.3 NooJ-grammatikák

Összhangban a hasonló témájú nemzetközi kutatásokkal [5] – a következő grammatikákat hoztuk létre:

- „név" elem: keresztnév, vezetéknév, eseleges középső név előfordulásának, a köztük levő összekötő elemeknek, valamint a szerzőség típusára (szerző, szerkesztő, szerkesztők) utaló elemek különféle változatai
- évszám: a publikáció kiadásának évére utaló információk (számok és egyéb összekötő karakterek, valamint betűk összessége)
- kiadó neve: a kiadó(k) nevét lefedő karaktersorozatok (betűk és köztük lévő kapcsolatok)
- kiadás helye: a kiadás helyére vonatkozó karakterkombinációk összessége
- oldalszámok: oldalszámok és határoló elemeiknek összessége
- fejezetcím: a fejezetcím elemeit összesítő gráf (a különféle változatok miatt – kisbetű, nagybetű, különféle határoló karakterek, mint a pont, vessző – a cím mező pontos behatárolása szinte lehetetlen)
- könyvcím: a fejezetcímhez hasonló elemek halmaza, amelyek detektálásában a cím melletti elemek segítenek
- évfolyamszámozás: az évfolyam különféle megjelenítésének módozatai

Az egyes grammatikák közül a „cím" mező annotálása a legnehezebb. A benne megjeleníthető bármilyen karakter (betű, szám és határoló írásjelek[2]) miatt szinte lehetetlen pontos grammatikát írni. Mindaddig azonban, amíg pontosan összeállítjuk a mellette álló grammatikák szerkezetét, a „cím" mező is viszonylag sikeresen parszolható: ez azonban inkább a pozícióból, mint a grammatika helyes szerkezetéből adódik.

Különböző problémák forrása lehet a hivatkozás „név" eleme is. Bár a „név" mező után a hivatkozások általában határolóelemmel (vessző, pont, zárójel) vannak elkülönítve, az egyes határolóelemek mégsem határolnak eléggé. Ennek fő oka, hogy az itt megjelenő elemeket több helyen is használják. A helyes parszolás számára támpontot nyújthatna a potenciális mezők egymásutánisága is. A név elem egyértelműsítésében legnagyobb segítség lehetne a „név" után következő „évszám" – azaz „névnek" minő-

[2] Cím esetében a többi mezőtől eltérően bármilyen karakter bármilyen kombinációja elképzelhető.

sül, ami a kiadás éve előtt van –, azonban nem minden hivatkozásstílus helyezi a „név" után az „évszámot" (ugyanis előfordul a hivatkozás végén is).

Egy további probléma a „név" elem belső sokszínűsége, összetettsége. A többszerzős tanulmányok jelölése, a keresztnév iniciáléval történő jelölése – főként több keresztnév esetén – és az esetleges középső név vagy iniciálé csak tovább növeli az elemzések bizonytalanságát. Az elemzésnek ugyanakkor fel kell készülnie a határolóelemek sokféleségére, illetve inkonzekvens használatára is (pl. vessző és pontosvessző keverése, vagy csak vessző használata).

A „név" elem bonyolultsága sokszor túlelemzéseket is eredményez. A „név" elem pontosabb bontása (vezetéknév, középső név, keresztnév), valamint a lehetséges találatok számának növekedése miatt a túlelemzések jelentősen megszaporod(hat)nak (csak reguláris kifejezések használatával, szótárak nélkül nehéz pontos találatot kapni).

Az egyes mezők pontosabb meghatározásában nagy segítség lehet a mezőtípusokra jellemző elemek beazonosítása (az viszont már problémaként merül fel, hogy a *vessző*, *pont* és *pontosvessző* határolóelemként is szerepelhet – ráadásul stílusonként változó szerepben). Különféle hivatkozásstílusok alapján végzett elemzések azt mutatják, hogy a hivatkozások belső struktúrájában összesen 13-féle jelölő használható fel, amelyek detektálása megkönnyítheti a szabályalapú elemzést[3]:

1. táblázat: Jelölők típusai.

Határolóelemek	Mező
1. Vessző ,	név, évfolyam, szám, oldalszám *határolóelem
2. Pont .	név, oldalszám *határolóelem
3. Pontosvessző ;	név *határolóelem
4. Kettőspont :	évfolyam, szám, oldalszám
5. Gondolatjel – -	évfolyam, szám, oldalszám
6. Kötőjel -	évfolyam, szám, oldalszám
7. Kerek zárójel () ( )	év, évfolyam, szám, oldalszám
8. Szögletes zárójel []	sorozat
9. Idézőjel „" " " " "	cím
10. Három pont…	cím
11. Kérdőjel ?	cím
12. Felkiáltójel !	cím
13. Aposztróf '	név, cím

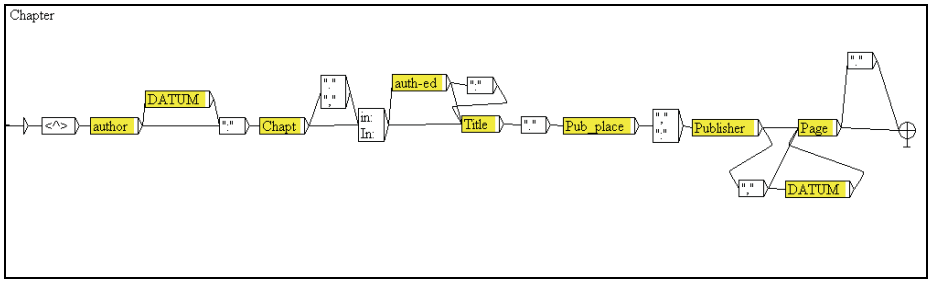
A grammatikák által lefedett mezők bizonyos, a grammatikára/hivatkozáselemre jellemző adatok – sztringsorok – alapján a hivatkozások egyértelműsíthetők. Ezek

[3] Bár a reguláris kifejezések használata elterjedt módszer a hivatkozások annotálására, egyes tapasztalatok szerint bonyolultabb struktúrák parszolására és rendezésére teljes mértékben alkalmatlanok [6].

elsősorban a névre, kiadás évére és helyére, valamint az oldalszámra vonatkozó adatok (pl. *eds*, *vol*, *in*, *kettőspont és szám kombinációja*).

Tesztjeink folyamán az alapelemek összekapcsolásával a következő hivatkozástípusokat parszoltuk teljes vagy részleges lefedettséggel: könyv, fejezet/tanulmány, folyóirat/folyóiratcikk, konferenciakötet – előadásunk a könyv és folyóirat-tanulmány felismerésének részeredményeit mutatja be. A tesztelés során szabadon választott 1027 darabból álló hivatkozásmintán a fenti elemek mellett a következő típusok fordultak elő: előadás, lexikon/szócikk lexikonban, szótár, kézirat/megjelenés alatt, diszszertáció/thesis.

A moduláris felépítés előnye, hogy az egyes grammatikák elemei beágyazhatók más-más grammatikákba (például a „név” mező, amely minden hivatkozástípusban azonos, és minden hivatkozástípusban előfordul), miáltal folyamatos fejlesztésük az összes hivatkozástípusra hatással van. Így a tesztek során, a találati lista lefedettségének növelése céljából azok folyamatosan módosíthatóak (pl. újabb – akár egyedi – hivatkozásstílus megjelenésekor). Természetesen ugyanazon elem többször is beágyazható ugyanazon grammatikába, akár opcionális változóként is (lásd a 1. ábra „dátum” elemét).



1. ábra. A „chapter” (fejezet) hivatkozást parszoló grammatika ábrája: a fő nodsor alatt és fölött lévő nodok kapcsolása fakultatív.

Mivel a hivatkozások egyes elemei formailag megegyeznek, megegyezhetnek (pl. a cím és az alcím, a tanulmány és az azt tartalmazó könyv szerzője), a parszolásban nemcsak a nodok tartalma (azaz a sztringsor és/vagy szótár), hanem az egyes elemek sorrendje is identifikál: pl. az 1. ábra „pubplace” eleme formailag megegyezik a „publisher”, vagy a „chapter” elemével, annotálásnál azonban a sorrend egyértelműsíti azt (az egyes grammatikák összetettsége, a határolóelemek többszöri előfordulása hosszabb – és egyben bonyolultabb – hivatkozásoknál csökkenti a parszolás eredményességét, elsősorban a határolóelemek hivatkozásmezőkben való előfordulása miatt).

Azonos hivatkozástípusok többféle realizációja (többféle stíluslap, illetve egyedi megoldások) miatt és a lehető legjobb eredmény elérése érdekében a fő hivatkozástípusoknak többféle változata is megjeleníthető: ezeket az elemzés XML-kimenete egyértelműsíti[4]. Sajnos a különféle stílusok bizonyos pontjainak átfedése, illetve a hivatkozások szerkezetének viszonylagos strukturálatlansága (a mintegy 200 folyóirat

[4] Az összes grammatika egyszerre futtatható, a keresés eredménye így a (részben vagy teljesen) annotált hivatkozás.

stíluskészlete közel sem nevezhető konzekvensnek – nemhogy egy folyóiraton belül, hanem sokszor egy tanulmányon belül sem) miatt előfordul, hogy a parszolás eredményeképpen ugyanazon hivatkozás több hivatkozástípusba is besorolódik. Ezeket az eseteket jelenleg manuálisan tudjuk csak egyértelműsíteni, azonban a megfelelő szótárak alkalmazása segítség lehet (ez utóbbi összeállítása folyamatban van, elsősorban a kiadók, a folyóiratok, valamint vezeték- és keresztnevek lexikonjainak kiépítését tervezzük). Az egyedi tartalmi és főleg stílusbeli megoldások miatt a grammatikákat folyamatosan építeni és finomítani kell: mindaddig, amíg el nem fogynak a különálló esetek.

4 Az eredmények

A tesztelések előtti kismintás pretesztelések folyamán fejlesztettük ki a jelenleg is használatban lévő grammatikákat (amelyek még tökéletesítésre szorulnak). Tesztjeinket az Akadémiai Kiadó által megjelentetett, általunk szabadon kiválasztott 16 társadalomtudományi folyóirat hivatkozásain végeztük. Az 57 753 db hivatkozásból véletlen mintavétellel kiválasztott 1027 darabos mintán[5] végzett NooJ-tesztelések a következő eredményeket és tapasztalatokat hozták.

Bár a minta nem reprezentatív (kiválasztásában nem játszottak szerepet az alapsokaság – mintegy 33 ezer darab hivatkozás – tulajdonságai), a tudományos folyóiratok leghangsúlyosabb magyarországi kiadójának társadalomtudományi kiadványait tekintve mindenképpen irányadó. A mintának választott hivatkozások típusai kézi szelekció után a következők:

2. táblázat: Teszthivatkozások tipizálása manuális annotálás után.

Hivatkozástípusok	db (N=1027)	% (N=100)
könyv (book)	411	40,0
folyóirat-tanulmány (article)	358	34,9
fejezet vagy tanulmány könyvben (chapter)	164	16,0
egyéb bibliográfiai tétel	94	9,1

A kézi válogatás után az egyes hivatkozáscsoportokon lefuttatott grammatikák a következő eredményt hozták.

Az első tesztelések során azt vizsgáltuk, milyen pontossággal ismerik fel a grammatikák az egyes hivatkozástípusokat. A hivatkozásokon belüli parszolás eredményei a hivatkozás hosszától és annak bonyolultságától függően eléggé szórtak. A hosszabb hivatkozások (elsősorban a már említett „név” és „cím” elemek összetettsége miatt) kevésbé pontosak, gyakoribb az egyes tételek „túlelemzése”.

[5] Az ábécésorrendbe rakott 57 753 hivatkozás minden 50. elemét kiválasztva kapott 1150 tételből manuálisan eltávolítottuk a nem hivatkozási elemeket, azaz a csonka tételeket, vagy a hivatkozások közé került nem referenciákat. Így kaptuk meg az 1027 db hivatkozást.

3. táblázat: Elemzések találatának pontossága.

	∑	T	jó találat (1)	jó találat (2)	pontosság (1) %	fedés (1) %	pontosság (2) %	fedés (2) %
teljes	769	192	87	119	45,31	11,31	61,98	45,31
article	358	33	30	-	90,91	8,38	-	-
book	411	192	89	-	46,35	21,65	-	-

A 3. táblázatban a grammatikák pontosságát és lefedettségét foglaltuk össze. Elemzésünkben csak a könyv és tanulmányok találati pontosságát érintjük, mivel – a már említett okok miatt – egy tételhez gyakran több elemzés is tartozik (192 hivatkozáshoz összesen 692 elemzés, tételenként átlag 3,6). Ezért kérdés, mit tekintünk helyes felismerésnek: ha egy tételnek van legalább egy helyes elemzése, vagy ha minden elemzése helyes (a típus felismerése szempontjából). A „teljes" nyelvtan esetében a tesztmintán egyszerre futtattuk le a rendelkezésünkre álló grammatikákat (az „article" és „book" esetében csak a folyóirat-tanulmányra és könyvre írt grammatikákat teszteltük). Kiértékelésénél szigorúbb módon jártunk el, csak azokat az eseteket fogadtuk el, ahol az adott hivatkozáshoz csak jó elemzések tartoztak. Az egyes grammatikák külön tesztelésénél erre nem voltunk tekintettel. Ezeknél az volt a célunk, hogy az adott ág teljesítményét önmagában nézzük, függetlenül attól, hogy esetleg más grammatikák (más típusba tartozóként) is felismernék az adott hivatkozást.

A „T" oszlop tartalmazza a tesztanyag azon tételeinek számát, amelyekre (legalább egyféleképp) illeszkedik a tesztelt nyelvtan. A „jó találat (1)" oszlop a ∑ oszlop (azaz a kézi annotálás eredményeit tartalmazó oszlop) hivatkozásain lefuttatott grammatikák találatát tartalmazza azon hivatkozásokra, ahol mindegyik elemzés jó volt a kérdéses hivatkozásnál. A „jó találat (2)" oszlop azon elemzések számát tartalmazza, ahol egy tételhez tartozó elemzések legalább egyike jó. A „pontosság (1)" értékei „jó találat (1)" és a „T" százalékos arányát, a „fedés (1)' értékei a „jó találat (1)" és a „∑" százalékos arányát, míg a „pontosság (2)" a „jó találat (2)" és a „T", a „fedés (2)" a „jó találat (2)" és a „∑" százalékos arányát mutatja.

4. táblázat: F-mérték.

	pontosság (1) %	fedés (1) %	pontosság (2) %	fedés (2) %	F-mérték (1) %	F-mérték (2) %
teljes	45,31	11,31	61,98	45,31	18,10	24,76
article	90,91	8,38	-	-	15,35	-
book	46,35	21,65	-	-	29,51	-

A teljes mintán lefuttatott elemzések értékeiből számított kiegyensúlyozás (F-mérték) értéke csupán kis mértékben tér el mindkét elemzés esetében (bár az értékek viszonylag alacsonyak). Abban az esetben, ha azokat a találatokat vesszük figyelembe, ahol az elemzés minden eredménye jó volt (1), a teljes elemzés F-mértéke 18%, míg abban az esetben, ha azokkal a találatokkal foglalkozunk, ahol (a túlelemzés

mellett) legalább egy helyes elemzés született, a teljes elemzés F-mértéke 25%. Az „article" hivatkozásokon lefuttatott, folyóirat-tanulmányok parszolására írt grammatika találati pontossága 15%, míg a „book" hivatkozásokon lefuttatott, könyvek parszolására írt grammatikáé 30%.

Tesztjeink során a NooJ-grammatikák pontossága elmaradt a jóval időigényesebb manuális elemzésnél (ez is erősíti a különféle szótárak összeállítását és az elemzésbe történő bekapcsolását). Az egyes grammatikák sikertelenségébe belejátszott az elemzett hivatkozások sajátosságainak (önálló, a folyóiratra jellemző stylesheet) tükröződése is.

5 Távlatok, tervek, fejlesztések

Bár a kismintás előtesztelések (azaz a grammatikák szerkesztése közbeni próbatesztek) minden esetben sikerekkel zárultak, a viszonylag nagyobb mintán elvégzett tesztek eredményei leginkább a lehetséges korrekciók kidolgozására és újabb módszerek kipróbálására ösztönöznek bennünket. A csupán mintaillesztés alapján működő parszolás úgy tűnik, a magyarországi hivatkozások esetében sem lesz sikeres (a Min-Yuh Day és munkatársai által elvégzett tesztekhez hasonlóan [6]), az elemzésbe szótárakat is be kell fűzni. Szótárakat elsősorban a „név", „kiadó" és „folyóirat" mezők esetében tudunk használni.

Egy további megoldás lehet a „név" mező egyszerűsítése (akár arra hivatkozva, hogy az évszám előtt álló karaktersorozatot tekintjük névnek – mindezt megszorításokkal, mivel nem minden hivatkozás szerepelteti az évszámot a szerző után –, és egy következő elemzés során egyértelműsítjük azt), mivel a tapasztalatok azt mutatják, hogy a „név" pontos parszolására való törekvés jelentős mértékben megnöveli az elemzések számát.

Az eddig elvégzett munkák haszna, hogy előrevetítik a további lépéseket. Amellett, hogy a hivatkozások adatbázisának építése a megfelelő ütemben halad (feldolgoztuk a kiválasztott folyóiratok tartalomjegyzékét és kidolgoztuk az adatbázis struktúráját), tökéletesíteni kell egyrészt a meglévő grammatikákat, illetve ki kell terjeszteni azokat a többi hivatkozástípusra is.

Szakirodalom

1. Bergmark, D.: Automatic extraction of reference linking information from online documents. TR2000-1821 (2000)
2. Constans, P.: Approximate textual retrieval, arXiv:0705.0751 (2007) http://arxiv.org/PS_cache/arxiv/pdf/0705/0705.0751v1.pdf
3. Constans, P.: A Simple Extraction Procedure for Bibliographical Author Field (2009) arXiv:0902.0755. http://arxiv.org/PS_cache/arxiv/pdf/0902/0902.0755v1.pdf
4. Giuffrida, G., Shek, E. C., Yang, J.: Knowledge-based metadata extraction from PostScript files. In: Proceedings of the Fifth ACM International Conference on Digital Libraries (2000) 77–84

5. Day, M.-Y., Tsai, T.-H., Sung, C.-L., Lee, C.-W., Wu, S.-H., Ong, C.S., Hsu, W.-L.: A Knowledge-based Approach to Citation Extraction. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2005). Las Vegas, Nevada, USA. (2005) 50–55
6. Day, M.-Y., Tsai, T.-H., Sung, C.-L., Hsieh, C.-C., Lee, C.-W., Wu, S.-H., Wu, K.-P., Ong, C.S., Hsu, W.-L: Reference metadata extraction using a hierarchical knowledge representation framework. Decision Support Systems Vol. 43 (2007) 152–167
7. Silberztein, M.: NooJ manual. Available at the website http://www.nooj4nlp.net (2003)

II. Párhuzamos korpuszok

Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására

Laki László János[1], Prószéky Gábor[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
1083 Budapest, Práter u. 50/a
laklaja@digitus.itk.ppke.hu
proszeky@itk.ppke.hu
[2] MorphoLogic
1116 Budapest, Kardhegy u. 5.
proszeky@morphologic.hu

Kivonat: Előadásunkban foglalkozunk a statisztikai gépi fordítás minőségének javításával, az egyre mélyebb hibridizáció alkalmazásával, majd az angol–magyar kísérletek mellett olyan, morfológiailag közelebb álló nyelvpárok bevonásával, mint a lovári cigány nyelv és a magyar. Az előadás második felében egy tisztán statisztikai alapon működő szövegannotáló rendszer létrehozásával és kiértékelésével foglalkozunk.

1 Bevezetés

Az informatika fejlődése szinte az összes tudományág számára új lehetőségek halmazát nyitotta meg, és ez nem volt másképp a nyelvészetben sem. Napjaink számítógépei segítségével képesek lettünk óriási méretű szöveges anyagok gyors és hatékony kezelésére, valamint feldolgozására. Könnyen belátható, hogy a szabályalapú módszerekkel nagyon nehéz a nyelvekben kifejezett kapcsolatokban meglevő törvényszerűségeket megfogalmazni, viszont kézenfekvő megoldásnak tűnik statisztikai módszereket használni ezen feladatok megoldására. Jelen munkánk célja a magyar és más nyelvek közti átalakítások vizsgálata, gyakorlati megvalósításuk, illetve a meglévő módszerek javítása és a témában lévő lehetőségek felmérése.

2 Statisztikai gépi fordítás

A statisztikai nyelvfeldolgozás elterjedt alkalmazása a gépi fordítás. A statisztikai gépi fordító (SMT) módszer nagy előnye a szabályalapú fordítással szemben, hogy az architektúra létrehozásához nem szükséges a nyelvek grammatikájának ismerete. A rendszer tanításához csupán egy kétnyelvű korpuszra van szükség, amelyből statisztikai megfigyelésekkel nyerjük ki a szabályokat. Az idegen nyelvpárokon elért komoly eredményekben bízva választottuk ezt a módszert, hogy létrehozzunk egy fordítórendszert – először az angol–magyar nyelvpárhoz.

A fordítás során az egyetlen, amit biztosan tudunk, az a mondat, amit le szeretnénk fordítani (forrásnyelvi mondat). Ezért a fordítást úgy végezzük, mintha a célnyelvi mondatok halmazát egy zajos csatornán átengednénk, és a csatorna kimenetén összehasonlítanánk a forrásnyelvi mondattal. Az a mondat lesz a rendszerünk kimenete, amelyik a legjobban hasonlít a fordítandó mondatra. Ez a hasonlóság lényegében egy valószínűségi érték, amely a nyelvi modellből és a fordítási modellből számolható. Rendszerünk felépítéséhez egy angol–magyar párhuzamos korpuszt használunk, amely a Hunglish [3] korpusz két alkorpuszából áll: a Literature és a Magazines nevű részekből (a továbbiakban LitMag). A LitMag korpusz 654 939 mondatot és 9 425 911 szót tartalmaz, így kisméretű korpusznak tekinthető.

Több módszert is megvizsgáltunk, melyek képesek párhuzamos korpuszból információt kinyerni. Végül az IBM modellek mellett döntöttünk, mivel, hatékony, viszonylag pontos, és a feladatnak nagyon jól megfelelő algoritmusnak bizonyultak. Ezen okból kezdtük használni a Moses keretrendszert [5], amelyik implementálja ezeket a modelleket. Ebben a rendszerben megtalálható a párhuzamos korpusz előfeldolgozása, a fordítási és nyelvi modellek létrehozása, a dekódolás, valamint a BLEU-pontra optimalizálás. A kiértékeléshez a BiLingual Evaluation Understudy (BLEU) módszert használtuk, amelynek lényege, hogy a fordításokat referenciafordításokhoz hasonlítja, majd pontozza őket. Eredményként egy 0 és 1 közötti valós számot kapunk. A fentebb jelzett tanítás után alaprendszerünk 0,1085 (10,85%) BLEU-pontot kapott.

Szembetűnő eredmény, hogy a más nyelvpárok közti fordítógépek eredményei sokkal jobbak az angol–magyar nyelvű fordítógépekhez képest: angol–francia 32%, spanyol–katalán > 40%, angol–spanyol 30% [2]. Ha jobban megvizsgáljuk az eredményeket, az is látszik, hogy az előbb felsorolt nyelvek között nagy hasonlóság van mind nyelvtanilag, mind a szavak szintjén. Kis eltérés van ugyanis a szórendben és a nyelvtani szerkezetekben. Ennek köszönhetően a gépi fordító rendszer nagyobb biztonsággal képes létrehozni a transzformációkat a két nyelv között, valamint a kapott fordítások is sokkal jobban fognak hasonlítani a referenciafordításokhoz. Ezzel szemben a magyar és az angol nyelv között igen jelentős a formai eltérés. Ennek köszönhetően, ha ugyanazt a rendszert használjuk angol–francia vagy angol–magyar viszonylatban, jelentős minőségi különbség figyelhető meg. A pusztán statisztikai rendszer helyett tehát célszerű valamilyen szabályalapú elemet is tartalmazó, ún. hibrid rendszert alkalmazni a fordítás minőségének javítása érdekében.

3 Hibrid rendszerek

3.1 Szótár hozzáadása a korpuszhoz

A fordítórendszerek kiértékelésénél megfigyelhető, hogy a szóösszerendelő nehezen találja meg az összetartozó szövegrészeket, ha azok a nyelvtani szerkezet miatt messze vannak egymástól, vagy nagyon eltérnek. Az első ötlet a minőség javítására, hogy az eredeti korpuszhoz hozzáadunk egy kétnyelvű szótárat. Ettől azt az eredményt reméltük, hogy a kifejezések pontos fordítása nem csak segít az összerendelőnek megtalálni az összetartozó kifejezéseket a mondatban, de csökkenti a lefordítatlan szavak számát is.

Ehhez a feladathoz egy egyszerű angol–magyar szótárat használtuk [10], melyet először átalakítottunk, hogy egy kifejezésnek csak egyetlen megfelelője legyen. Így 344 924 darab kifejezéspárt kaptunk. Az elkészült szótárat többször is beleraktuk a korpuszba abból a célból, hogy a helyes előfordulások minél nagyobb súllyal forduljanak elő a fordítási modellben. Ezzel párhuzamosan viszont folyamatosan csökken az eredeti korpusz fontossága, csökken a többszavas kifejezések súlyozása a fordítási modellben, és romlik a nyelvi modell minősége. Ennek érdekében meg kell találni azt a számot, hogy hányszor éri meg a szótárat hozzáfűzni a korpuszhoz. E célból oly módon tanítottuk be az SMT rendszert, hogy az eredeti korpuszhoz egyszer, kétszer, háromszor, négyszer és ötször hozzáadtuk a kétnyelvű szótárat. Így a következő eredményeket kaptuk:

1. táblázat: Különböző rendszerek BLEU-eredményei.

Rendszer	**BLEU-érték**
Alaprendszer fordítása:	10,85%
Alap+1xszótár rendszer fordítása:	11,18%
Alap+2xszótár rendszer fordítása:	11,01%
Alap+3xszótár rendszer fordítása:	10,88%
Alap+4xszótár rendszer fordítása:	10,87%
Alap+5xszótár rendszer fordítása:	10,87%

Az 1. táblázatból látszik, hogy az alaprendszer (10,85% BLEU) értékéhez képest 0,33%-os javulás figyelhető meg, amely mértéke a behelyezett szótárak számától függően folyamatosan csökken. Ez a görbe azért tetőzik az első esetben, mert a szótár mérete összemérhető az eredeti korpusz méretével (fele az eredeti korpusznak), és emiatt annak ismétlése viszonylag hamar eltolja a súlyokat. A többszörös szótárhozzáadástól várt javuláshoz szükséges lenne egy nagyobb méretű párhuzamos mondatszintű korpusz is, de erőforrási problémák miatt nem tudtunk ilyet használni.

A teszthalmazból kiválasztott példamondat fordításait a 2. táblázat tartalmazza. Az első sorban az eredeti angol mondat olvasható, a másodikban ennek a referenciafordítása, majd az alaprendszer, végül a több szótárral kiegészített SMT fordítások eredményei.

Rögtön az első kifejezés elemzésénél feltűnik az *i wonder* fordításában való eltérés. Mind az alaprendszer, mind a legjobb eredményt nyújtó első rendszer *csak tudnám*-ra, míg a többi a *kíváncsi vagyok*-ra fordítja. Annak ellenére, hogy mind a két fordítás helyes, az automatikus kiértékelő mégis más eredményt ad a két fordításra, mivel a referenciafordításban a *kíváncsi vagyok* szerepel.

A következő érdekes kérdés a *teaching us* elemzése. A fordítás vizsgálatából kiderült, hogy az alaprendszer a *teaching*-et az *a tanítást*-ra fordította, ami a mondatbeli jelentéstől nem is áll messze. Ezzel szemben a szótárral kiegészített rendszerekben egységesen a *tanított nekünk* kifejezés érte el a legnagyobb valószínűséget, amely az *us* fordítását (*nekünk*) jobban tükrözi. Sőt, kissé elvont értelmezéssel az eredeti jelentéshez is közelebb áll, a szó szerinti fordítás viszont eltávolodott. A legnagyobb probléma itt is az, hogy nem egyezik meg a referenciával, ezért sem kap nagyobb BLEU-értéket.

2. táblázat: Különböző rendszerek eredményeinek összehasonlítása.

Angol referencia-fordítás:	" i wonder who 'll be teaching us ? " said hermione as they edged into the chattering crowd .
Magyar referenciafordítás:	- kíváncsi vagyok , ki tartja a tanfolyamot - morfondírozott hermione , miközben barátaival befurakodtak a tömegbe .
Alaprendszer fordítása:	- csak tudnám , ki lesz a tanítást ? - kérdezte hermione , mikor ő az .
Alap+1xszótár rendszer fordítása:	- csak tudnám , ki lesz tanított nekünk ? - szólt hermione , mikor elindult a jóvoltából .
Alap+2xszótár rendszer fordítása:	- kíváncsi vagyok , aki tanított nekünk ? - szólt hermione , mikor elindult a zsibongó tömeg .
Alap+3xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .
Alap+4xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .
Alap+5xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .

A *said* fordításánál hasonló jelenség figyelhető meg. Az alaprendszer *kérdezte*, míg a szótáras módszerek a *szólt* fordítást adták, amely annak tudható be, hogy a szótárban ez volt a megfeleltetése.

Nézzük a példa második részét. Látható, hogy az alaprendszer eredménye viszonylag gyenge (*mikor ő az* .). Ez amiatt van, hogy a szóösszerendelő a hosszabb mondatok második felét gyakran hozzákapcsolja valamelyik szóhoz, így torzul a fordítási modell. Ebből kifolyólag a dekódoló sem tud megbirkózni a hasonló szövegrészekkel. Így fordulhat elő, hogy a program „összecsapja" a fordítandó mondatok végét. Ezzel szemben a szótáras esetekben megfigyelhető változások bizonyítják a szóösszerendelő minőségének javulását. Az 1xszótár esetben már egy kerek mondatot kapunk, 2xszótár esetben megjelenik a *zsibongó tömeg*, 3xszótár után pedig a *mikor elindult az összeverődött tömegen* . kifejezés lett a rendszer szerinti legjobb fordítás.

A statisztikai gépi fordítórendszerek egyik nagy hiányosságát tükrözi, hogy az angolban az *into* prepozíció egy külön egységnek felel meg, de a fordító nem találja a helyes magyar fordítást. Mivel a magyar nyelv toldalékokat használ, ezért a főnévhez kapcsolódó különböző ragok más-más jelentéssel bíró szavakat hoznak létre, melyek közül a fordítómodul általában nem a helyes toldalékkal ellátottat választja ki. Ennek tudható be az a jelenség, hogy az *into* az első három esetben mintha nem is jelenne meg a fordításban (*tömeg*), a 3xszótáras rendszertől pedig megjelenik a *tömegen*, ami már ragozott alak, ám a program nem a helyes toldalékot találta meg.

Igaz, hogy a BLEU módszerrel való kiértékelés hatékonysága vitatható, de SMT rendszerek összehasonlítására alkalmas. Emiatt megvizsgáltuk a különböző rendszerek 1-9-gramos kifejezésekre vonatkozó BLEU %-ait (3. táblázat). Ebből megfigyel-

hető, hogy az alaprendszerhez képest (1) a szótárral kiegészített rendszerek 1-4-gram esetén mind jobb eredményt értek el. Ez jól mutatja, hogy a szótárban túlnyomórészt egy-két, de maximum négy-öt szóból álló kifejezések voltak, és emiatt ezek fordítása is egyre jobb lett. Látható, hogy a legjobb eredményt elérő 1xszótár (2) rendszer szinte az összes mérési esetben jobb lett, mint az alaprendszer, tehát ebben az esetben közelítettük meg legjobban a korpusz és a szótár méretének optimális arányát. E szint felett kezdenek az egy-két szavas kifejezések túl dominánssá válni, ami lerontja a magasabb n-gram értékeket. Ezért van az 5xszótár (4) esetben, hogy az 1-gram értéke sokkal magasabb még az 1xszótáras rendszerénél is, de már 2-gram esetén alacsonyabb lesz nála, míg 5-gram esetén már az alaprendszernél is.

3. táblázat: Különböző rendszerek n-gramonkénti eredményeinek összehasonlítása (BLEU-%-ban).

	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram
1	47,05	16,29	7,07	3,54	1,94	1,14	0,74	0,57	0,46
2	47,60	16,62	7,35	3,78	2,02	1,19	0,75	0,57	0,43
3	47,55	16,46	7,25	3,75	2,09	1,25	0,81	0,60	0,46
4	47,32	16,33	7,09	3,64	1,94	1,09	0,68	0,47	0,33
5	47,74	16,43	7,19	3,63	1,93	1,08	0,68	0,51	0,39

Sorai: 1. Alaprendszer; 2.Alap+1xszótár rendszer fordítása; 3. Alap+2xszótár rendszer fordítása; 4. Alap+3xszótár rendszer fordítása; 5. Alap+5xszótár rendszer fordítása

Az eredmények fontossága abban rejlik, hogy rámutatnak: a szóösszerendelés javításával lehet javítani a fordítórendszer minőségét.

3.2 Joshua

Következő lépésként a további hibridizáció lehetőségeit vizsgáltuk. Látható, hogy a távoli nyelvek fordítása esetén – amilyen például a magyar és az angol nyelvpár – az egytagú kifejezések közti statisztikákon felül szükségünk van más fogódzók kihasználására is. Ilyen tulajdonságok lehetnek a nyelvtani szabályok.

A fenti célok elérésére nyújthat lehetőséget a Joshua keretrendszer [6], mely nem pusztán szó- vagy frázisszintű statisztikai valószínűségi modelleket használ, hanem bizonyos nyelvtani jellemzők előfordulását is figyelembe veszi. A Joshua rendszer további nagy előnye, hogy képes ezen generatív szabályok közti fordításra oly módon, hogy megadhatóak a szabályok mind a forrásnyelvre, mind a célnyelvre, valamint az is megadható, hogy mekkora valószínűséggel transzformálhatók át a szabályok egymásba. Ennek köszönhetően alkalmasabb egymástól morfológiailag és szintaktikailag távoli nyelvpárok közötti fordításra.

A feladat során a következő egyszerű szabályrendszert használtuk, hogy meg tudjuk becsülni, hogy a módszer mennyire alkalmas az elvárt feladat megoldására:

```
[S] ||| [X,1] ||| [X,1] ||| 0 0 0                          (1)
[S] ||| [S,1] [X,2] ||| [S,1] [X,2] ||| 0.434294482 0 0    (2)
```

A Joshua rendszert a 2. rész 2-es bekezdésben leírt korpusszal tanítottuk be, hogy össze tudjuk hasonlítani a Moses rendszer eredményeivel. Az eredményt a 4. táblázat mutatja.

4. táblázat: A Joshua rendszer eredményének összehasonlítása.

Rendszer	**BLEU-érték**
Alaprendszer	10,85%
LitMag+Joshua+OOV	9,85%
LitMag+Joshua	11,06%

A Joshua rendszer alapértelmezetten minden szót, amelyet nem tudott lefordítani, megjelöl az OOV (Out Of Vocabulary) jellel. Látható, hogy így a fordítás minősége rosszabb, mint az alaprendszeré. Ez annak tudható be, hogy például a tulajdonneveket is megcímzi a fordításban, és hiába lett volna helyes a fordítás, ezzel mégis elrontja azt. Ennek elkerülése érdekében utólag leszedtük ezeket a címkéket, és megkaptuk a 11,06%-os értéket, amely a rendszer nagymértékű javulását mutatja.

Az 5. táblázat jól szemlélteti, hogy akár már egy egyszerű szabály bevezetésével is hogyan változik a fordítórendszer eredménye. A *„For a little while only”* szerkezetet az alaprendszer egyszerűen *„egy darabig csak”*-nak fordította, míg a Joshua rendszer a rekurziós szabály alkalmazásával megtalálta a helyes *„csak egy kis ideig”* fordítást. Ebből a példából még az is látszik, hogy az emberi kiértékelő számára mind a két fordítás elfogadható, de a gépi kiértékelés számára az első minimális, míg a második fordítás maximális értéket kapott.

5. táblázat: Példafordítás a Joshua rendszerrel.

Angol referenciafordítás:	" for a little while only , " said the voice quietly .
Magyar referenciafordítás:	- csak egy kis ideig - mondta a hang csendesen .
Alaprendszer fordítása:	- egy darabig csak - mondta a hang .
Joshua rendszer fordítása:	- csak egy kis ideig nyugodtan - mondta a hang .

Annak ellenére, hogy rendkívül ígéretesnek tűnik ez az új rendszer, eredményeinkben mégis kevés helyen tüntettük fel. Ennek az az oka, hogy nagyobb korpusz esetén túlságosan nagy lett a rendszer erőforrásigénye, amit a közeljövőben szeretnénk megszüntetni.

3.3 Cigány-magyar SMT rendszer

Statisztikai fordítórendszerünket kipróbáltuk egy, a magyarhoz morfológiai gazdagságában közelebb álló nyelv esetében is. Korpuszként Vesho-Farkas-féle lovári cigány nyelvű Újszövetségét [7], illetve a Káldi-Neovulgáta magyar fordítást használtuk [8].

6. táblázat: A lovári-magyar rendszerek eredményeinek összehasonlítása.

Rendszer	BLEU-érték
Lovári-magyar (Moses)	30,53%
Lovári-magyar (Joshua)	29,20%
Magyar-lovári (Moses)	30,38%
Magyar-lovári (Joshua)	35,88%

A 6. táblázatból olvashatók a fordítórendszerek által elért eredmények. Megfigyelhető, hogy a magyar–angol nyelvpárhoz képest sokkal jobb eredményt sikerült elérni, aminek számos oka lehet. A legszámottevőbb ok, hogy a teszt- és a tanítókorpusz is egyaránt ugyanabból a szövegből (az Újszövetségből) került ki. Ebből következik, hogy a létrehozott fordítógép túlságosan téma- és stílusspecifikus lett. Ismeretes, hogy a négy evangélium esetében több alkalommal is előfordul tartalom- és szövegismétlődés, amikor az evangélisták ugyanazt a történetet írják le, sokszor nagyon hasonlóan. Emiatt fordulhat elő, hogy a tesztfordítások között 100%-os fordítás is van, mert egyszerűen más helyen ezek a mondatok benne voltak a korpuszban.

Ennek ellenére az eredmények vizsgálatából megfigyelhető (7. táblázat), hogy a fordítórendszer sokkal jobb fordításokat generált, és az emberi kiértékelés számára sokkal olvashatóbb eredményt kaptunk, mint az angol–magyar esetben.

7. táblázat: Példamondat-fordítás a különböző rendszerekkel.

Cigány referencia-fordítás:	le but manusha pale tele sharadine penge gada po drom , kavera pale kranzhi phagrenas tele pa kasht haj po drom rispisarnaslen .
Magyar referencia-fordítás:	a hatalmas tömeg pedig leterítette ruháit az útra , mások meg ágakat vagdostak a fákról és az útra szórták .
Moses fordítása:	a nép pedig le terítették ruháikat az úton , mások pedig ágakat phagrenas le a fa , és az úton rispisarnaslen .
Joshua fordítása:	a nép pedig le terítették ruháikat az úton , mások pedig ágakat phagrenas le a fa és az úton rispisarnaslen .

4 Statisztikai szövegelemző

Munkánk során egy másik témakörrel is foglalkoztunk, ugyanis az SMT rendszerrel végzett kísérleteink során szükségünk volt a korpuszunk morfológiai elemzésére, és ez adta az ötletet a statisztikai módszerek egy újabb felhasználási területére. A szövegelemzés feladata is megfogalmazható két nyelv közti transzformációként, ha rendelkezésünkre áll a sima szöveget és annak elemzését tartalmazó párhuzamos korpusz is.

Számos módszer létezik szövegelemzésre, melyek közül a két leggyakrabban használt a szabályalapú és a gépi tanuláson alapuló módszerek. Mind a két módszernek számos előnye van, de a szabályalapú rendszer számára – a gépi fordításhoz hasonlóan – rendkívül nehéz és körülményes megfogalmazni a megfelelő szabályokat. A gépi tanulásos megoldásnak is megvannak a nehézségei: igaz, hogy egyes

szabályokra nagyon pontosan betanítható, de ha az összes szabályra szeretnénk alkalmazni, túl komplex és lassú rendszert kapunk.

Ezzel szemben a statisztikai módszer a betanítás során az összes általa felismerhető szabályt figyelembe veszi. Ehhez csak egy megfelelő és elég nagy méretű korpuszra van szükség, cserébe egy online rendszert kapunk, ami viszonylag gyorsan képes ezután elemzést végezni. Ezen felbuzdulva megvizsgáltuk az SMT rendszer alkalmazhatóságát szövegelemzésre.

Ennek a rendszernek a felépítéséhez a Szeged Korpusz 2.0-t használtuk [1], melynek talán egyetlen hibája, hogy viszonylag kis méretű. Ennek ellenére alkalmasnak tűnt a felhasználásra. Mivel a szófaji címkék korlátozott számúak, elvben kisebb méretű korpuszban is elég nagy gyakorisággal szerepelhetnek. A rendszert kiértékeltük a BLEU módszerrel (8. táblázat), és kiszámítottuk a pontosságát is (9. táblázat).

8. táblázat: A szövegelemző rendszer automatikus kiértékelése I.

Rendszer	**BLEU-érték**
Szeged+Moses	90,97%
Szeged+Joshua	90,96%

9. táblázat: A szövegelemző rendszer pontossága I.

Rendszer	**Pontosság**
Szeged+Moses+helyes	90,29%
Szeged+Moses+helytelen	9,71%

A kiértékelésénél szembetűnt a rendszer néhány hibája. Az első és talán legfontosabb probléma a korpusz szerkezetéből fakad. Az elemzett korpuszban egymás után szerepelnek a szavak szótövei, amikhez hozzákapcsolódnak az elemzést tartalmazó címkék, de a több tagból álló kifejezések esetekben (pl. többtagú tulajdonnevek, igei szerkezetek) a címke csak a kifejezés utolsó szaván vagy utána helyezkedik el. Az egy szófaji egységbe tartozó kifejezések jelölésének hiánya a statisztikai módszerben félrevezető fordítási modellt eredményez. Ennek köszönhetően az első rendszer a tulajdonnévi címkéhez hozzácsatolt egy *„)_[PUNCT]”* szöveget, így gyengébb eredményt kaptunk. Az eredmény javítása érdekében minden önálló címkét hozzácsatoltunk az előtte álló szóhoz, így kaptuk a 10. táblázatban látható eredményt.

10. táblázat: A szövegelemző rendszer eredménye II.

Rendszer	**BLEU-érték**	**Helyes**	**Helytelen**
Moses	90,97%	90,80%	9,20%
Joshua	90,96%	90,72%	9,28%

A 10. táblázatból látszik, hogy változatlan BLEU-értékek mellett a rendszer pontossága 0,5–0,6 százalékkal javult. Ezt annak köszönhetjük, hogy nem kerültek a fordításba felesleges elemek, de a többtagú kifejezések fordítása továbbra sem megoldott. A probléma megoldásához elengedhetetlen ezeknek a kifejezéseknek az ösz-

szekapcsolása például a tulajdonnevek felismerésével. Nem volt célunk ilyen rendszer kifejlesztése, viszont az elmélet igazolása érdekében összekötöttük a korpuszban ezeket a kifejezéseket. A tanítás után a 11. táblázatban látható eredményt kaptuk.

11. táblázat: A szövegelemző rendszer eredménye III.

Rendszer	BLEU-érték	Helyes	Helytelen
Moses	90,96%	91,05%	8,95%
Joshua	90,77%	91,07%	8,93%

A 11. táblázatból megfigyelhető, hogy az összetartozó szavak összekapcsolása tovább javította rendszerünk pontosságát, annak ellenére, hogy BLEU-értéke alacsonyabb lett, mint az elődeinek. A hibás fordítások vizsgálatából kiderült, hogy a hibák két fő csoportba sorolhatók. Az első, amikor a rendszer nem ad fordítást, hanem visszaadja az eredeti szót. Az esetek túlnyomó részében ezek a szavak nem szerepelnek a korpuszban, így a fordítási modellben sincs megfeleltetésük. A másik hibafajta a helytelen elemzések halmaza. Itt is két fő hibakategória különíthető el. Az egyik esetben a szófajt helyesen beazonosítja, csak annak a további elemzésében ront; a súlyosabb hiba pedig, mikor már a szófajt sem találja el.

Felmerül a kérdés, hogy egyértelműsítést tartalmaz-e a rendszer és alkalmas-e rá. A választ maga a statisztikai módszer tulajdonsága adja. Amelyik jelenségre van elegendő példa, arra az SMT rendszer nagyon jól fog működni. Ha minden többértelmű kifejezés elég sokszor szerepel a tanítóhalmazban, a rendszer helyesen fogja megítélni a többértelmű szavakat is. Sajnos ilyen korpusz egyelőre nem állt rendelkezésünkre, és valószínűleg a közeljövőben nem is fog. Mivel a felhasznált Szeged Korpusz viszonylag kis méretű, ezért ez a rendszer nagy valószínűséggel nem oldja meg jól az egyértelműsítést, tehát ha a kérdéses szövegkörnyezet nincs benne a korpuszban, általában a legvalószínűbb elemzést rendeli hozzá a szóhoz.

A 12. táblázatban egy példán keresztül bemutatott rendszer javításának további lehetőségeit is vizsgáltuk. Az előzőekben leírtakból következik, hogy megfelelő méretű korpusz segítségével bármilyen szabály jól betanítható. Mivel jelen esetben korpuszunk fix méretű, minőségi javulás akkor érhető el, ha csökkentjük a komplexitást. A mi esetünkben ez úgy érhető el, ha a sima szöveget csak a szófaji címkék „nyelvére” fordítjuk, tehát nem írjuk ki eléjük maguknak a szavaknak a szótöveit. Mivel ezekből a címkékből sokkal kevesebb van, mint a magyar szavakból, kisebb korpuszból is felépíthető egy relatíve pontos rendszer. Másfelől: ha elhagyjuk a szótöveket az elemzésből, és csak a címkékre fordítunk, sokkal nagyobb súllyal szerepel majd a szófajok mondatban elfoglalt sorrendje mind a nyelvi, mind a fordítási modell esetében.

A rendszer eredményei a 13. táblázatban láthatóak. Ismét megfigyelhető a BLEU-érték csökkenése az eredeti rendszerhez képest, ám a pontosság az eddigi legjobb lett (bár ezt szótövesítés nélkül oldottuk meg). A módszer azonban használható, mert ha egy másik rendszert a sima szöveg szótöves változatára tanítunk be, akkor hasonló jó hatásfokkal tudjuk fordítani, és a kapott eredményeket összekapcsolva az elemző minősége is javítható lehet.

12. táblázat: Példa a szövegelemző működésére.

Sima szöveg:	mindenképp kötelességtudó szeretnék lenni , de azért nem olyan fanatikus szinten , mint egyes felnőttek , hogy még a családot is feláldozza a kötelesség miatt .
Referencia elemzés:	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] ._[Punct]
SMT elemző:	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] ._[Punct]

13. táblázat: A szövegelemző rendszer eredménye IV.

Rendszer	**BLEU-érték**	**Helyes**	**Helytelen**
Moses	88,65%	91,22%	8,77%
Joshua	88,57%	91,09%	8,91%

5 Összefoglalás

Kutatásunkban statisztikai módszerek és kétnyelvű párhuzamos korpusz segítségével igyekeztünk megoldani olyan feladatokat, ahol a cél megfogalmazható volt két nyelv közti transzformációként. Sikerült javítani az angol–magyar statisztikai gépi fordítórendszer minőségét részben kétnyelvű szótár hozzáadásával, részben a rendszer hibridizációjával. Sikeresen betanítottunk egy lovári–magyar statisztikai gépi fordító rendszert, amely eredményei segíthetnek az angol–magyar fordító rendszerek minőségét javítani. Eredményeink rávilágítanak, hogy bizonyos mértékű hibridizáció segítségével az SMT rendszerek minősége javítható. A kutatás folyamán létrehoztunk egy statisztikai szövegelemző rendszert is, és megvizsgáltuk annak minőségét. A kapott eredmények bíztatóak voltak, és rámutattak, hogy ebben a kutatási irányban is rejlenek további lehetőségek. Az eredmények figyelmeztettek arra is, hogy önmagukban a statisztikai módszerek nem elégségesek ezen feladat megoldásra: mindenképp szükséges valamilyen hibridizáció.

Hivatkozások

1. Csendes D., Hatvani Cs., Alexin Z., Csirik J., Gyimóthy T., Prószéky G., Váradi T. : Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In: I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Egyetem (2003) 238–247
2. EuroMatrix-táblázat. http://www.statmt.org/matrix/ (2007)

3. 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)
4. Koehn P.: Moses – A Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models. User Manual and Code Guide (2009)
5. 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. Association for Computational Linguistics, Prague (2007) 177–180
6. Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W.N.G., Weese, J., Zaidan, O. F.: JosHUa: An Open Source Toolkit for Parsing Based Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. Athens, Greece (2009) 135–139
7. Suntoiskirpe Nyevo Teshtamento (ford.: Vesho-Farkas Zoltán). Szent Jeromos Bibliatársulat, Budapest (2003)
8. Újszövetségi Szentírás a Neovulgáta alapján. Szent Jeromos Bibliatársulat, Budapest (1997)
9. Varga D., Halácsy P., Kornai A., Nagy V., Németh L., Trón V.: Parallel Corpora for Medium Density Languages. Recent Advances in Natural Language Processing Conference (2005) 590–596
10. Vonyó A.: A mindenki által keresett ingyenes angol–magyar magyar–angol köznapi, műszaki és szlengszótár. http://almos.vein.hu/~vonyoa/SZOTAR.HTM (1999)

Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban

Héja Enikő[1], Sass Bálint[1]

[1] MTA Nyelvtudományi Intézet
{eheja,sass.balint}@nytud.hu

Kivonat: Jelen cikk célja annak vizsgálata, hogy a párhuzamos korpuszokból fordítási ekvivalensek kinyerésére használt módszer kiterjeszthető-e többszavas kifejezésekre is. A kísérletben a többszavas kifejezéseket kizárólag igei szerkezetek alkották. Első lépésként a párhuzamos korpusz forrásnyelvi és célnyelvi oldalából külön-külön nyertük ki az igei szerkezeteket. A jelenlegi fázisban a kinyerés félig automatikus módon történt: előre meghatározott forrásnyelvi igékhez és ezek célnyelvi fordításaihoz tartozó igei szerkezeteket kerestünk, melyekből kézzel válogattuk ki a céljainknak megfelelőeket. A következő lépésben ezeket egytagú kifejezésekké vontuk össze a párhuzamos korpuszban. Az összevont igei szerkezetek már az illesztési algoritmus bemeneteként szolgálhattak. Eredményeink azt mutatják, hogy az alkalmazott módszer jól használható igei szerkezetek fordítási ekvivalenseinek detekciójára.

1 Bevezetés

Jelen cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX[1] projekt része. A projekt azt vizsgálja, hogy a nyelvtechnológiai módszerek és eszközök – különös tekintettel a párhuzamos korpuszokra – mennyiben járulhatnak hozzá a szótárkészítési folyamathoz. A szótárkészítés automatikus támogatása elsősorban a kevéssé használt nyelvek esetében bír jelentőséggel, hiszen az ilyen nyelvpárokra készült szótárak iránti kereslet alacsony, így a szükséges munkálatok finanszírozása is korlátozott. A projekt célkitűzése középméretű (kb. 15,000 szócikk), általános célú szótárak létrehozása volt a magyar-litván, illetve a francia-holland nyelvpárokra.

Jelenleg nem létezik olyan módszer, amely lehetővé tenné a szótárak *teljesen* automatikus előállítását. Ezért egy megfelelő lefedettségű és pontosságú lexikai erőforrás előállítása mindenképpen igényel emberi utószerkesztési munkálatokat is. Ennek fényében úgy fogalmazhatjuk meg feladatunkat, hogy célja a lexikográfusok számára olyan erőforrásokat biztosítani, amelyek a lehető legjobban csökkentik a teljes értékű, emberi felhasználásra alkalmas szótárak elkészítéséhez szükséges munkát. A fenti elvárásoknak megfelelő automatikusan generált erőforrásokat protoszótáraknak fogjuk nevezni a cikk hátralevő részében. Jelen írás az alábbi szerkezetet követi: a bevezetés után röviden ismertetjük az egytagú fordítási ekvivalensek kinyerésére használt

[1] http://www.efnil.org/projects/efnilex

módszert, illetve ennek előnyeit és hátrányait (2). Ezt követően vázoljuk a munkafolyamatot (3), amely két fő lépésre bontható: az igei szerkezetek kinyerésére (3.1), valamint a protoszótár létrehozására (3.2). Majd eredményeinket mutatjuk be (4), végül pedig a konklúziókat és a további teendőket (5).

2 Az alkalmazott módszer – előnyök és hátrányok

A statisztikai gépi fordítás térhódításával jelentősen megnőtt a párhuzamos korpuszok szerepe a nyelvtechnológiában. Már legalább 16 éve használnak különféle statisztikai algoritmusokat forrásnyelvi és célnyelvi szópárok kinyerésére, hogy így bővítsék a gépi fordítás bemenetéül szolgáló szótárakat (pl. [8]).

Érdekes módon a lexikográfusok között a mai napig sem eldöntött kérdés, hogy érdemes-e párhuzamos korpuszokra támaszkodni az emberi felhasználásra készülő szótárak készítése során. Például [1] szerint „*Túl magas az az ár, amit akkor kéne fizetni, ha a szótárszerkesztők kétnyelvű korpuszokat használnának. [...] Az ellenérvek: túl sok fordítási ekvivalens, a szerkesztés során mindegyik fordítási ekvivalens egyformán fontosnak tűnik a lexikográfus számára, [...], a bejegyzések a legtöbb felhasználó számára túl sok részletet tartalmaznak.*"

Eddigi kísérleteink [4] azt mutatták, hogy ha előfeldolgozásként szóillesztést végzünk, akkor a szóillesztés során kapott szópárok és ezek fordítási valószínűségei már megfelelő kiindulópontként szolgálhatnak egy kétnyelvű szótárhoz. Így tehát a fenti ellenvetések nem állják meg a helyüket. Először is, a forrásnyelvi szó és célnyelvi ekvivalense közötti fordítási valószínűség, valamint a forrásnyelvi szó és célnyelvi megfelelőjének gyakoriságai alapján szűrhetjük a fordítási jelölteket, így csökkentve a lehetséges fordítási ekvivalensek számát. Továbbá, a fordítási valószínűségek alapján sorrendezhetjük a fordítási jelölteket, ami biztosítja, hogy a leggyakrabban használt fordítás szerepeljen a szótári bejegyzés elején. Az általunk javasolt módszer további előnyei közé tartozik, hogy egy megfelelő méretű párhuzamos korpusz garantálja, hogy a legfontosabb fordítási ekvivalensek szerepelni fognak a szótárban. Ezenfelül, a párhuzamos korpusz gazdag tárháza a valódi nyelvből vett példamondatoknak, amelyek alapján a lexikográfus vagy a szótárhasználó kiválaszthatja a számára legmegfelelőbb fordítást a lehetséges fordítások közül. A fenti jellemzők miatt a javasolt módszer különösen alkalmas aktív[2] szótárak készítésének támogatására.

A [4]-ben javasolt módszer hátránya, hogy nem kezeli a többszavas kifejezéseket, így jelen állapotában alkalmatlan a több szóból álló fordítási ekvivalensek kiszűrésére. Ennek a feladatnak a megoldását mind a forrásnyelvi, mind a célnyelvi oldalon kiemelten fontosnak tartjuk, hiszen a többszavas kifejezések és szokásos fordításainak kinyerésével biztosíthatjuk, hogy a gyakori szófordulatok szerepeljenek a szótárban. Ez teszi lehetővé, hogy a szótár alapján természetesen hangzó célnyelvi szöveget hozhassunk létre. Jelen cikkben a többszavas kifejezések kezelését célzó első kísérletet ismertetjük. A cikkben ismertetett kísérlet célja *ige + bővítmény* szerkezetek fordítási ekvivalenseinek automatikus felismerése a francia-holland nyelvpárra. Reménye-

[2] Az aktív szótárak célja, hogy a forrásnyelvi anyanyelvi beszélőt segítsék a célnyelvi megnyilatkozások létrehozásában.

ink szerint az általunk javasolt módszer az ige+bővítmény szerkezetek fordítási ekvivalenseinek automatikus meghatározásával elősegíti a szótári tételek mikrostruktútájának kialakítását.

3 Munkafolyamat

A munkafolyamat két fő szakaszból áll. Az első lépésben a francia és holland igei szerkezetek félig automatikus kinyerésével hozzuk létre a vizsgálandó igei szerkezetek listáját (3.1). A második lépésben a kiválasztott többszavas igei szerkezeteket összevonjuk, így ezek az illesztés bemenetül szolgálhatnak. Eredményül egy többszavas igei szerkezeteket tartalmazó protoszótárat kapunk (3.2).

3.1 Francia és holland igei szerkezetek félig automatikus kinyerése

A kísérlethez a TLT-Centrale által fejlesztett Holland Párhuzamos Korpusz (DPC – Dutch Parallel Corpus) francia-holland alkorpuszát használtuk [5]. Az összesen 6,820,547 tokenes párhuzamos korpusz 186,945 illesztett egységet tartalmaz[3].

Első lépésben kézzel kiválasztottunk 20 gyakori, általunk poliszémnek gondolt francia igét (pl.: *mettre* 'tesz, helyez'). Ezek mindegyikéhez hozzárendeltünk egy-egy alapértelmezettnek tűnő holland fordítást (a *mettre* esetében ez a *leggen*). Az 1. táblázatban tételesen felsoroljuk a kiválasztott francia igéket, ezek holland megfelelőit, illetve magyar fordításukat. Megadtuk továbbá, hogy a következő lépés eredményeként hányféle különböző igei szerkezetet találtunk az adott igére.

1. táblázat: Francia igék és holland fordításaik.

Francia ige	Különböző igei szerkezetek	Holland fordítás	Különböző igei szerkezetek	Magyar fordítás
donner	12	*geven*	31	adni
effectuer	3	*teweegbrengen*	0	előidéz, véghezvisz
enlever	0	*verwijderen*	1	eltávolít
faire	31	*doen*	12	csinálni
mener	2	*leiden*	4	vezet
mettre	26	*leggen*	5	tenni

[3] Mivel a párhuzamos korpusz tartalmaz egy-a-többhöz, illetve több-az-egyhez megfeleltetéseket, a mondatok száma helyett a korpusz méretét az illesztett egységek számával adjuk meg.

montrer	4	*wijzen*	1	mutatni
obtenir	5	*behalen*	2	megszerezni
offrir	1	*aanbieden*	2	kínálni
ouvrir	1	*openen*	1	nyitni
passer	3	*vergaan*	0	eltölteni
porter	3	*brengen*	14	hozni
prendre	23	*nemen*	23	(el)venni
recevoir	2	*krijgen*	12	kapni
rendre	3	*maken*	19	tesz (vmilyenné)
rester	0	*blijven*	1	maradni
tenir	4	*houden*	11	tartani
traiter	1	*behandelen*	2	bánni
trouver	3	*vinden*	6	találni
voir	0	*zien*	0	látni

Amint azt a táblázat is mutatja, a kiválasztott igék némelyikéhez (*unlever*, *rester*, *voir, vergaan, zien*) egyetlen feltételeinknek megfelelő igei szerkezetet sem találtunk.

A következő lépésben automatikusan kinyertük a releváns francia, illetve holland ige+bővítmény szerkezeteket a párhuzamos korpusz megfelelő egynyelvű részéből. A [7]-ben leírt igei szerkezetek kinyerésére szolgáló módszert alkalmaztuk. Ez a módszer tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszon dolgozik. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzés pedig meg kell hogy állapítsa a tagmondat igéjét, a bővítmények fejét, valamint a bővítmények igéhez való szintaktikai viszonyát. A szintaktikai viszonyt a megfelelő esetrag vagy egy elöljárószó jelöli.

A módszer a tagmondatokban az ige mellett meglévő jellegzetes bővítménykereteket határozza meg, a gyakori részkeretek rendszerezett összeszámlálása révén. Előnye abban rejlik, hogy automatikusan felismeri, hogy melyik bővítménynél lényegi elem a konkrét fej és melyiknél csak az ige-bővítmény viszony, azaz egyszerre képes meghatározni az összetett igéket és a vonzatkereteket is. A *hasznot húz vmiből* szerkezet esetén például felfedezi, hogy a lexikálisan kötött tárgy mellett egy *-ból*/*-ből* esetragos vonzat szerepel az igei keretben.

Az algoritmus vázlata a következő. Vesszük a korpusz összes tagmondatát. Előállítjuk a tagmondatoknak megfelelő szerkezeteket, melyekben a bővítményi fejeket minden variációban, váltakozva töröljük, illetve megtartjuk. Hossz szerint csökkenő sorba rendezzük a kapott szerkezetlistát, majd sorra elhagyjuk azokat a szerkezeteket, melyeknek a gyakorisága 5-nél kisebb, és ezek gyakoriságát a megfelelő, illeszkedő rövidebb keret gyakoriságához adjuk. A megmaradó szerkezetek gyakoriság szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

Említettük, hogy az algoritmus bemenetként tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszt vár. Jelen kísérletben mindkét elemző lépést egyszerű szabályok használatával közelítettük a francia és a holland nyelv esetében is. Tagmondathatárt jelent nyilvánvalóan a mondathatár. Ezen kívül a kötőszó, az alárendelt tagmondatot bevezető holland *te*, ill. francia *pour*, a vonatkozó névmás és bizonyos írásjelek (vessző, kettőspont és pontosvessző) is, amennyiben a legutóbbi tagmondathatár óta szerepelt a mondatban ige. Bővítményi fejek a főnevek (valamint a reflexív igék miatt a holland *zich* és a francia *se*); a bővítményi viszonyt pedig a szó előtt álló prepozíció jelzi. Prepozíció híján az ige előtt alanynak, az ige után pedig tárgynak vettük az adott bővítményt.

Eredményként olyan összetett igei szerkezeteket kapunk, mint a francia *mettre accent sur*... vagy holland megfelelője, *a leggen nadruk op*... (magyarul mindkettő: 'hangsúlyt helyez vmire').

Az így kinyert kifejezésekből kézzel válogattuk ki azokat az igei szerkezeteket, amelyekről azt gondoltuk, hogy nem, vagy csak részben kompozicionálisak. Mivel fordítási feladatról van szó, a kompozicionalitás ebben az esetben nem önmagában, hanem egy másik nyelv függvényében értelmezhető. Így a nem transzparens kifejezések mellett intézményesült kifejezések kinyerésére is törekedtünk. Például bár a *mettre l'appareil hors tension* (szó szerint 'feszültségen kívül helyezni a készüléket' – 'áramtalanítani') francia kifejezés kompozicionális szerkezetnek tekinthető, bekerült a listánkba, mivel a holland fordítás – *uitschakelen* – már nem őrzi meg a kifejezés eredeti szerkezetét. Így tehát az automatikusan kinyert igei szerkezetek közül azokat vettük fel a listánkba, amely az alábbi kritériumok közül bármelyiknek megfelelt:

(1) A kifejezés jelentése az eredeti nyelven nem transzparens (pl. *faire mouche* 'célt érni')
(2) Ha feltételezhető, hogy az igei szerkezet fordítása nem tükörfordítás
    a. Az igei szerkezet intézményesült
    b. Az igei szerkezet magyar fordításában a főnév igemódosítóként[4] jelenik meg (l. [3]).

Fontos hangsúlyozni, hogy az igei minták közül a kollokációszerű mintákat választottuk ki, azaz azokat, melyek az ige mellett tartalmaztak egy konkrét főnevet is. A konkrét főnév az alanyeset kivételével bármilyen esetben állhat. Az igei szerkezetek kiválasztásakor nem törekedtünk a teljes igei vonzatkeret megőrzésére, így bizonyos esetekben a kitöltetlen – vagyis tipikus főnévi lemma nélkül álló – esetragokat el-

[4] Igemódosítónak azokat a névelőtlen névszókat tekintettük, amelyek az igekötővel nagyjából hasonló disztribúciót mutatnak és semleges, tagadást nem tartalmazó mondatban közvetlenül az ige előtt állnak.

hagytuk. Ennek oka egyfelől, hogy az igei szerkezetek összevonásával csökkenthetjük az adathiány problémáját, másfelől pedig az, hogy mivel az illesztés bemeneti korpusza nem tartalmazott sem részleges szintaktikai elemzést, sem tagmondatfelismerést, az esetek egy jelentős részében lehetetlen volt pontosan azonosítani a megfelelő prepozíciót.

Az illesztés bemenetéül az alábbi szintaktikai mintákra illeszkedő igei szerkezetek szolgáltak (V: ige; N_ACC: a főnévi lemma szintaktikai funkciója tárgy; ACC: kitöltetlen tárgy; N_PREP: a főnévi lemma valamilyen prepozícióval szerepel; PREP: kitöltetlen prepozíció):

(1) V + N_ACC
(2) V + N_PREP
(3) V + ACC + N_PREP
(4) V + N_ACC + PREP
(5) V + N_PREP + PREP
(6) V + N_PREP + N_PREP + PREP

A harmadik lépésben következik ezen igei szerkezetek korpuszbeli azonosítása, összevonása és illesztése.

3.2 A protoszótár létrehozása

A továbbiakban a kiválogatott többszavas igei kifejezéseket egy szóként kezeltük, és így közvetlenül alkalmaztuk az eredeti szavakon működő illesztő algoritmust.

Mint említettük, az illesztés bemeneti korpusza lemmatizált, de nem tartalmaz sem tagmondat-információt, sem részleges szintaktikai elemzést. Ezen munkaszakasz első lépése a minták felismerése a korpuszban, majd ezek összevonása egytagú kifejezéssé. A 126 francia igei szerkezet összesen 7805-ször, míg a 146 holland igei szerkezet 8029-szer fordult elő a párhuzamos korpuszban.

A szóillesztést a GIZA++ szoftverrel végeztük [6], amely a szóillesztés során fordításjelölteket hoz létre, úgy, hogy a forrásnyelvi és célnyelvi lemmapárokhoz fordítási valószínűséget rendel. A fordítási valószínűség a célnyelvi és forrásnyelvi szópár feltételes valószínűségének közelítése – $P(\text{szó}_{\text{cél}}|\text{szó}_{\text{forrás}})$ – az EM (expectation maximization) algoritmus alapján [2].

A protoszótárak kiindulási alapját az így kinyert fordítási jelöltek és fordítási valószínűségeik képezték. Mivel a fordítási valószínűség 0-tól 1-ig bármilyen értéket felvehet, ebben a szakaszban még sok helytelen fordítási jelöltünk van. Ezért szükség van olyan szűrők bevezetésére, amelyek lehetővé teszik a legjobb fordításjelöltek automatikus kiválasztását a lehető legtöbb helyes fordításjelölt megtartásával. Eddigi tapasztalataink azt mutatták [4], hogy a fordítási valószínűségek és a forrásnyelvi, illetve célnyelvi korpuszgyakorisági adatok együttesen már jól használhatóak az eredmények szűrésére. Így a protoszótárban az alábbi adatok szerepelnek:

2. táblázat: A protoszótárban szereplő adatok.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
mettre_à_jour 'frissít'	*actualiseren* 'frissít'	0.0472766	105	39

A már elvégzett kiértékelések alapján (magyar-litván, magyar-szlovén) az alábbi általános feltételeket fogalmazhatjuk meg a protoszótárban szereplő tételekkel szemben:

(1) A forrásnyelvi és a célnyelvi szónak is legalább 5-ször elő kell fordulnia a párhuzamos korpuszban. Ez a feltétel szükséges ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez.

(2) Mivel a fordítási valószínűség a célnyelvi szó feltételes valószínűsége a forrásnyelvi szó mellett, további követelmény, hogy a két lemma (vagy kifejezés) gyakorisága ne legyen túl különböző. Gyakori célnyelvi lemmák esetében az illesztési algoritmus magas fordítási valószínűséget rendelhet rossz fordítási jelöltekhez is, ha a forrásnyelvi lemma ritkán fordul elő a korpuszban és a lemmák gyakran fordulnak elő illesztett mondatokban. A 3. táblázatban egy ilyen fordítási jelöltpárt mutatunk be.

3. táblázat: Rossz fordítási jelöltpár.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
mettre_vie_en_danger 'veszélyezteti az életét'	*rekening_houden* 'figyelembe vesz'	0.876763	24	577

Mivel a munka célja emberi felhasználásra szánt szótárak automatikus előállítása, a kiértékelés során a rossz és a jó fordítási jelöltpárok helyett *lexikográfiailag hasznos* és *lexikográfiailag nem hasznos* fordítási jelölteket különböztettünk meg. Az egyszavas kifejezések kiértékelése azt mutatta, hogy a vizsgált paraméterek mellett (a forrásnyelvi szó és a célnyelvi szó gyakorisága legalább 5 és a fordítási valószínűség legalább 0.5) a találatok mintegy 90%-a lexikográfiailag hasznos fordítás. Tekintve, hogy a pontosság fordítottan korrelál a lefedettséggel, és utólagos kézi feldolgozásra mindenképpen szükség van, a paraméterek olyan beállítására lesz szükség, amely kb. 60%-os pontosságot eredményez. Egy további megfigyelés szerint a gyakoribb lemmák esetén már a jóval kisebb fordítási valószínűséggel rendelkező párok jelentős része is a lexikográfiailag hasznos kategóriába tartozik, így ezek figyelembevétele tovább növeli a lefedettséget. Mivel a jelen munkaszakasz feladata nem a megfelelő szűrési paraméterek beállítása, hanem annak vizsgálata, hogy az eredeti módszer kiterjeszthető-e többszavas kifejezésekre is, a paramétereket úgy állítottuk be, hogy a lehető legtöbb jó fordítási jelöltet megtartsuk, még akkor is, ha ez alacsony pontossághoz vezet. Ezért minden olyan fordítási jelöltpárt megtartottunk, ahol a pár mindkét tagjának gyakorisága legalább 5, a fordítási valószínűséget azonban 0.02-re csökkentettük. A következő szakaszban eredményeinket mutatjuk be.

4 Eredmények

Az eredmények részleges kiértékelése azt mutatta, hogy a használt módszer alkalmas igei szerkezetek fordításainak kinyerésére.

A fenti paraméterekkel összesen 906 olyan fordítási jelöltet kaptunk, amelynek legalább egyik tagja többszavas kifejezés, ebből 632 esetben a forrásnyelvi kifejezés többszavas. Ez 113 különböző francia igei szerkezetet jelent (a fordításjelöltek között 127 különböző holland igei szerkezet szerepel). A részleges kiértékelés során a francia többszavas kifejezésekre koncentráltunk.

294 fordítási jelöltpárt értékeltünk ki kézzel. A kiértékelés során a *teljesen jó* és a *rossz fordítási* párok mellett megkülönböztettünk *részlegesen jó* fordítási párokat is. A részlegesen jó fordítási párok esetében a jelöltpár valamelyik tagja egy fel nem ismert többszavas kifejezés, így csak részleges illesztés történt. A kiértékelt párok közül 57 fordítás volt tökéletes (19%) és 28 fordítási jelölt bizonyult részlegesen jó fordításnak. A választott paraméterek mellett 189 fordítási jelölt volt teljesen rossz. Mivel ebből 132 csak egy mondatpárban fordult elő a teljes párhuzamos korpuszban, a jövőben egy további paraméterként azon mondatpárok számát is felvesszük, amelyben a fordítási jelöltek előfordulnak. Ezen szűrő hasznosságát támasztja alá, hogy a 149 fordítási jelölt közül, amelyek csak egy mondatpárban fordultak elő, csak 17 volt lexikográfiailag hasznos fordítás.

Az eredmények kézi ellenőrzése során azt találtuk, hogy a rossz fordítási jelöltek száma jelentősen csökkenthető lenne, ha a részleges szintaktikai elemzésre és a tagmondathatárra vonatkozó információkat az illesztés során is figyelembe vennénk.

Fontos hangsúlyozni, hogy jelen cikk célja nem a lehető legnagyobb pontosság elérése, hanem a módszer alkalmazhatóságának vizsgálata igei szerkezetek kinyerésére. A paraméterek átállításával a pontosság jelentősen növelhető. A 4. és 5. táblázatban két példával illusztráljuk, hogy a megközelítés hogyan használható fordítások kinyerésére.

4. táblázat: Első példa.

$\text{Kifejezés}_{\text{forrás}}$	$\text{Kifejezés}_{\text{cél}}$	$P(\text{szó}_{\text{cél}}\|\text{szó}_{\text{forrás}})$	$\text{Gyak}_{\text{forrás}}$	$\text{Gyak}_{\text{cél}}$
mettre_à_jour	*bijwerken*	0.0649992	105	60
FR: Comment les **met**-on **à jour**?				
NL: Hoe worden ze **bijgewerkt**?				
H: Hogyan lehet ezeket **frissíteni**?				
mettre_à_jour	*actualiseren*	0.0472766	105	39
FR: De plus, un PGR **mis à jour** doit être soumis:				
NL: Bovendien dient een **geactualiseerd** RMP ingediend te worden :				

H: Ezenfelül, egy **frisített** PGR-t kell elküldeni:				
mettre_à_jour	*aanpassing*	0.0372093	105	442
FR: **Mise à jour** de la liste des produits admis au remboursement				
NL: **Aanpassing** van de lijst van de voor vergoeding aangenomen producten				
H: A költség-visszatérítésre elfogadott termékek listájának **kiigazítása**				
mettre_à_jour	*update*	0.029671	105	34
FR: Toutes les informations au sujet du changement y ont été publiés avec de fréquentes **mises à jour** .				
NL: Alle informatie met betrekking tot de omslag is erop gepubliceerd , met regelmatige **updates**.				
H: Minden változásra vonatkozó információ ott van közzétéve, rendszeres **frissítésekkel**.				

5. táblázat: Második példa.

$Kifejezés_{forrás}$	$Kifejezés_{cél}$	$P(szó_{cél}\|szó_{forrás})$	$Gyak_{forrás}$	$Gyak_{cél}$
prendre_en_considération	*nemen_in_aanmerking*	0.186464	93	73
FR: Les offres qui dérogent à cette date ne sont pas **prises en considération**.				
NL: Offertes die hiervan afwijken worden niet **in aanmerking genomen**.				
H: A megadott dátumoktól eltérő ajánlatokat nem **vesszük figyelembe**.				
prendre_en_considération	*houden_rekening*	0.166621	93	438
FR: Concernant votre apport financier personnel , vos revenus sont **pris en considération**.				
NL: Voor uw eventuele persoonlijke financiële bijdrage wordt **rekening gehouden met** uw inkomsten.				
H: Az Ön személyes anyagi hozzájárulásánál **figyelembe vesszük** a jövedelmét.				
prendre_en_considération	*nemen_in_overweging*	0.0221903	93	35
FR: La date de conclusion à **prendre en considération** pour le choix …				
NL: De datum van sluiting die **in overweging** moet worden **genomen** voor de keuze …				
H: A zárás időpontja, amelyet a választáshoz **figyelembe** kell **venni** ….				

5 Konklúziók és további teendők

Jelen cikk célja annak vizsgálata volt, hogy a párhuzamos korpuszokból fordítási ekvivalensek kinyerésére használt módszer kiterjeszthető-e többszavas kifejezésekre is. A kísérletben a többszavas kifejezéseket kizárólag igei szerkezetek alkották. Első lépésként a párhuzamos korpusz célnyelvi és forrásnyelvi oldalából külön-külön nyertük ki az igei szerkezeteket. A jelenlegi fázisban a kinyerés félig automatikus módon történt: az előre meghatározott forrásnyelvi igékhez és ezek célnyelvi fordításaihoz tartozó igei szerkezeteket kerestünk, melyekből kézzel válogattuk ki a céljainknak megfelelő igei szerkezeteket. A következő lépésben ezen igei szerkezeteket egytagú kifejezésekké vontuk össze a korpuszban. Így az összevont igei szerkezetek az illesztési algoritmus bemeneteként szolgálhattak. Annak ellenére, hogy a kiértékelés során a szűrésre használt paramétereknek alacsony értékeket adtunk meg, és így eredményeink meglehetősen alacsony pontosságúak, a kinyert igei szerkezetek egyértelműen mutatják, hogy a módszer jól használható igei szerkezetek fordítási ekvivalenseinek detekciójára.

További feladataink közé tartozik a részleges szintaktikai elemzés és a tagmondathatár-felismerés minőségének javítása a forrásnyelvre és a célnyelvre egyaránt, valamint ezen információ figyelembevétele a szóillesztés bemeneti korpuszában is.

A lefedettség jelentősen növelhető lenne, ha nemcsak előre kiválasztott igékhez tartozó igei szerkezeteket illesztenénk, hanem minden igei szerkezetet, amely gyakran fordul elő a párhuzamos korpuszban. Célunk továbbá, hogy a célnyelvi igei szerkezetek detekciójánál ne a forrásnyelvi igék feltételezett fordításából induljunk ki, hanem ezeket egymástól függetlenül nyerjük ki. Elvárásaink szerint egy ilyen lépés csökkentené a részleges illesztések számát is, és növelné a teljesen jó fordításai ekvivalensek arányát.

További tervezett kutatási irány a módszer kiterjesztése a kollokációkra is.

Bibliográfia

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
2. Dempster, A. P., Laird, N. M., Rubin. D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B Vol. 39 No. 1 (1977) 1–22
3. É. Kiss, K.: Mondattan. In: É. Kiss, K., Kiefer, F., Siptár, P. (szerk.): Új magyar nyelvtan. Osiris Kiadó, Budapest (2003) 15–184
4. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography.In: Proceedings of the LREC2010 Conference. La Valletta, Malta (2010) 2798–2805
5. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007. Birmingham, United Kingdom (2007)
6. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics Vol. 29 No. 1 (2003) 19–51

7. Sass, B.: A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In: Angelova G. et al. (szerk.): Proceedings of RANLP 2009. Borovec, Bulgária (2009) 399–403
8. Wu, D.: Learning an English-Chinese Lexicon from a Parallel Corpus. In: Proceedings of AMTA'94 (1994) 206–213

Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban

Vincze Veronika[1], Felvégi Zsuzsanna[2], R. Tóth Krisztina[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola,
Angol Alkalmazott Nyelvészet PhD Program
felvegi@gyakg.u-szeged.hu
[3] MTA-SZTE Oktatáselméleti Kutatócsoport
tothkr@inf.u-szeged.hu

Kivonat: A természetes nyelvi feldolgozásban az egyik legnehezebb problémát a többszavas kifejezések azonosítása és megfelelő kezelése jelenti. Ezt megkönnyítendő, a SzegedParalell angol–magyar párhuzamos korpusz egy részében kézzel bejelöltük a félig kompozicionális szerkezeteket. A szerkezeteket mindkét nyelven annotáltuk, lehetővé téve ezáltal az angol és magyar szerkezetek automatikus párosítását. Az annotált korpusz jól használható tanuló adatbázisként mind egynyelvű, mind többnyelvű alkalmazásokban, de a kontrasztív nyelvészetben, stilisztikában, illetve a nyelvoktatásban is hasznosítható.

1 Bevezetés

A természetes nyelvi feldolgozásban, különösen a gépi fordítás és fordítástámogatás területén az egyik legnehezebb problémát a többszavas kifejezések megfelelő kezelése jelenti. A többszavas kifejezések sikeres kezelésének első lépése, hogy felismerjük, többszavas kifejezéssel van dolgunk, hiszen például a többszavas kifejezések szintaktikai felépítése hasonlít más, produktív szerkezetek felépítéséhez (*veri az ördög a feleségét – veri a szomszéd a kutyáját*). Ennek automatikus eldöntése igen nehéz feladat, a feladatot tovább nehezíti, ha a felismerés mellett különböző nyelven írt többszavas kifejezések megfeleltetését tűzzük ki célul. A probléma nehézségét mutatja, hogy egy többszavas kifejezés idegen nyelvű megfelelője a legritkább esetben szó szerinti fordítása az eredetinek. Például a félig kompozicionális szerkezetek esetén – melyek olyan, főnévből és igéből álló többszavas kifejezések, ahol a szemantikai fej a főnév, míg az ige pusztán csak a szerkezet igeiségéért felelős –, csak a kifejezés egy részét fordíthatjuk szó szerint (a főnévi tagot: *szerződést köt – make a contract*), az idiómák azonban a bennük szereplő szavak szintjén semmiképpen sem feleltethetők meg egymásnak (*ez nem az én asztalom – it's not my cup of tea*), ezért problémát jelent(h)e(t)nek a különféle számítógépes alkalmazások számára. Az azonosításukra képes algoritmusok fejlesztéséhez és teszteléséhez annotált korpuszokra van szükség.

2 A többszavas kifejezések

E fejezetben bemutatjuk a többszavas kifejezések főbb tulajdonságait, amelyek megnehezítik a többszavas kifejezések automatikus felismerését. Példáink az angol és a magyar nyelvből származnak, ezzel is kiemelve a probléma általánosságát (azaz nyelvektől való függetlenségét).

2.1 A többszavas kifejezések idioszinkratikus tulajdonságai

A többszavas kifejezések olyan lexikai egységek, melyek több szóból állnak, és szintaktikai, szemantikai, pragmatikai vagy statisztikai szempontból idioszinkratikus sajátosságokat mutatnak [1, 4, 8]. A lexikálisan idioszinkratikus kifejezések részei nem helyettesíthetők más, azonos (vagy hasonló) értelmű szavakkal anélkül, hogy elvesztenék eredeti jelentésüket [5, 7]. Például a *fűbe harap* idiómában nem cserélhetjük ki a *fű* szót *pázsit*ra (*pázsitba harap*) a jelentés megőrzése mellett. Hasonlóan az angolban: a *kick the pail* 'felrúgja a vödröt' csakis szó szerint értelmezhető, míg a 'meghal' jelentést csak a *kick the bucket* idióma képes hordozni, noha a *pail* és *bucket* szavak önmagukban szinonimák.

A szintaktikailag idioszinkratikus kifejezések szintaktikai tulajdonságai nem következnek a részek szintaktikai tulajdonságaiból. A magyarban például a *kerek perec* kifejezés határozóként viselkedik, noha szigorúan véve egy melléknév és egy főnév kapcsolatáról van szó. Az *all of a sudden* angol többszavas kifejezés szintén határozószói szerepet tölt be, azonban egy névmás, egy prepozíció, egy névelő és egy melléknév alkotja.

A többszavas kifejezések jelentése többnyire nem (teljesen) kompozicionális, azaz nem számítható ki pusztán részeinek jelentésére és azok kapcsolódási módjára támaszkodva (szemantikai idioszinkrázia). Tipikus példák az idiómák: a fenti *fűbe harap* kifejezés 'meghal' jelentése semmiképpen sem számítható ki a *fű*, a *harap* és a *–bA* toldalék jelentéséből. A kompozicionalitás hiányára az is rámutat, hogy a fenti kifejezés helyes angol változata a *kick the bucket*, amely szavak szintjén semmiképpen sem jó fordítása az eredeti kifejezésnek, viszont jelentés szintjén tökéletesen megfelel egymásnak a két kifejezés.

A pragmatikailag idioszinkratikus kifejezések többnyire egy adott helyzetben vagy adott körülmények között használatosak: például a *Jó étvágyat!* mondat jellemzően étkezés előtt hangzik el, a *How do you do?* kifejezés pedig bemutatkozáskor használatos.

A statisztikai idioszinkrázia azt takarja, hogy a többszavas kifejezés tagjai statisztikailag szignifikánsan nagy valószínűséggel együttesen fordulnak elő, ezeket nevezik kollokációknak [8]. Megjegyezzük azonban, hogy a kollokációkra nem (feltétlenül) jellemző a szemantikai vagy szintaktikai sajátos viselkedés. Néhány példa: a *fekete-fehér* melléknév *fehér-fekete* sorrendben is ugyanazzal a jelentéssel rendelkezne, azonban jelentősen többször fordul elő *fekete-fehér* alakban, mint fordítva, illetve az angol *pepper and salt* 'bors-só' jelentése egyenértékű a *salt and pepper* 'só-bors' szókapcsolatéval, mégis utóbbi számít a bevett kifejezésnek.

Természetesen nem minden egyes idioszinkratikus tulajdonság érvényes minden többszavas kifejezésre: léteznek olyan többszavas kifejezések, melyek például szintaktikailag szabályosan viselkednek, azonban jelentésük nem kompozicionális (ilyen a legtöbb idióma).

2.2 A többszavas kifejezések szintaktikai viselkedése

A többszavas kifejezések szintaxisuk szerint lehetnek kötöttek, félig kötöttek, illetve produktívak [8, 11]. A kötött kifejezések nem mutatnak szintaktikai változatosságot: mindig ugyanabban a formában fordulnak elő [5]. Ilyenek például a közmondások (l. alább). Az automatikus felismerést tovább nehezíti, hogy a félig kötött kifejezések bizonyos fokig módosíthatók: az igék például ragozhatók az idiómákon belül, az összetett főnevek pedig többes száma tehetők. A legnehezebb problémát a szintaktikailag produktív többszavas kifejezések jelentik – a félig kötött kifejezéseknél említett módosítások mellett –, melyek szabadabban változtathatók: tagjaik módosíthatók (például egy jelző módosíthatja a félig kompozicionális szerkezetek főnévi tagját), sőt a szerkezet tagjai nem is feltétlenül szerepelnek egymás mellett a mondatban.

2.3 A többszavas kifejezések típusai

A többszavas kifejezések több csoportba sorolhatók egyrészt az őket alkotó szavak szófaja alapján, másrészt szintaktikai viselkedésük alapján. Az alábbiakban a magyarra és angolra egyaránt jellemző, gyakran előforduló többszavas kifejezések, azaz az összetett szavak, idiómák, közmondások és a félig kompozicionális szerkezetek azon tulajdonságait ismertetjük röviden, amelyek megnehezítik azok automatikus cél- és forrásnyelvi megfeleltetését.

Az összetett szavak olyan lexikai egységek, melyek két vagy több önállóan is létező szóból állnak. Helyesírásukat tekintve léteznek egybeírt összetett szavak (*iskolaigazgató*), kötőjellel írt összetett szavak (*időjárás-jelentés*), illetve szóközt tartalmazó összetett szavak (jellemzően az angol nyelvben, például *power plant*, míg a magyar helyesírási gyakorlat az összetett szavak határterületéhez tartozónak minősíti a különírt állandósult szókapcsolatokat, például *kútba esés*). Természetesen az összetett szavak nem csak főnevek lehetnek, léteznek összetett melléknevek (*red haired*, *nagyotmondó*), összetett határozószók (*above all*, *csakazértis*), összetett prepozíciók (*in front of*) és összetett kötőszavak (*in order that, nehogy*) is. A legtöbb összetett szó szintaktikailag szabályos viselkedést mutat, de pontos jelentésük – azaz az összetétel tagjai közti viszony jellege – még azonos szintaktikai felépítés esetén is változhat: a *repceolaj* repcéből készül, de a *babaolaj* babák számára.

Az idiómák jelentése nem határozható meg részeinek jelentéséből [6, 8], noha szintaktikai viselkedésük általában véve szabályos, szemantikájuk teljességgel megjósolhatatlan: a *fekete bárány* 'a megszokottól eltérően viselkedő (nemkívánatos) személy' jelentésének semmi köze sem az állathoz, sem a fekete színhez. Az idiómák általában mutatnak morfológiai változatosságot (például a bennük szereplő ige ragozható).

A közmondások a legtöbb ember által igaznak tartott állításokat fejeznek ki, leginkább egy teljes mondat formájában (*Ki korán kel, aranyat lel*). Emiatt általában ugyanabban a formában fordulnak elő, szemben az idiómákkal.

A félig kompozicionális főnév + ige szerkezetekben (pl.: *tanácsot ad, döntést hoz, virágba borul*) a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában [14]. Mivel jelentésük nem teljesen kompozicionális, a szerkezet elemeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Emellett a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*) [2], így azonosításuk nem valósulhat meg pusztán szintaktikai mintákat figyelembe véve. Végül, mivel a szerkezet szintaktikai és szemantikai feje nem azonos, a szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni – az angol vonzatos igékhez (phrasal verbs) hasonlóan.

A fenti okokból kifolyólag a többszavas kifejezések kezelése különleges figyelmet érdemel a természetes nyelvi alkalmazásokban. Ennek első lépéseként azonosítani kell a többszavas kifejezéseket, mely célhoz különféle algoritmusok fejlesztése segíthet hozzá. Ebben sikeresen hasznosítható egy kézzel annotált tanító adatbázis: korpuszunk építésekor ezt a szempontot tartottuk szem előtt.

3 A SzegedParalell párhuzamos korpusz

A SzegedParalell korpusz építésére két, nemzetközi viszonylatban is alkalmazott alternatíva merült fel: 1) a korpusz építéséhez már meglévő, egynyelvű annotált korpuszt használnak fel, annak szövegeit lefordítják a célnyelvre, majd a lefordított szövegeket is feldolgozzák; 2) a korpusz építésekor két nyelven elérhető szövegeket keresnek, majd ezeket a nyers szövegeket mindkét nyelven annotálják. A SzegedParalell építése során az utóbbi megoldást választottuk, mert a fordítás hosszadalmasabb és költségesebb procedúra, mint a kétnyelvű szövegek gyűjtése. A korpusz felépítését tekintve az alábbi témákból tartalmaz szövegeket.

1. táblázat: A SzegedParalell korpusz felépítése.

Témakör	ME (db)
tankönyvi mondatok	3.496
Európai Unióról szóló szövegek	1.518
kétnyelvű magazinok	5.320
irodalmi és történelmi alkotások	88.716
egyéb	695
összesen	***99.745***

Megjegyzés: Mondatszinkronizációs egységek (ME), l. lentebb.

Az első korpuszrész többnyire Dévainé Angeli Mariann *Angol nyelvtani gyakorlatok* és Dohár Péter *Kis angol nyelvtan* című könyvének különálló párhuzamos mondataiból áll, melyek az angol nyelvtan sajátosságait hivatottak reprezentálni. A nyelvtankönyvi mondatok mellett autentikus szövegeket is beépítettünk a párhuzamos korpuszba, így biztosítva az egyensúlyt a mesterkélt és a természetes nyelvi szerkezetek között. Az autentikus szövegek elsősorban kétnyelvű magazinokból, interneten található általános nyelvezetű, hétköznapi, gazdasági és jogi témájú szövegeket tartalmaznak, és változatos szókinccsel rendelkeznek. A kétnyelvű magazinok alkorpuszát a *Horizon Magazin*, illetve a *Resource Ingatlan Info* c. újságok alkotják. A Horizon Magazin (a Malév fedélzeti magazinja) sokféle, hétköznapi témát ölel fel, mint kultúra, utazás, interjúk hírességekkel, nevezetesebb városok bemutatása, a Resource Ingatlan Info pedig elsősorban területfejlesztésről, építőipari beruházásokról, logisztikáról, és az ezekhez kapcsolódó témakörökről közöl cikkeket. A korpusz tartalmaz továbbá Európai Unióról[1] szóló rövid ismeretterjesztő cikkeket (pl.: az EU történetéről, zászlójáról, himnuszáról, pénzneméről stb).

Az irodalmi szövegeket tartalmazó korpuszrész a Hunglish korpusz [3] irodalmi műveinek egy részét tartalmazza. Elsősorban modern angol irodalmi műveket építettünk be a korpuszunkba. Továbbá a Magyar Elektronikus Könyvtár weboldalán elérhető kétnyelvű szövegeket (történelmi és irodalmi művek) emeltünk be a korpuszba.

Az egyéb kategóriában rövid terjedelmű szövegek szerepelnek, melyek korábbi internetes gyűjtések eredményei: rövid beszámoló kulturális eseményről, tudományos feltárások és receptek.

A szövegek párhuzamosítása során először a szövegeket normalizáltuk, majd ellenőriztük a fordítás helyességét, szükség esetén javítottuk azt. Az egyes magyar és angol nyelvi szövegeket külön fájlokban mentettük. A fájlok szinkronizálása után a parallel szövegek *bekezdésszintű* összerendelését automatikusan végeztük, mert szakfordítói tapasztalatokra alapozva a forrásnyelvi és a célnyelvi szövegek egyenlő számú bekezdést tartalmaznak, és ezek sorrendje nem felcserélhető.

A *mondatillesztés* alapaspektusa az a fordítási tény, hogy a fordítási egységek nem nyúlhatnak át bekezdések határán. A mondatok illesztése a bekezdés-összerendeléssel szemben nem egy kölcsönösen egyértelmű reláció, melynek okai:

(1) a bekezdésben szereplő mondatok sorrendje felcserélhető, illetve
(2) egy forrásnyelvi mondat a célnyelven több mondatnak is megfelelhet, így a célnyelvi egységek adott esetben túlnyúlhatnak egy mondat határán.

A párhuzamos szövegek feldolgozása során az alábbi összerendelési lehetőségek (mondatszinkronizációs típusok) fordultak elő:

- 1:1 egy kiindulási nyelvi mondatnak egy célnyelvi megfelelője van (*egyezés*)
- 1:0 a kiindulási nyelvi mondatnak nincs tartalmi megfelelője a célnyelven (*kihagyás*)
- 0:1 a célnyelvi mondat nem szerepel a forrásnyelvi szövegben (*betoldás*)
- 1:N egy kiindulási nyelvi mondatnak több célnyelvi megfelelője van (*szétbontás*)
- N:1 több kiindulási nyelvi mondatnak egy célnyelvi mondat felel meg (*összevonás*)

[1] A http://europa.eu.int weboldalról és a Wikipédia weboldaláról.

- N:M több kiindulási nyelvi mondatnak több célnyelvi mondat felel meg (ez többnyire szépirodalmi művekben fordul elő)

Ezeket az egymásnak megfeleltetett egységeket mondatszinkronizációs egységeknek nevezzük, melyek segítségével feltérképezzük a magyar és angol nyelvi félig kompozicionális szerkezetek előfordulási gyakoriságát, illetve az angol és magyar nyelvű kifejezéseinek automatikus megfeleltetése során felmerülő problémákat.

4 A korpusz annotálása

A többszavas kifejezések természetes nyelvi feldolgozását megkönnyítendő a SzegedParalell angol–magyar párhuzamos korpuszban [10] kézzel bejelöltük a félig kompozicionális szerkezeteket (rövidítve: FX) [13].

4.1 Az annotált korpusz

A korpusz szövegei közül elsődlegesen az újságcikkeket, Európai Unióról szóló szövegeket és a tankönyvi mondatokat annotáltuk, merthogy a félig kompozicionális szerkezetek nagyszámú előfordulása a sajtónyelvben és a gazdasági-politikai doménben várható [14], illetve a nyelvkönyvekben fordításra szánt mondatok valószínűleg nagy arányban tartalmaznak többszavas kifejezéseket (azaz olyan nyelvi elemeket, amelyeknek a fordítása nem szó szerinti). Az összehasonlítás végett azonban néhány irodalmi szövegben is jelöltük a félig kompozicionális szerkezeteket. Így egyrészt a különféle témájú szövegekből nyert adatok segítségével kvantitatívan alá tudjuk támasztani (vagy meg tudjuk cáfolni) a fenti feltételezéseket, másrészt pedig a nyelvek és domének közti összevetésből olyan minőségi jellegű kérdések megválaszolására is sor kerülhet, hogy miként fordítja a szépirodalmi fordítás vagy egy gazdasági-jogi témájú szöveg a félig kompozicionális szerkezeteket (azaz a félig kompozicionális szerkezetnek szintén szerkezet felel-e meg a másik nyelvben, avagy szabadabban fordítják).

Az annotált korpusz méretét a következő táblázat mutatja:

2. táblázat: Az annotált korpusz felépítése és mérete.

Téma	Szövegek száma (db)	ME (db)	FX - magyar	FX - angol
EU	30	1518	295 (19,4%)	227 (15%)
Újságcikkek	151	5320	477 (9%)	400 (7,5%)
Tankönyvi mondatok	7	3496	85 (2,4%)	131 (3,7%)
Irodalmi szövegek	3	3232	247 (7,6%)	336 (10,4%)
Egyéb	5	695	8 (1,2%)	6 (0,8%)
Összesen	196	14261	1112 (7,8%)	1100 (7,71%)

A korpuszban levő angol és magyar félig kompozicionális szerkezetek száma megközelítőleg ugyanaz, emiatt a mondatszinkronizációs egységek (ME) közel ugyanakkora hányada tartalmaz félig kompozicionális szerkezetet (l. százalékos értékek). Ez azonban nem jelenti azt, hogy minden egyes félig kompozicionális szerkezetnek megvan a másik nyelvű megfelelője – más szóval, vannak olyan szerkezetek, amelyek csak a magyar, illetve csak az angol korpuszban szerepelnek.

4.2 Annotációs elvek

A félig kompozicionális szerkezeteket angol és magyar nyelven egyaránt annotáltuk. A szövegeken három annotátor dolgozott egységes annotációs útmutató alapján egy nyelvész szakértő irányításával. Ugyanazon szöveg forrás- és célnyelvi változatában ugyanaz a személy jelölte be a félig kompozicionális szerkezeteket.

Mivel a félig kompozicionális szerkezetek szintaktikailag produktívak (vö. 2.2), többféle formában is előfordulhattak a szövegekben. Hasonlóan a Szeged Korpusz korábbi annotálási elveihez [13], itt is a következő altípusokba soroltuk a szerkezeteket az annotáció során:

Főnév + ige kombinációja (VERB): *bejelentést tesz, igénybe vesz, take a look, pay a visit*

Igenevek (PART): *gondot viselő, kézbe véve, photos taken, taking part*

Főnévi változat (NOM): *szerződéskötés, bérbe vétel, service provider, decision-maker*

Különálló szerkezet (SPLIT): *előadást fog tartani, kapott tőlük engedélyt, contest is held, effort you make*

A szerkezetek altípusainak megoszlása a következő táblázatban látható:

3. táblázat: A félig kompozicionális szerkezetek altípusainak megoszlása.

	VERB		**PART**		**NOM**		**SPLIT**	
	angol	magyar	angol	magyar	angol	magyar	angol	magyar
EU	132	158	30	76	24	32	41	29
Újságcikkek	281	330	48	86	16	23	55	38
Tankönyvi mondatok	106	62	4	10	13	3	8	10
Irodalmi szövegek	222	196	13	11	5	0	96	40
Egyéb	4	7	1	0	0	0	1	1
Összesen	*745*	*753*	*96*	*183*	*58*	*58*	*201*	*118*

Míg az igei és a főnévi alakok száma nagyrészt megegyezik a két nyelvben, addig a melléknévi igenevek és a különálló szerkezetek száma jelentősen eltér egymástól. Ez a különbség valószínűleg nyelvtani okokra vezethető vissza: a SPLIT kategóriába sorolt elemek nagy része az angolban például passzív szerkezetet alkot, ahol is a szerkezet főnévi komponense alanyi funkciót tölt be, ezáltal nem szomszédos az igével. A PART kategória esetében pedig előfordul, hogy míg a magyarban a főnévi komponensnek előmódosítója van, mely megköveteli az igei komponens melléknévi igenév formájában való jelenlétét is, addig az angolban utómódosítót találunk, amely elől elmaradhat az igenév, például:

*az emberi jogokba **vetett** hit*
a belief in human rights

Természetesen további részletes vizsgálatok más tendenciákra is fényt deríthetnek e különbségek elemzésében.

5 A kétnyelvű szerkezetek párosítása

A kétnyelvű annotáció (l. alábbi példa) lehetővé teszi az angol és magyar szerkezetek automatikus párosítását, mivel a mondat szinten párhuzamosított, annotált korpuszban egyszerűen megtalálható az adott kifejezés másik nyelvű megfelelője, ha a mondatszinkronizációs egységek egy-egy félig kompozicionális szerkezetet tartalmaznak:

*A mulatozások fő időszaka a 15-16. századra tehető, amikor ezek a bálok **fontos szerepet játszottak** a párválasztásban.*
*Such revelry can be claimed to have reached its height in the 15th and 16th centuries, when the dances **played an important role** for those in search of a good match.*

A szerkezetek automatikus szinkronizálása természetesen külön kézi ellenőrzést igényel, amennyiben az adott mondaton belül több félig kompozicionális szerkezet is előfordul. Sass Bálint [9] beszámol egy igei szerkezetek párhuzamos korpuszból való kinyerésére szolgáló eljárásról, mely egy korábbi, igéket és azok bővítményeit kinyerő algoritmusra épül. A módszer lényege, hogy a tagmondatok igéit egymás mellé rendelve egy komplex ige jön létre, melyhez a bővítményeket halmazként rendeljük hozzá, felcímkézve őket aszerint, hogy melyik nyelvű részkorpuszból származnak. Az így kapott reprezentációból az eredeti algoritmus segítségével lehet kigyűjteni az egyes nyelvekre jellemző igei szerkezeteket. A módszer előnye, hogy nemcsak a szerkezet-szerkezet párokat képes megtalálni, hanem azokat az eseteket is, amikor a szerkezetnek ige felel meg. A módszer nyelvfüggetlen, tehát korpuszunkra is alkalmazható, a kézi annotációnak köszönhetően pedig a kiértékelés is egyszerűbb.

6 Eredmények

A különböző doméneket tekintve elmondhatjuk, hogy valóban a gazdasági-jogi témájú, Európai Unióról szóló szövegekben fordul elő arányukban a legtöbb félig kompozicionális szerkezet. Ezzel ellentétben a tankönyvi példamondatok elenyésző részében találhatunk félig kompozicionális szerkezetet, vagyis vélhetően nem annyira a lexikai, mint inkább a nyelvtani szempontok domináltak a mondatsorok összeállításakor. További érdekesség, hogy az angol irodalmi szövegek jelentősen nagyobb arányban tartalmaznak félig kompozicionális szerkezetet, mint magyar megfelelőik, ez különösen Swift *Gulliver utazásai* c. regényére igaz (a regény mondatainak 16%-a tartalmaz félig kompozicionális szerkezetet, az angol nyelvű szövegek közül ez a legmagasabb arány). Mivel a regény 1726-ban jelent meg[2], a korai XVIII. századi angol nyelvállapotot tükrözi, azonban – kellő számú adat híján – elhamarkodott lenne arra a következtetésre jutni, hogy a korabeli angol nyelvben jóval több félig kompozicionális szerkezet szerepel, mint a mai angolban: ennek alátámasztásához további – nyelvtörténeti – vizsgálatokra van szükség.

Ha a szövegek témája szerint vizsgáljuk a félig kompozicionális szerkezetek megfeleltetését, azt találjuk, hogy az újságcikkekben és az EU-ról szóló szövegekben nagy arányú a megfelelés a szerkezetek között (azaz egy magyar félig kompozicionális szerkezet angol párja is nagy valószínűséggel félig kompozicionális szerkezet), azonban a szépirodalmi szövegre ez nem áll. Egyrészt az angol irodalmi szövegekben szám szerint jóval több a félig kompozicionális szerkezet, mint a magyar szövegekben (a Mark Twain-regény kivételével, ahol kiegyenlített a számuk), másrészt igen gyakori az a jelenség, hogy a félig kompozicionális szerkezetnek a másik nyelvi megfelelője nem szerkezet (sőt nem is mindig ige), például:

[…] during which time, the emperor ***gave orders*** *to have a bed prepared for me.*
[…] ez idő folyamán, a császár ***parancsára****, fekvőhelyet készítettek nekem.*

Mindebből az következik, hogy a szépirodalmi szövegek kevésbé használhatók mint tanító, illetve tesztadatbázis a szerkezetek automatikus szinkronizálásához, mint például a gazdasági-jogi jellegű szövegek vagy újságcikkek.

Az angol és a magyar nyelvű félig kompozicionális szerkezetek összehasonlításával megállapítható, hogy számos forrásnyelvi félig kompozicionális szerkezetnek szintén egy célnyelvi szerkezet felel meg, ezáltal lehetővé válik az automatikus párosítás (l. fent). Vannak azonban olyan esetek is, amikor az egyik nyelvben félig kompozicionális szerkezetet találunk, a másik nyelv viszont igét alkalmaz:

I don't usually moan or ***make any special requests****.*
Nem vagyok nyűgös, ***nincsenek extra kívánságaim****.*

[2] A SzegedParalellben Vajdafy Ernő 1906-os fordítása szerepel, amely azonban tudatosan archaizáló nyelvezetű, így a keletkezésük között eltelt közel két évszázad ellenére is összemérhetőnek ítéltük meg a két szöveget.

Ennek speciális esete az, amikor jellemzően a szerkezet főnévi komponensével azonos tőből származó, azonos jelentésű igei variáns [12] szerepel a célnyelven:

*It **decided** to welcome 10 more countries to join the EU on 1 May 2004.*
*A Tanács **meghozta a döntést** arról, hogy 2004. május 1-jén 10 új államot vesznek fel az Unió tagállamai sorába.*

Egy másik érdekesség, hogy az angol passzív szerkezetek magyar megfelelői időnként a *kerül* igét tartalmazó félig kompozicionális szerkezetek:

*The song "Auld Lang Syne" **was** partially written by Robert Burns and **published** after his death in 1796.*
*A híres „Auld Lang Syne" („Régóta már") című dalt részben Robert Burns írta, és halála után, 1796-ban **került kiadásra**.*

E nyelvek közti különbségek elemzése mind a gépi fordítás, mind a fordítástudomány számára haszonnal bírhat.

7 A korpusz felhasználhatósága

Az annotált korpusz jól használható tanuló adatbázisként más mind egynyelvű, mind többnyelvű alkalmazásokban (például többnyelvű információ-visszakeresés), de a kontrasztív nyelvészetben, stilisztikában, illetve a nyelvoktatásban is hasznosítható.

Az adatbázis oktatási és kutatási célokra ingyenesen elérhető a Creative Commons licenc alatt a www.inf.u-szeged.hu/rgai/nlp címen.

8 Összegzés

Cikkünkben bemutattuk a SzegedParalell angol-magyar párhuzamos korpusznak félig kompozicionális szerkezetekre annotált verzióját. Az elkészült adatbázis mintegy 1100 szerkezetet tartalmaz mind angol, mind magyar nyelven (noha a forrásnyelvi többszavas kifejezés célnyelvi megfelelője nem minden esetben többszavas kifejezés). Az annotált korpusz hasznosítható különféle, a többszavas kifejezések automatikus felismerésére készített algoritmusok tanításában és kiértékelésében, ezenkívül a gépi fordítás és a többnyelvű információ-visszakeresés számára is haszonnal bírhat, de nyelvészek is sikeresen építhetik be kutatásaikba az itt elért eredményeket.

Köszönetnyilvánítás

Szeretnénk köszönetet mondani a korpusz annotátorainak áldozatos munkájukért.

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002). Las Palmas (2002) 1934–1940
2. Fazly, A., Stevenson, S.: Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. Association for Computational Lingustics (2007) 9-16
3. Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T., Vonyó A.: A hunglish korpusz és szótár. In: Alexin Z., Csendes D. (szerk.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2005) 134–142
4. Kim, S.N.: Statistical Modeling of Multiword Expressions. PhD thesis, University of Melbourne (2008)
5. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
6. Nunberg, G., Sag, I. A., Wasow, T.: Idioms. Language, Vol. 70 (1994) 491–538
7. Oravecz, Cs., Nagy, V. Varasdi, K.: Lexical idiosyncrasy in MWE extraction. In: Proceedings from the Corpus Linguistics Conference Series, Vol. 1, No. 1. Birmingham (2005)
8. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) Proceedings of Conference on Intelligent Text Processing and Computational Linguistics 2002. Mexico City (2002)
9. Sass B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 102–110
10. Tóth, K., Farkas, R., Kocsor, A.: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. Acta Cybernetica Vol. 18, No. 3 (2008) 463–478
11. Váradi T.: Többszavas kifejezések kezelése MT szótárban. In: Alexin Z., Csendes D. (szerk.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2005) 233–244
12. Vincze V.: Angol–magyar főnév + ige szerkezetek és igei párjaik. In: Váradi T. (szerk.): II. Alkalmazott Nyelvészeti Doktorandusz Konferencia. MTA Nyelvtudományi Intézet, Budapest (2009) 113–123
13. Vincze V.: Félig kompozicionális szerkezetek a Szeged Korpuszban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (MSzNy 2009). Szegedi Tudományegyetem, Szeged (2009) 390–393
14. Vincze, V., Csirik, J.: Hungarian Corpus of Light Verb Constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, China (2010) 1110–1118

Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból

Sass Bálint

MTA Nyelvtudományi Intézet
e-mail: sass.balint@nytud.hu

Kivonat Jelen dolgozatban egy egynyelvű korpuszra kifejlesztett, igei szerkezeteket kinyerő eljárást alkalmazunk holland-francia párhuzamos korpuszra, a korpuszreprezentáció alkalmas átalakításával, kétnyelvű, párhuzamos igei szerkezetek kinyerése céljából. Az igei szerkezetek közül kiemelendők a vonzatos komplex igék, melyekben az ige mellett egy névszói kollokátum, valamint vonzat is áll (pl.: *részt vesz vmiben*). A nyelvtechnológiai alkalmazások (pl.: a gépi fordítás) lexikális erőforrásainak tartalmaznia, ismernie kell ezeket a kifejezéseket, hogy magas nyelvi minőségű kimenetet adhassanak. Ezek a szerkezetek ugyanakkor sok esetben más nyelvre lefordítva teljesen más formát mutatnak. Bár a szükséges elemző lépések során alkalmazott egyszerű közelítő módszerek, valamint a feladat nehézsége miatt a kinyerés pontossága nem kiemelkedő, jelen dolgozatból világos, hogy az algoritmus képes különféle, akár aszimmetrikus, párhuzamos szerkezetek feltérképezésére is.

1. Bevezetés

A többszavas kifejezésekkel foglalkozó szakirodalom legnagyobb része a kételemű, két tagból álló kifejezésekkel foglalkozik [1]. Siepmann [2, 412. oldal] szerint általánosan elfogadott a kutatók között, hogy a kollokációk bináris egységek. Számtalan asszociációs mértéket dolgoztak ki [3], melyekkel két tag közötti kapcsolat szorossága mérhető. A kettőnél több tagú kifejezések kezelésével ritkábban foglalkoznak, az ide tartozó módszerek három csoportra oszthatók [4, 5.1 fejezet]: egyrészt megpróbálhatjuk az asszociációs mértékeket kettőnél több elemre kiterjeszteni; alkalmazhatunk iteratív kollokációkinyerő módszereket, ahol a már kinyert kéttagú kollokációk a következő iterációban összevont elemként egy nagyobb kiterjedésű kollokáció részét képezhetik; valamint a kinyert bigramokat utólag feldolgozva is következtethetünk bizonyos többtagú kollokációk meglétére.

A többelemű kifejezések között speciális csoportot alkotnak a vonzattal is bíró komplex igék. Ide tartozik a *kilátásba helyez vmit* vagy a *részt vesz vmiben*. Ezekben a szerkezetekben *négy* egységet különíthetünk el: az igét, a vonzatot (magyarban esetrag képviseli), a komplex ige névszói elemét, valamint e névszói elem esetragját. A nyelvekre általában jellemző, hogy a komplex igék névszói elemét és a vonzatot *ugyanazokkal* a nyelvi eszközökkel kapcsolják az igéhez, legyen az esetrag, névutó, elöljáró, igei partikula vagy akár sorrendi megkötés, mint az angol tárgy esetében. Emiatt ezek a „négyelemű kollokációk” speciális

kezelést igényelnek: az őket megcélzó lexikai kinyerő eljárásnak fel kell ismernie, hogy az adott bővítményi elem lexikálisan kötött módon a komplex ige része-e (*kilátásba*, *részt*), vagy pedig vonzat, mely esetben a konkrét szó nem része a szerkezetnek, csupán a viszonyjelölő (*vmit*, *vmiben*).

Nyilvánvalónak tűnik, hogy ezek a szerkezetek csak a vonzatukkal együtt teljesek, csak teljes formájukban tudnak hozzájárulni nyelvtechnológiai alkalmazások teljesítményének javításához, például tipikusan egy gépi fordítóban használt lexikai adatbázis elemeként. Mégis a korábbi kutatásokra jellemző, hogy elfogadják helyes eredménynek a hiányos szerkezeteket is. A kollokációkutatók sokszor megfeledkeztek arról, hogy a kollokációknak vonzatuk is lehet, amint ez az [5] cikkben idézett *zur Verfügung stellen* 'rendelkezésre bocsát' szerkezet esetében is kitűnik. Ebben a cikkben csak az elöljáró+főnév+ige típusú szerkezeteket vizsgálták, ennek megfelelően a fenti szerkezet inherens részét képező *tárgy* megtévesztő módon elmarad. Siepmann [2, 416. oldal] is hangsúlyozza: ,,az igei kollokációk és a vonzatok szorosan összefüggnek, számos ige+főnév kollokáció a vonzatok adott disztribúcióját kívánja meg . . . a vonzatuktól megfosztott ige+főnév kombinációk nem tekinthetők teljes értékű szerkezetnek".

Visszatérve a gépi fordításos példánkra, gondolhatnánk, hogy a tárgy elmaradása nem is jelent nagy problémát, mert amit az egyik nyelv tárggyal fejez ki, azt ,,nyilván" a másik is ugyanúgy tárggyal jeleníti meg. Ez azonban egyáltalán nem mindig igaz, és még kevésbé igaz az egyéb esetragokra/elöljárókra, melyek a legváltozatosabb mintázatokban felelhetnek meg egymásnak két nyelv viszonylatában.

Rendelkezésünkre áll egy korábban kifejlesztett nyelvfüggetlen lexikai kinyerő eljárás, mely képes feltérképezni egy egynyelvű korpuszban található különböző bonyolultságú igei szerkezeteket az egyszerű vonzatkeretektől (pl.: *alkalmazkodik vmihez*) a bonyolultabb, akár vonzatos, komplex igékig (pl.: *vállat von*, *örömét leli vmiben*) [6]. Az eljárást sikerrel alkalmaztunk egy egynyelvű szótár előállítása során [7].

Egy gépi fordításban közvetlenül hasznosítható kétnyelvű lexikai adatbázis vagy szótár összeállításához azonban kétnyelvű, *párhuzamos* igei szerkezetekre van szükség. Jelen dolgozatban azt vizsgáljuk, hogy hogyan adaptálható a [6]-ban leírt eljárás párhuzamos korpuszra. Azaz arra a feladatra, hogy bemenetként párhuzamos korpuszt dolgozzon fel, eredményként pedig párhuzamos igei szerkezeteket (igei szerkezeteket és a fordításukat) szolgáltasson. A távlati cél kétnyelvű lexikai adatbázis létrehozása, mely az igei szerkezetek szintjén mutatja be a két nyelv egymásnak megfelelő elemeit. Mivel az algoritmus az igei szerkezetek teljes spektrumát lefedi, azt várjuk, hogy szükség esetén képes lesz párba állítani *különböző* felépítésű szerkezeteket is: képes lesz megragadni azokat az eseteket is, amikor az egyik nyelv egyszerű igét használ ugyanarra, amit a másik nyelv komplex ige segítségével ír körül.

2. Módszer

Az eredeti lexikai kinyerő eljárás [6] tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszt vár bemenetként. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzésnek pedig meg kell határoznia a tagmondat igéjét, a bővítmények fejét, valamint az ige és a bővítmények közötti viszonyjelölőket. Az ilyen formátumú bemenetet feldolgozva a gyakori bővítménykereteket számbavéve az algoritmus automatikusan állítja elő a jellegzetes igei szerkezetek listáját. Az eredeti algoritmus a következőképpen működik:

1. Először is vesszük a korpuszból az összes tagmondatot. A maximum két bővítményt tartalmazó tagmondatokból váltakozva töröljük a bővítményi fejeket, így előállítjuk a tagmondatoknak megfelelő lehetséges szerkezeteket. A *Társasház jön létre,* tagmondatból kialakuló lehetséges szerkezetek: *társasház jön létre*, *vmi jön létre*, *társasház jön vmire* és *vmi jön vmire*. (Ezek közül nyilván a második a kívánt helyes szerkezet.) Erre az átalakításra azért van szükség, hogy a szerkezetlistában megjelenhessenek a szabad viszonyjelölőt (azaz esetleges vonzatot) tartalmazó szerkezetek.
2. Hossz szerint csökkenő sorba rendezzük az igei szerkezetek 1. lépés szerint kiegészített teljes listáját. Egy szerkezet hosszát a benne található esetek és fejek összesített száma adja.
3. A leghosszabbtól kezdve sorra elhagyjuk azokat a szerkezeteket, melyeknek a gyakorisága 5-nél kisebb. Az elhagyott szerkezetek gyakoriságát az első olyan rövidebb keret gyakoriságához adjuk hozzá, mely illeszkedik az eredeti keretre. (A *vmi jön létre* 3 hosszúságú keret például illeszkedik a *társasház jön létre* 4 hosszúságú keretre.) A listán még egyszer végighaladva ellenőrizzük, hogy az elhagyott szerkezetek gyakorisága mindig valóban a lehető legspecifikusabb megmaradó szerkezethez rendelődjön hozzá.
4. Végül a megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

3. A módszer alkalmazása párhuzamos korpuszra

Jelen munkálathoz a Dutch Parallel Corpus (Holland Párhuzamos Korpusz) [8] francia-holland részét használtuk. Ez egy könnyen hozzáférhető, morfológiailag elemzett korpusz, mely 3,2 millió holland és 3,6 millió francia tokent tartalmaz. A nyelvválasztás lehetőséget ad arra, hogy az eredetileg magyar nyelvre használt algoritmus nyelvfüggetlenségét alátámasszuk.

Az előfeldolgozás első lépéseként elvégeztük a tagmondatra bontást mindkét nyelvre. Egyszerű, szabályalapú módszerünk a következő szabályokat tartalmazta. A mondathatáron kívül tagmondathatárt jelentett a kötőszó, az alárendelt tagmondatot bevezető holland *te*, ill. francia *pour*, a vonatkozó névmás és bizonyos írásjelek (vessző, kettőspont és pontosvessző) is, amennyiben a legutóbbi tagmondathatár óta szerepelt a mondatban ige. A részleges szintaktikai elemzést szintén egyszerű szabályok használatával valósítottuk meg. A tagmondatokban

lévő főnevek (illetve a reflexív igék miatt a holland *zich* és a francia *se*) lettek a bővítményi fejek, az elöljárók pedig a viszonyjelölők. A francia *à* elöljáró + *le* névelő összevonásából keletkező *au* szócska szótövét a korpuszban lévő *au*-ról *à*-ra javítottuk, így egységesen kaptuk meg az összes *à* elöljárós bővítményt; hasonlóan jártunk el a *de* + *le* = *du* esetében is. Ha nem találtunk a fej előtt elöljárót, akkor az ige előtt alanyként, az ige után pedig tárgyként kezeltük a szóban forgó bővítményt.

Az így előállított két elemzett „félkorpuszból" a következő módon alakítottuk ki a kétnyelvű bemeneti korpuszt:

1. Az igét tartalmazó tagmondatokat fordítási egységenként sorra egymáshoz rendeltük (a fordítási egység első holland tagmondatához a megfelelő fordítási egység első francia tagmondatát stb.). Ha a fordítási egység nem azonos számú tagmondatot tartalmazott, akkor a fennmaradó(ka)t figyelmen kívül hagytuk.
2. Az egymáshoz rendelt tagmondatok holland, ill. francia igéjéből egy igepárt hoztunk létre (pl.: *gaan+aller* 'megy'), ez játssza majd az eredeti eljárás igéjének szerepét.
3. A tagmondatpárban található bővítményi csoportokat (mindkét nyelvűeket) egy halmazként soroltuk fel az igepár mellett, az egyes bővítményeket a megfelelő nyelv kódjával megjelölve.

A fenti lépések során egyfajta metakorpuszt alakítottunk tehát ki, mely párhuzamos tagmondatokból áll, a két eredeti tagmondat igéje egy metaigét alkot, a bővítmények pedig egy egyesített halmazként állnak a metaige mellett. A reprezentációt az 1. ábrán látható példa szemlélteti.

holland tagmondat: *Ze geloofde in de grote liefde*
francia tagmondat: *Elle croyait au grand amour*
magyar fordítás: 'Hitt a nagy szerelemben'

reprezentáció: geloven+croire in_{nl}:liefde à_{fr}:amour

1. ábra. Példa a kétnyelvű bemeneti korpuszból. Az igepárt '+' jel kapcsolja össze, az elöljárókat alsóindex sorolja a megfelelő nyelvhez, kettőspont után pedig az elöljáróhoz tartozó főnévi fej szerepel.

Ezek után az így kialakított kétnyelvű reprezentációra közvetlenül futtattuk az eredeti algoritmust. Mindössze két apróbb szükséges változtatást tettünk meg:

– Az algoritmus eredetileg két bővítményi pozíciót kezelt, ezt most *négyre* bővítettük, hogy hogy megkaphassuk azokat a párhuzamos szerkezeteket is, melyben mindkét nyelvben 2-2 (tehát párhuzamos szerkezetenként összesen négy) lényeges bővítmény van.

– A három és négy pozíciót tartalmazó keretek közül a vonzatos komplex ige formájúak hosszához hozzáadtunk egy 0,2-t. Így ezeknek a szerkezeteknek az esélyét megnöveltük, hogy az algoritmus 3. lépésében (104. oldal) a kiesőktől gyakoriságot örökölhessenek. E heurisztika hatására a végső listában több vonzatos komplex igét kaptunk.

4. Kiértékelés

A bemeneti kétnyelvű metakorpuszban 20-szor vagy annál többször előforduló 1356 igepárra futtattuk az algoritmust. Bár számos egy vagy két egyszerű vonzatot tartalmazó szerkezet is került az eredménylistára (l. 2. ábra), a kiértékelést mégis csak a legizgalmasabb részre, a (leggyakoribb) vonzatos komplex igékre korlátoztam.

párhuzamos szerkezet: geven+donner OBJ_{nl} aan_{nl} OBJ_{fr} à_{fr}
holland szerkezet: *geven OBJ aan*
francia szerkezet: *donner OBJ à*
magyar megfelelő: 'ad vmit vkinek'

párhuzamos szerkezet: geloven+croire in_{nl} à_{fr}
holland szerkezet: *geloven in*
francia szerkezet: *croire à*
magyar megfelelő: 'hisz vmiben'

2. ábra. Példák egyszerű vonzatot tartalmazó szerkezetekre. A párhuzamos szerkezetekből egyszerűen levezethetők a holland és francia szerkezetek, így a párhuzamos szerkezet közvetlenül megmutatja az adott igével használandó megfelelő elöljárót.

Összesen 67 olyan, legalább 15-ös gyakorisági értékkel bíró szerkezetet kaptunk, melyben vonzati pozíció és lexikálisan kötött bővítményi pozíció is volt. Az alábbi szempontok alapján fogadtam el egy párhuzamos szerkezetet helyesnek:

– Ami értelmes, az helyesnek számít, függetlenül attól, hogy idiomatikus-e a jelentése vagy sem.
– A holland *van*, ill. francia *de* általában az elemzés által egyáltalán nem kezelt birtokos szerkezetek miatt jelent meg. Ezeket nem vettük figyelembe, nem befolyásolták a szerkezetek helyességét.
– Az alany és a tárgy megállapítása nem tökéletes, ezért az alany és a tárgyat egymás helyett is elfogadtuk.
– Helyesnek fogadtuk el a szerkezetet akkor is, ha határozószó hiányzott belőle, mivel az elemzés nem kezelte a határozószókat.
– A hiányos szerkezetek nem jók, a helyességhez szükséges minden lényeges elem megléte.

1. táblázat. A kinyert 34 helyes vonzatos komplexige-szerkezet. A második és harmadik oszlopban a párhuzamos szerkezetből derivált holland, illetve francia szerkezet olvasható. A negyedik oszlopban a párhuzamos szerkezet gyakorisági értéke található. Az elöljárót a hozzá tartozó szóhoz kettőspont kapcsolja.

#	holland szerkezet	francia szerkezet	gyak	magyar megfelelő	megjegyzés
1.	*gaan om*	*agir se de*	114	'szó van vmiről'	'agir se de' 1. megfelelője
2.	*zijn OBJ*	*agir se de*	69	'vmi van'	'agir se de' 2. megfelelője
3.	*houden rekening(OBJ) met*	*tenir compte(OBJ) de*	40	'számításba vesz vmit'	met ∼ számol vmivel, tenir ∼ számon tart vmit
4.	*hebben OBJ*	*avoir besoin(OBJ) de*	39	'szükség van vmire'	holland: határozószó ('nodig') hiányzik
5.	*bestaan uit*	*composer se de*	35	'áll vmiből'	aszimmetrikus
6.	*stellen te:beschikking van*	*mettre à:disposition de*	31	'rendelkezésére bocsát'	a tárgy már nem fért bele a 4 pozícióba
7.	*spelen rol(OBJ) in*	*jouer rôle(OBJ) dans*	30	'szerepet játszik vmiben'	
8.	*bedoeld in:artikel*	*viser OBJ à:article*	30	'hivatkozik paragrafusban'	
9.	*doen beroep(OBJ) op*	*faire appel(OBJ) à*	29	'fellebbez vkihez'	
10.	*betreffen OBJ*	*agir se de*	27	'kb. illeti'	'agir se de' 3. megfelelője
11.	*zijn stad(SBJ) OBJ*	*être ville(SBJ) OBJ*	26	'a város vmilyen'	
12.	*vermelden in:artikel*	*viser OBJ à:article*	24	'említ paragrafusban'	
13.	*maken deel(OBJ) van*	*faire partie(OBJ) de*	24	'részét képezi vminek, tartozik vmihez'	
14.	*gaan over*	*agir se de*	24	'szó van vmiről'	'agir se de' 4. megfelelője
15.	*zien afbeelding(OBJ)*	*voir figurer(OBJ) de*	23	'lásd az ábrát'	
16.	*zijn van:toepassing op*	*appliquer se à*	22	'érvényes, vonatkozik vmire'	'appliquer se à' 1. megfelelője, aszimmetrikus
17.	*gelden voor*	*appliquer se à*	22	'érvényes, vonatkozik vmire'	'appliquer se à' 2. megfelelője, aszimmetrikus
18.	*nemen deel(OBJ) aan*	*participer à*	21	'részt vesz vmiben'	aszimmetrikus
19.	*richten zich tot*	*adresser se à*	19	'megcéloz, megszólít vkit'	
20.	*kennen voordeel(OBJ)*	*octroyer avantage(OBJ) de*	19	'megvan az előnye vminek'	
21.	*houden rekening(OBJ) met*	*prendre en*	19	'számításba vesz vmit'	ti. en:compte/considération
22.	*hebben betrekking(OBJ) op*	*concerner OBJ*	19	'vonatkozik vmire'	aszimmetrikus
23.	*zijn op:zoek naar*	*être à:recherche de*	18	'keres vmit'	
24.	*heten*	*appeler se OBJ*	18	'hívják vhogy'	
25.	*hebben effect(OBJ) op*	*avoir effet(OBJ) sur*	18	'(vmilyen) hatása van vmire'	
26.	*zijn in:België*	*être en:Belgique de*	17	'van Belgiumban'	
27.	*vergaderen*	*réunir se de*	17	'találkozót tart, összeül'	
28.	*zijn OBJ*	*etre OBJ à:foi*	16	'egyszerre van'	'à la fois' = ugyanakkor + holland határozószó
29.	*stoppen*	*arreter se de*	16	'befejeződik'	
30.	*liggen aan:basis van*	*etre à:base de*	16	'vminek az alapja'	
31.	*branden*	*allumer se de*	16	'ég (pl. lámpa)'	
32.	*bedragen euro(OBJ)*	*élever se à*	16	'(vmennyi euró) össszeget tesz ki'	aszimmetrikus, hiányzik a francia 'euro'
33.	*zijn OBJ*	*faire objet(OBJ) de*	15	'vmi alanya lesz' ???	
34.	*spelen rol(OBJ)*	*jouer rôle(OBJ) de*	15	'szerepet játszik'	'vmiben' nélküli változat

A fenti szempontok miatt 9 szerkezet egy másik szerkezettel egybeesett, ezeket kizártuk az értékelésből, sem helyesnek, sem helytelennek nem számítottuk. A megmaradó 58 szerkezetből a kiértékelés során 34 bizonyult helyesnek, ez *58,6 százalékos* pontosságot jelent. Ez természetesen jócskán elmarad az eredeti cikkben közölt, egynyelvű, magyar korpuszon mért 94 százalékos pontossági értéktől. Jelen feladat nyilvánvalóan jóval nehezebb: sokkal több elemet kell helyesen megtalálni, hogy a kapott párhuzamos szerkezet valóban teljes legyen. A 34 helyes vonzatos komplexige-szerkezetet az 1. táblázat tartalmazza.

5. Példák

A bevezető végén elővételeztük, hogy az algoritmusunk várhatóan leghasznosabb tulajdonsága az lesz, hogy olyan párhuzamos szerkezetek felfedezésére is képes, ahol a két nyelv teljesen más felépítésű szerkezetet használ az adott jelentés kifejezésére. Ezeket a párhuzamos szerkezeteket *aszimmetrikus*nak nevezzük. Gyengén vagy ,,tartalmilag" aszimmetrikus egy párhuzamos szerkezet, ha ugyanannyi szabad, illetve lexikálisan kötött bővítménye van, de a bővítmények nem az alapértelmezett módon felelnek meg egymásnak: tárgynak nem tárgy felel meg, vagy a kötött szavaknak, illetve a viszonyjelölőknek nem a szokásos fordítása szerepel. Erősen vagy ,,formailag" aszimmetrikus egy párhuzamos szerkezet, ha a bővítmények közvetlenül nem feleltethetők meg egymásnak, vagy a bővítmények száma nem is egyezik a két nyelvben. Az 1. táblázatban aszimmetrikusként megjelölt szerkezetek közül a legérdekesebb a következő három:

- A 18. sorszámú szerkezet klasszikus példája az egyszerű és komplex ige megfelelésének: a *részt vesz* fogalmát a holland nyelv a magyarhoz hasonlóan komplex igével (*nemen deel(OBJ)*) fejezi ki, a francia pedig egy szóval (*participer*).
- A 22. sorszámú szerkezet aszimmetriáját az (is) okozza, hogy a francia tárgy a hollandban nem tárgynak, hanem *op* elöljárós bővítménynek felel meg.
- A legbonyolultabb a 16. sorszámú szerkezet: itt a francia részen reflexív igével (*appliquer se*) találkozunk, a hollandban pedig egy létigés komplex igével (*zijn van:toepassing*).

Itt térhetünk ki annak a felvetődő kérdésnek a megválaszolására, hogy eredetileg miért nem úgy fogtunk neki a feladatnak, hogy külön-külön ellőállítottuk volna a holland és francia szerkezeteket, majd a két szerkezettárat illesztettük volna össze. A válasz az, hogy azért, mert így megkaphatjuk azokat a párhuzamos szerkezeteket is, amelyek két oldala formailag egyáltalán nem hasonlít egymásra.

Az eredmények jól mutatják az ismert tényt, hogy a különböző nyelvek egyes nyelvi elemei csak nagyjából felelnek meg egymásnak: sokszor van példa arra, hogy az egymás fordításának vélt szavak csak bizonyos környezetben fordításai egymásnak, vagy bizonyos környezetben nem fordításai egymásnak. Másképp fogalmazva a nyelvi elemek (például igék vagy elöljárók), a kifejezések különböző részhalmazait fedik le, és két nyelv viszonylatában ezek a részhalmazok szinte

soha nem esnek pontosan egybe, az átfedés mértéke széles határok között változik. Mikor egy párhuzamos szerkezetben egy tartalmas szónak nem a szokásos fordítása van jelen, máris egy gyengén aszimmetrikus szerkezettel van dolgunk.

A párhuzamos szerkezetek szépen megadják az igék egy-egy „jelentését", pontosabban azt, hogy adott környezetben, az adott elemek mellé éppen melyik ige illik. A szerkezet többi része sok esetben „szó szerinti" fordítás, és pontosan az ige az, amely kifejezésről kifejezésre más-más, nem kikövetkeztethető, megtanulandó, idiomatikus. Így van ez a 9. és a 13. szerkezet (l.: 1. táblázat) esetében, mikor a 'csinál' jelentésű francia *faire* az egyik kifejezésben a hasonló jelentésű holland *doen*-nal áll párban, máskor pedig a szintén hasonló jelentésű *maken*-nel, de nem felcserélhető módon.

Hasonlóan viselkednek az elöljárók is, gyakran kevéssé megjósolható módon. A nagyjából *-on/-en/-ön* vagy *-ra/re* szerepű elöljárók közül valamikor az *op–à* (16. szerkezet), máskor pedig az *aan–à* (18. szerkezet) áll párban, ugyanakkor az *op*-nak a *sur* is megfelelhet (25. szerkezet).

6. Összefoglalás

Az ismertetett módszer korpuszvezérelt módon, kétnyelvű igei szerkezetek hasznos gyűjteményét képes előállítani. Lényeges tulajdonsága, hogy képes felfedezni a formailag egymásra nem hasonlító, de egymásnak megfelelő, egymás fordításaiként kezelendő igei szerkezeteket. Nehéz feladat a párhuzamos szerkezetekben lévő számos elem mindegyikét megtalálni, ezért gyakran előfordul, hogy a kapott szerkezetek hiányosak. Ilyen esetben kézi utószerkesztéssel lehet javítani a hibákat.

A nyelvenkénti 3-3,5 millió szavas korpusz ilyen feladatra kicsinek számít, ezért viszonylag alacsony a kapott szerkezetek száma. A párhuzamos korpuszok előállítási költsége magas, ezért a közeljövőben maximum ennél egy nagyságrenddel nagyobb párhuzamos korpuszokra számíthatunk. Ezek használata azonban már jelentősen növelhetné a kinyerhető párhuzamos szerkezetek mennyiségét.

Amint a fentiekben láttuk, rendre egyszerű közelítő módszereket alkalmaztunk az előkészítő, elemző lépések során. Az e lépések során előforduló különféle hibáktól, hiányosságoktól függetlenül egyértelművé vált a módszer képessége az egymásnak megfelelő igei szerkezetek közvetlen megragadására. Az elemzési lépések fejlesztése nagy mértékben javíthatna a végső eredmény minőségén, de az a jelen dolgozatból így is látszik, hogy maga az algoritmus megfelel a kívánt célnak.

Ha egy párhuzamos szerkezet két oldalán 1-1 vonzati pozíció van, akkor azok a legtöbb esetben egymás megfelelői. Ritkán előfordulhat több ilyen pozíció (pl.: *őrizetbe vesz vkit vmi miatt*), ekkor egy külön módszerrel kell meghatározni, hogy melyik vonzati pozíció melyiknek felelhet meg, például az előforduló tartalmas szavak élősége/élettelensége alapján. Ennek kidolgozása a jövő feladata, mint ahogy a további, egyszerűbb típusokra kiterjedő kiértékelés elvégzése is.

Hivatkozások

1. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart (2005)
2. Siepmann, D.: Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. International Journal of Lexicography **18**(4) (2005) 409–444
3. Pecina, P.: A machine learning approach to multiword expression extraction. In: Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions, Marrakech, Morocco (2008) 54–57
4. Seretan, V.: Collocation extraction based on syntactic parsing. PhD thesis, University of Geneva (2008)
5. Evert, S., Krenn, B.: Methods for the qualitative evaluation of lexical association measures. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, Toulouse, France (2001)
6. Sass, B.: A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. In: Proceedings of RANLP 2009, Borovets, Bulgaria (2009) 399–403
7. Sass, B., Pajzs, J.: FDVC – creating a corpus-driven frequency dictionary of verb phrase constructions. In: eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Cahiers du CENTAL 7. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium (2010) 263–272
8. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007, Birmingham, United Kingdom (2007)

III. Szemantika

Vonzatkeretlisták helyett polaritásos hatásláncccsaládok – avagy a ℜeALIS σ függvénye

Alberti Gábor[1], Kilián Imre[2]

[1] PTE BTK Nyelvtudományi Tanszék
ℜeALIS Elméleti és Számítógépes Nyelvészeti Kutatócsoport
albi@btk.pte.hu
[2] ℜeALIS ESzNyK / PTE TTK Informatika Tanszék
mindkét cím: 7624 Pécs, Ifjúság útja 6.
kilian@gamma.ttk.pte.hu

Kivonat: A számítógépes fordításra [5] és más intelligens nyelvfeldolgozási feladatokra [6] irányuló kutatásaink során korábban *esetkeretlisták* formájában tároltuk az igék és más régensek legalapvetőbb vonzatszerkezeti tulajdonságaira vonatkozó információt. Mára megérett a lehetőség a ℜeALIS elmélet nyújtotta dinamikus diskurzus-szemantikai alapokon [10] arra, hogy elméleti nyelvészeti szempontból jóval igényesebb struktúrában (*polaritásos hatásláncccsaládok* formájában) rögzítsük a rokonítható vonzatszerkezet-változatokat [2], amiből sokkal több információ nyerhető ki a vonzathelyek betöltésére vonatkozóan [1]. Mindezt az „élethossziglani" ('lifelong') ℜeALIS keretbe ágyazva tudjuk elhelyezni, egy szövegelemző interpretáló információállapotának részeként [7]. A technológiai oldalon a Prolog nyelv alkalmazásának előnyeiről és mikéntjének fogásairól számolunk be.[1]

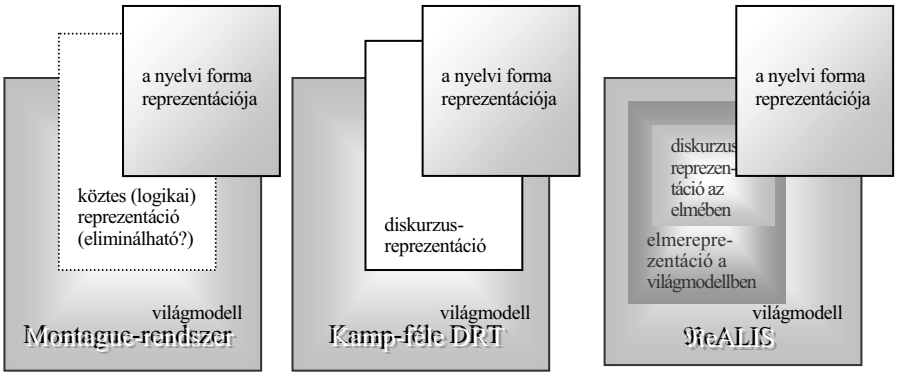
1 A ℜeALIS mint nyelvelméleti keret

A címben szereplő ℜeALIS elméletet a Montague-féle formális szemantika [11] és az abból kisarjadó, kisdiskurzusok jelentésábrázolására kidolgozott (Kamp-féle reprezentacionalista dinamikus szemantikai) DRT elmélet [12] továbbfejlesztéseként szoktuk bemutatni [6, 4].

Az alábbi 1. ábrán összevetjük a megemlített három jelentéselméletnek a nyelvi forma és a világmodell kapcsolatáról felállított modelljét. Ez alkalmat ad arra, hogy felvázoljuk a ℜeALIS kidolgozása melletti döntő metaelméleti érvet. Montague elméletileg igazolta, hogy nincsen szükség a nyelvi forma interpretációja során sem a korábban alkalmazott logikai reprezentációra, sem másfajta közvetítő reprezentációra; de e matematikai tételt nem tudta maradéktalanul átültetni a gyakorlatba. Ez követőinek sem sikerült, sőt egyenesen kialakult egy *reprezentacionalista* irányzat, mely a

[1] A szerzőket e cikk megírásában az OTKA T60595 sz. projektje támogatta. Értékes megjegyzéseikért elsősorban a ℜeALIS ESzNy Kutatócsoport további tagjainak szeretnénk köszönetet mondani: Kleiber Juditnak, Károly Mártonnak és Farkas Juditnak.

közvetítő reprezentáció kiküszöbölhetetlenségét mutatta be nyelvi adatokon, és funkciójára is magyarázatot talált [12]: a *diskurzusábrázolás* önálló szintjeként meghatározva azt. A ReALIS-modell képes feloldani a matematikai elmélet és a nyelvi gyakorlat közötti látszólagos ellentmondást egy további tényező, a kettős természetű *interpretálói elme* reprezentációjának a bevonásával.



1. ábra A nyelvi forma és a világmodell kapcsolatának háromféle formális modellje

Az önálló diskurzusreprezentációs szint úgy küszöbölhető ki, ha (nyelvészetileg releváns) információtartalmát beágyazzuk a világmodellbe, mégpedig a hangfolyamot strukturált diskurzusként felfogó interpretálói elme reprezentációjának részeként. Az interpretálói elmék pedig nyilván részei a világmodellnek, hiszen tartalmuk éppúgy beszédtárgyat jelent (ki mit tud, hisz, tervez?), mint mondjuk egy lakás részletei, egy állat felépítése, egy város struktúrája. Ami megkülönbözteti a reprezentálandó objektumok körében az elmét az utóbb említett dolgoktól, az az, hogy az elme kettős természetű a reprezentáció folyamatának szempontjából: nemcsak tárgya, de létrehozója is az ábrázolatoknak.

A ReALIS-modell ennek az alapgondolatnak a formalizált kidolgozását és nyelvészeti alkalmazását jelenti. A mintegy 40 oldalnyi definíciós állomány (http://lingua.btk.pte.hu/realispapers) ismertetésére nyilván ehelyütt nincs mód; most a nyelvtechnológiai relevancia mellett kell érvelnünk. A döntő általános érvünk az, hogy az intelligens nyelvtechnológiai feladatok (például számítógépes jelentésreprezentáció, illetve fordítás) végső soron a kommunikáló emberi elmék lehető legfinomabb modellezését kívánják meg. Idézünk néhány ezt szemléltető közismert példatípust egy másik cikkünkből [7] az alábbi (1) pontban.

Hogy *a miniszter*-re való visszautalás (1a) egy angol fordításban *he* vagy *she* legyen-e, annak eldöntése a szövegkontextus világában való tájékozottságot tesz szükségessé. A magyar szövegben pedig az elnökre utaló *az* névmás jól formáltsága az adott helyen a *diskurzus* szerkesztésének módján múlik: *topikváltás* történik a második mondatban [17].

Az (1b) azt szemlélteti, hogy egy szöveg jólformáltsága (és értelmezési lehetősége) múlhat olyan *ontológiai* ismereteken, mint a bernáthegyi és a kuvasz besorolása a kutyák kategóriájába, a papagájé pedig a madarakéba.

Az (1c) a Szegmentált DRT nevű irányzat [8] egyik iskolapéldája arra, hogy egy szöveg mondatai között (kötőszavak híján is) fel kell tárnunk a *retorikai* relációkat, melyek az *időstruktúrára* is befolyással lehetnek. A nyelvtechnológia szempontjából az most a megfigyelni való, hogy ha a második mondatban az első mondat *okát* ismerjük fel, akkor fordíthatjuk angolra Past Perfect idővel; *narratív* retorikai reláció feltételezése mellett viszont nem (ami az erőltetettebb eset). Az *ok* reláció felismerése viszont azt a világismeretet igényli, hogy az *elesésnek* a véletlen megbotlás mellett egy akaratlagos ellökés is lehet a kiváltója.

a. Megpillantottam az elnököt. ***Az*** viszont nem láthatott meg engem. (**2**)
b. Van egy bernáthegyim, egy kuvaszom és egy papagájom.
   b.1. ... ***A két kutya*** gyakran felbosszantja ***a szegény madarat***!
   b.2. ... ****A kutya*** gyakran felbosszantja ***a két szegény madarat***!
c. Péter elesett. Jancsi csúnyán meglökte.
d. A kalózvezér ***elásatta*** a rabolt kincseket.
   ... ***Az emberei*** napokig küszködtek ***a fagyos földdel***,
   de még így sem sikerült ***kellően mély gödröket*** csinálniuk ***a hitvány ásóikkal***.

Végül az (1d) diskurzus a cikk központi témáját jelentő *hatásláncсaládok* háttérismeretét szemlélteti, amit egyfajta kiterjesztett *lexikai tudás* részének tartunk. Az *elásat* igealak explicit módon csak egy Okozó szereplő (a kalózvezér) és egy Páciens jellegű szereplő (a kincsek) relációját mutatja be. A második mondat kézenfekvő értelmezése viszont feltételezi az alábbi hatáslánc felidézését (ami a jelentésábrázolás vagy -kivonatolás során is releváns): az Okozó *közvetlen* hatást (mondjuk egy kiadott parancs révén) az Ágensekre gyakorol (az embereire), akik Eszközeik (a hitvány ásók) segítségével (fagyos) földdarabokat (szintén Páciens?) ásnak ki eredeti helyükről (Kezdőpont?), egy vagy több gödröt létrehozva; a kincsek végül is idekerülnek az Okozó akaratából. Az említett szereptípusok a *thematikus szerepek elméletének* [15] az eszköztárából származnak, amelynek *absztrakt szerephierarchiákra* épülő továbbfejlesztett modelljét [2, 4] vesszük alapul e projektben.

Azt az álláspontot képviseljük tehát, hogy az intelligens nyelvtechnológiai feladatok olyan modellezését kívánják meg a kommunikációban álló emberi elméknek, ami hatalmas (és dinamikusan fejleszthető) kulturális/enciklopédikus, ontológiai és erősen kiterjesztett lexikai információs bázison alapul. A matematikailag formalizált ℜeALIS-modell (***Re**ciprocal **a**nd **L**ifelong **I**nterpretation **S**ystem*) éppen ezt kínálja, „kölcsönös” és „élethossziglani” jellegéből adódóan: a kölcsönösség a kommunikációs interakció megragadását biztosítja, az élethossziglaniság pedig hatalmas, folyamatosan bővíthető adatbázisokat jelent. Egyszerűen fogalmazva, a számítógépes jelentéskivonatolás és fordítás jövőjét abban látjuk, hogy a gép egykor úgy fog majd működni, mintha a megfigyelő/fordító (interpretáló) ember tevékenységét szimulálná.

A ℜeALIS interpretáló modelljének *szimultán rekurzív* definícióját röviden és informálisan úgy vázolhatjuk fel, mint egy élethossziglani folyamat leírását, amelynek során egy kezdetben (a születés idealizált pillanatában) strukturálatlan referenshalmazon négy reláció terjeszkedik ki, mintegy leképezve a környező világ hatásait. Az α *horgonyzófüggvény* az ugyanazon szereplőre mutató referenseket társítja egymással (az interpretáció során leginkább a régensekhez megtalált vonzatok és a határozott főnévi kifejezésekhez megtalált antecedensek alapján). Egy interpretálói elmében α

relációban áll például rengeteg olyan referens, ami W. A. Mozartra vonatkozó információ formulájában szerepel, a híres zeneszerzőre utalva. A λ *szintfüggvény* a klaszszikus szemantikából ismert konstans–változó megkülönböztetés általánosításaképpen fogható fel, e kételemű reláció helyett egy gazdag részbenrendezési struktúra osztályaiba („világocskákba") sorolva a külvilág entitásaira utaló referenseket és az interpretálói hiedelmek, vágyak, feltételezések, szándékok és tervek „fiktív" referenseit. A κ *kurzor* az interpretáló koncentrált figyelmét kívánja modellezni, pillanatonként változó módon kijelölve idő-, tér-, topik- és eseményreferenseket. Végül a σ *eventuális* függvény feladata „összerakni" a referensekből azokat a formulákat, amelyek megragadni hivatottak az eseményeket és állapotokat (azaz „eventualitásokat"). Ez teszi lehetővé a *robusztusságot*: külön-külön is információt adnak a referensek, nemcsak egyben létezik egy-egy propozícióformula, mint a korábbi elméleteknél; így a hiányos mondatokhoz is rendelhető valamilyen reprezentáció.

Az alábbi (2) eventuális formula például nagyjából egy olyasféle eseményt ($e_{\text{elásatás}}$) ír le a DRT „nyelvén" [12], hogy Long John Silver elásatott velem valamilyen kincseket egy szigeten egy bizonyos időpontban ($t_{2010\text{-}09\text{-}11}$).

$$e_{\text{elásatás}} : p_{\text{elásat}}\ t_{2010\text{-}09\text{-}11}\ r_{\text{LongJohnSilver}}\ r_{\text{kincsek}}\ r_{\text{én}}\ r_{\text{sziget}} \qquad \textbf{(3)}$$

Ezt a ℜeALIS (kétváltozós) σ függvénye úgy ragadja meg, hogy egy referenshez (pl. $e_{\text{elásatás}}$) (ami ezáltal *eventuális referensnek* osztályozódik) és egy szerepparaméterhez hozzárendel egy predikátumot (pl. $p_{\text{elásat}}$), egy időreferenst (t_x) és 0 vagy több argumentumreferenst (r_y). Ez idáig jóformán csak technikai újítás, a lényeg most jön: a szerepparaméter értékvektora választ néhány szereplőt egy *polaritásos hatáslánccsaládból*, beállítva az alábbi 1. táblázat bal oldalán felsorolt összes tényezőt a jobb oldali lehetőségek szerint. A következőkben bemutatjuk, hogy ezzel a megközelítéssel milyen széles körben meg tudjuk ragadni az ige (vagy más régens) körül kibontakozó mondat(rész) grammatikai jellemzőit, lehetőségeit, járulékos hangtani és jelentéstani tulajdonságait. Mindenekelőtt azonban megismerkedünk a polaritásos hatáslánccsaládokkal.

1. táblázat: Az argumentumszerepet meghatározó paraméterek és értékeik.

HATÓKÖRI SORREND	1, 2, 3, 4, 5, ...
HATÁSLÁNCCSALÁDBELI SZEREP(CÍMKE)	pl. Ág, Pác, Pác', Pác", Ok, Esz, Idő, Tér
ESETPROMINENCIA [2]	centrális (→), nem centr. vonzat (●), szabad hat. (○)
INFORMÁCIÓS SZEREP [14]	{T,K}*^{F,Q}*^(M)^C
REFERENCIALITÁSI FOKOZAT [1]	+hat > [-hat,+spec] > [-spec,+ref] > [-ref,+exp] > ∅
BESZÉDAKTUSBELI RÉSZVÉTEL (SAP) [6]	1Sg > 2Sg > ... > 3Pl

2 A polaritásos hatáslánccsaládok

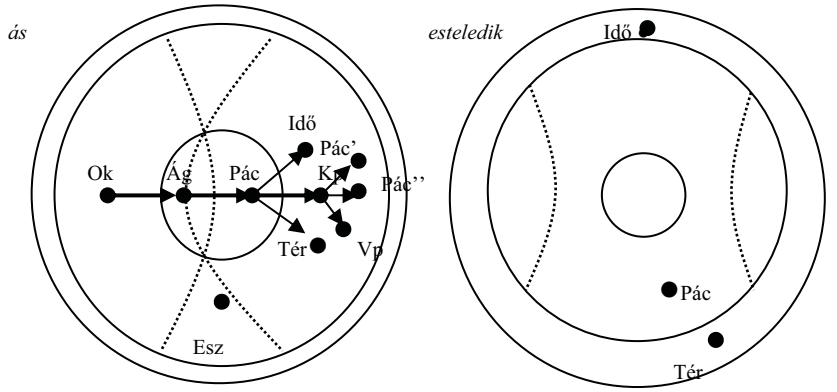
Az alábbi 2. táblázat thematikus szerepekkel megadott vonzatszerkezet-változatokat mutat, a termékeny *ás* és a nem igazán termékeny *esteledik* igető családjából.

Hogy a fenti 1. táblázatban *centrálisnak* nevezett alanyi és tárgyi grammatikai funkciót mely vonzatszerepekhez társíthatjuk, azt univerzális szabályok korlátozzák

[2, 3], az adott családra jellemző karakter rögzítésén túl, amit az alábbi 2. ábra mutat be vizuálisan. A lényeg az, hogy a pontsorral lehatárolt bal oldali körszeletben vannak a potenciális tranzitív alanyok („ágensi karakter"), a jobb oldaliban pedig a tárgyak („páciensi karakter") – ez adja a *polaritásos* jelleget; továbbá olyan ⟨Alany, Tárgy⟩ vonzatkeret nem lehetséges, amelyben a Tárgy felől mutatna a hatásirány nyila az Alany felé – ez ugyanis megsértené a hatáslánc elvét.

2. táblázat: Két magyar igető néhány vonzatkerete.

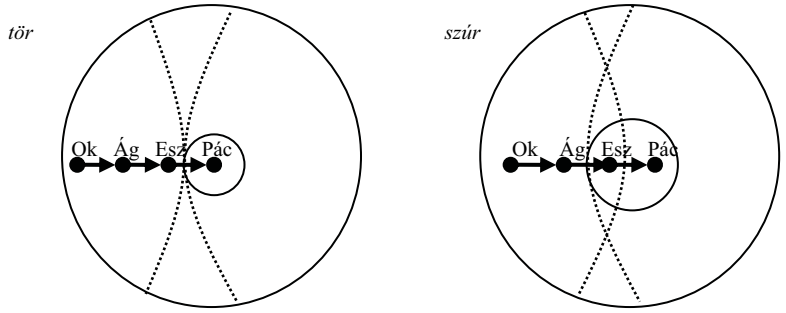
CENTRÁLIS ESETKERET	*esteledik / ás*
⟨ ⟩ (Páciens, Idő, Tér)	1. a. Esteledett. / b. Ránk esteledett (8-kor a tóparton).
⟨**Ágens**, Páciens⟩	2. **A kalózok** (kétségbeesetten) ásták az agyagos földet.
⟨**Ágens**, Tér/Idő⟩	3. a-b. **A kalózok** felásták a szigetet / végigásták a hétvégét.
⟨**Ágens**, Kezdőpont⟩	4. **A kalózok** ástak egy mély gödröt.
⟨**Ágens**, Végpont⟩	5. **A kalózok** ástak egy sírt.
⟨**Ágens**, Páciens' / Pác.''⟩	6. a-b. **A kalózok** elástak / kiástak egy értékes kincset.
⟨**Ágens**, Eszköz⟩	7. – (***A kalózok** kétségbeesetten ásták a hitvány ásóikat.)
⟨**Okozó**, Pác' / Pác'''/Pác ⟩	8. a-b. a-b. **A vezér** elásatta / kiásatta a kincset.
⟨**Okozó**, Ágens⟩	9. ??**A vezér** (álló nap) ásatta az embereit.



2. ábra. Két magyar igető (részleges) polaritásos hatásláncsaládja: *esteledik* és *ás*.

Az imént meghivatkozott cikkeinkben részletesen ismertetjük, hogy hogyan lehet néhány nyelvi adatból meghatározni egy-egy igető esetében a *polaritásos hatásláncsaládot*. Feltételezzük, hogy az „ideális interpretáló" is hasonló módon figyeli meg a hatásirányt, továbbá az ágensi/páciensi karaktert a vonzatszerepek esetében; ezek alapján aztán felépíti magában a polaritásos hatásláncsaládot, ami a továbbiakban meghatározza, hogy milyen vonzatszerkezet-változatokat ítél majd jól formáltnak (ellentmondó adatok esetén esetleg módosítva a hatásláncsalád struktúráját). A nyelvtechnológiai alkalmazás ezek után kézenfekvő, összhangban az interpretáló szimulálásának elvével: vonzatkeretlisták helyett polaritásos hatásláncsaládok formájában tároljuk a releváns lexikai információt.

Újdonság értékű megállapításunk [2, 3] az, hogy a lehetséges vonzatszerkezet-változatok nem (közvetlenül) a thematikus karakterből adódnak, hanem a *hatásláncirányon* kívül egyrészt a polaritásos hatáslánccsaládok *primitív magjából*, amit a 2-3. ábrán a legbelső kör jelöl, másrészt a (pontsorral jelölt ívek által lehatárolt bal / jobb oldali) *ágensi*, illetve *páciensi pólusból*. E négy tényező részben véletlenszerűen alakul ki a nyelv története során egy-egy igető esetében: ebből adódik az alábbi 3. táblázatban szemléltetett különbség a *tör* és a *szúr* lehetséges vonzatkeret-változatai között, miközben a vonzatszerepek thematikus karakterében erőltetett lenne bármi különbséget feltételezni (3. ábra). Arról van szó, hogy a *szúrás* Eszköze tranzitív alanyként és tárgyként egyaránt szerepelhet, míg a *törés* Eszköze tárgyként nem.



3. ábra. Két hasonló magyar igető (részleges) polaritásos hatáslánccsaládja: *tör* és *szúr*.

3. táblázat: A *tör* és a *szúr* vonzatkeret-változatai.

tör	CENTRÁLIS ESETKERET	*szúr*
Betört egy ablak. *primitív mag!*	⟨Páciens⟩	–
Ez a kalapács még vastagabb ablakot is betörne.	⟨**Eszköz**, Páciens⟩	**Egy szög** megszúrta a kezemet. *primitív mag!*
–	⟨**Ágens**, Eszköz⟩	**Péter** beleszúrt egy szöget az abroncsba.
Péter betörte az ablakot egy kalapáccsal.	⟨**Ágens**, Páciens⟩	**Péter** kiszúrta az abroncsot egy szöggel.
–	⟨**Okozó**, Eszköz⟩	**Mari** beleszúratott egy szöget az abroncsba.
Mari betörette az ablakot (Péterrel / egy kalapáccsal).	⟨**Okozó**, Páciens⟩	**Mari** kiszúratta az abroncsot (Péterrel / egy szöggel).
–	⟨**Okozó**, Ágens⟩	–

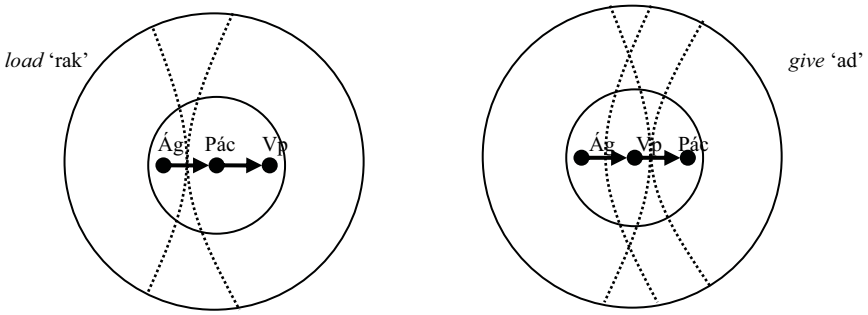
A ℜeALIS modelljében tehát a hagyományos thematikus szerepek [15] csupán címkeként szerepelnek, szemantikai karakterük a polaritásos hatáslánccsalád struktúrájában betöltött pozíciójuk révén definiáltatik.

Térjünk vissza az elméleti kitérő után a 2. ábrához! Az *esteledik* primitív magja üres, amint azt a T2.1.a. példamondat mutatja a fenti 2. táblázatban. Az Idő és a Tér szabad határozói (l. a külső gyűrűben!) csatlakozhatnak ehhez az igetőhöz, illetve vonzatként egy sajátos Páciens (T2.1.b.), nem alanyi vagy tárgyi (azaz centrális) funkcióban, hanem *-rA* ragos alakban. Az *esteledik* hatáslánccsaládja ennyi.

Az *ás*-é viszont annál gazdagabban burjánzik! A legszűkebb jelentés, a primitív mag földdarabok mozgatását írja le (T2.2.). E földmozgató tevékenység kumulálód*hat* alternatív módokon: behatottá téve egy földterületet (T2.3.a.), egy időszakot (T2.3.b.), vagy *gödröt* eredményezve a Kezdőpontból távozó földdarabok helyén (T2.4.). A gödörásásból kiindulva is többféle jelentésbővülést tapasztalhatunk, újabb tranzitív vonzatkeret-változatok formájában: a gödör szolgálhat sírként (T2.5.), vagy kincsek rejtekhelyeként – ez utóbbi esetben az *ásás* tevékenysége kétféle célt is szolgálhat (T2.6.a-b.). Az *ás* polaritásos hatásláncc*saládjában* a pontsorívek helyzete megmutatja, hogy az *ásás* Eszköze alanyként / tárgyként nem fejezhető ki (T2.7.), az Ágens tárgyi kifejezése pedig „határeset" (T2.9.). További jelentésbővülést egy Okozó bevonásával érhetünk el (T2.8.a-b.; l. még (1d) és (2)), ami mellett tárgyként bármely korábban szóba került tárgyat is felvehetünk (pl. *sírt ásat*).

A szakaszban tárgyalt vonzatelmélet univerzális. Ennek illusztrálása végett most (kommentárok nélkül) bemutatjuk az egytárgyas és a kéttárgyas angol igetípus tárgyalternációjának és passzivizálásának lehetőségeit, illetve az ezeket meghatározó polaritásos hatásláncc*saládokat*.

4. táblázat/ábra. Két angol igető (részleges) polaritásos hatásláncc*saládja*.



load 'rak'	CENTRÁLIS ESETKERET	*give* 'ad'
He loaded hay onto the wagon. ő rak-Past széna -rA a kocsi 'Szénát rakott a kocsira.'	⟨**Ágens**, Páciens⟩	**He** gave a book to Mary. ő adott egy könyv Mari 'Adott egy könyvet Marinak.'
He loaded the wagon with hay. ő rak-Past a kocsi -vAl széna 'Megrakta a kocsit szénával.'	⟨**Ágens**, Végpont⟩	–
–	⟨**Ágens**,**Végpont**,Páciens⟩	**He** gave **Mary** a book. ő adott Mari egy könyv 'Adott Marinak egy könyvet.'
The hay was loaded onto wagons a széna volt rak-PastP -rA kocsi-Pl 'A szénát kocsikra rakták.'	⟨Páciens⟩	The book was given to Mary. egy könyv volt ad-PastP -nAk Mari 'A könyvet odaadták Marinak.'
–	⟨**Végpont**, Páciens⟩	**Mary** was given a book. Mari volt ad-PastP egy könyv 'Marinak adtak egy könyvet.'
The wagon was loaded with hay. a kocsi volt rak-PastP -vAl széna 'A kocsit megrakták szénával.'	⟨Végpont⟩	–

3 A σ függvény grammatikai paramétervektora

Mint az 1. szakasz végén leszögeztük, a ℜeALIS modelljében egy esemény vagy állapot reprezentációját a σ eventuális függvény szervezi meg, hozzárendelve egy eventuális referenshez a predikátumot képviselő referenst és az argumentumokat képviselő referenseket (valamint egy időreferenst).

Ami a *predikátum* kijelölését illeti, arra az előnyös megközelítésre szeretnénk felhívni a figyelmet, hogy amennyiben az interpretáló például a fenti 2. táblázat valamely mondatában szembesül az *ás* igetővel, nem kell egyből döntenie, hogy melyik vonzatkeret-változatról van szó, hanem annyit kell megállapítania, hogy a 2. ábrán bemutatott bal oldali családot „hívja elő" az *ás* hangalak, rögzítve valamiféle alapjelentést (a potenciális bővítésekkel együtt). Az adott vonzatkeret-változat meghatározása azután egy specifikáló procedúrára van bízva.

E specifikáció az argumentumreferensek *paramétervektorának* beállításával folytatódik (ami technikailag a kétváltozós σ függvény egyik argumentumhelyére kerül, az eventuális referens mellé). Ennek egyik fontos eleme, amint ezt az 1. szakasz végén közölt 1. táblázat 2-3. sora is mutatja, *szerepcímkék* segítségével kiválasztani a polaritásos hatásláncsaládból a predikátummal társítandó argumentumokat, és specifikálni, hogy melyikük kapjon *centrális* funkciót, azaz alanyi / tárgyi megjelenítést.

Az alábbi 5.a-b. táblázat egyetlen vonzatkeret-választást szemléltet a számtalan megengedett változatból: az Okozónak és az elásandó kincsre utaló Páciens' címkéjű szereplőnek biztosítunk centrális grammatikai funkciót; a hatáslánc elvéből adódóan az már determinált, hogy az előbbi alanyként, az utóbbi tárgyként jelenítendő meg.

5. táblázat: Két argumentumkiosztás az *ás* predikátumreferens mellett.

a. *ás*	hatókör	szerepcímke	esetprom.	inf. szerep	ref. fokozat	SAP
+ref	1	Ok	→	T	+hat	3Sg
$(+ref)_F$	2	Tér	○	F	-hat, +ref	3Sg
+ref	3	Pác'	→	C	+hat	3Pl
+ref	4	Ág	●	C	+hat	1Sg

A 'kalózvezér egy ''lakatlan szigeten ásatta el velem a rabolt kincseket.
?A 'kalózvezér egy ''lakatlan szigeten ásatta velem el a rabolt kincseket.
?A 'kalózvezér egy ''lakatlan szigeten ásatta el a rabolt kincseket velem.
??A 'kalózvezér egy ''lakatlan szigeten ásatta velem a rabolt kincseket el.
??A 'kalózvezér egy ''lakatlan szigeten ásatta a rabolt kincseket velem el.
???A 'kalózvezér egy ''lakatlan szigeten ásatta a rabolt kincseket el velem.

b. *ás*	hatókör	szerepcímke	esetprom.	inf. szerep	ref. fokozat	SAP
$(+ref)_F$	1	Tér	○	F	+hat	3Sg
$(+ref)_F$	2	Ag	●		-hat, +ref	3Sg
+ref	3	Ok	→	C	+hat	3Sg
+ref	4	Pác'	→	C	+hat	3Sg

?''Madeirán ásatta el a kalózvezér a rabolt kincseket egy ''tússzal.
? ''Madeirán ásatta el egy ''tússzal a kalózvezér a rabolt kincseket.
?? ''Madeirán ásatta egy ''tússzal el a kalózvezér a rabolt kincseket.
??? ''Madeirán ásatta el a kalózvezér egy ''tússzal a rabolt kincseket.

Az 1. táblázat azt is megmutatja, hogy egy mondatba kerülő argumentum formai tulajdonságait számos más paraméter értékelése is befolyásolja, a thematikus szerepén és a grammatikai funkcióján kívül. Ezeket érdemes egyetlen paramétervektorban

egyesíteni, mivel ily módon minden nyelvészeti megszorítást e vektor (részben nyelvspecifikus, részben univerzális) jólformáltsági feltételeként fogalmazhatunk meg, illetve építhetünk be a gépi nyelvfeldolgozó rendszereinkbe. A fordítás problémája innen nézve annak kérdése, hogy egy adott paramétervektor egyaránt jól formált-e a forrás- és a célnyelvben; ha pedig nem, akkor a forrásnyelvi paramétervektorhoz milyen jól formált célnyelvi paramétervektor áll a legközelebb. A kérdéskör nyelvészeti oldaláról a [11] cikk nyújt részletes tájékoztatást; ehelyütt csak szemléltetjük a lehetőségeket az 5.a-b. táblázat alapján.

A magyarban az argumentumok *hatóköri* sorrendje független a thematikus szerepüktől és a grammatikai funkciójuktól. A T5.a. verzióban ilyen hatóköri sorrendet tekintünk: Okozó > Tér > Páciens' > Ágens; a T5.b. verzióban pedig ilyet: Tér > Ágens > Okozó > Páciens'.

Az *információs szerkezeti* szereposztás viszont erősen függ a hatóköri sorrendtől, bár a korábbi Topikfélék > Kvantorfélék > Fókusz sorrendet az újabb kutatások [14] „liberalizálták" (lényegében az 1. táblázat 4. sorában megadott módon). A T5.a. verzióban az Okozó topik funkciót kap, a Tér argumentuma pedig fókusz funkciót. A T5.b-ben egy ⟨Tér,Ágens⟩ tükörfókusz konstrukciót szemléltetünk. A 'C' ('komplementum') jelölés arra utal, hogy egy argumentum nem kapott semmilyen kitüntetett szerepet az információs szerkezetben. Az 'M' a generatív szakirodalomban sokat tanulmányozott igemódosítói / igevivői pozícióra utal. Az alábbi 6. táblázatban felidézzük a logikai jelentéstöbbletet adó információs szerepek jelentésrendszerét:

6. táblázat: A magyar operátorjelentések rendszere.
($R_n = R \setminus R_m$, ahol R_m: a megemlített szereplők, R: minden olyan szereplő, amely eljátszhatta volna a megemlített szereplők által eljátszott szerepet)

	P(x)	¬P(x)
$\exists x \in R_n$	Q_{is}: *is* kvantor *Meglep, hogy **a nővéremet is** meghívtad.*	K: kontrasztív topik ***A nővéremet** bezzeg meghívtad!*
$\forall x \in R_n$	Q_{mind}: *mind* kvantor *Meghívhattál volna **mindannyiunkat**!*	F: azonosító–kirekesztő fókusz *Bánt, hogy **csak Annát és Beát** hívtad meg.*

Az argumentumok *referencialitási fokozatát* korlátozhatja a thematikus szerep (l. [1]: pl. **Alakult az énekkar a klubban*), de a korlátokat semlegesítheti az információs szerep (pl. egy fókusz funkció: *A klubban alakult az énekkar*). Az ezzel kapcsolatos nyelvspecifikus, illetve univerzális tudás is beépíthető a σ függvény paramétervektorának jólformáltsági feltételeibe. Új ötlet, hogy a szám / személy értéket is vegyük figyelembe (SAP: 'speech act prominence'), mivel korrelációt mutat például a topikválasztással [10].

Ahogy az 5.a-b. táblázatok alatt megadtuk, a paramétervektor teljes specifikálását követően preferenciasorrendbe álló szórendi és hangsúlyozási változatok egész listája adódik. Gyakran több kifogástalan szórend is van, például abból adódóan, hogy a magyar kvantor nem feltétlenül áll az ige elé megmutatni a hatóköri elsőbbségét. A 6. táblázat megfelelő példamondata így is elhangozhatna: *Meglep, hogy meghívtad **a nővéremet is***.

4. Implementáció: egy megvalósítás sarokkövei

A ℜeALIS szintaxisfelbontó hozzáállása – a szemantikai relációk felépítését nem a szintaxistól függetlenül, hanem a kétirányú hozzáállást keverve, együtt, közösen végezzük, és a szemantikai információkat is a lexikon elemeihez kötjük –, komoly felkészültséget és gondos tervezést követel. Ilyen bonyolultsági fokozatot csakis valamiféle szimbolikus programozási nyelv, a „mesterséges intelligencia" programnyelvek valamelyikének választásával érhetünk el. Az előzetes tapasztalatok alapján minderre a Prolog programozási nyelv valamelyik dialektusa alkalmas.

Bemenet és kimenet. Egy program tervezésénél az első lépés: a bemenet és kimenet specifikációja, vagyis a program által kiértékelt ún. „fekete doboz" függvény értelmezési tartományának és értékkészletének a megadása. Ez kiindulópontul szolgál olyan értelemben is, hogy egy részleges megvalósítás (pl. többdimenziós halmazok esetén) ennek részhalmazát, vagy függvényösszetétel esetén a teljes feldolgozási függvény valamilyen összetevőjét számolja ki.

A ℜeALIS-megvalósítás célkitűzése a szöveg és a diskurzusreprezentáció közötti reláció kiszámítása. Ez (Prolog-szerű értelmezésben) mindkét irányú kapcsolatot jelenti. Ha a szöveg adott, akkor a program azt a reprezentációs kifejezést számítja ki, amely az adott logikai rendszerben és az interpretáló belső információállapotát leíró tudásbázisban (ontológiában) kiértékelhető, bizonyítható vagy hozzávehető a tudásbázishoz. Az ellenkező irányban: ha a tudáskezelő összetevő által (pl. egy kérdésre adott válaszként) egy logikai kifejezést kapunk, akkor végeredményben a szöveg képét állítja elő.

Nemdeterminizmus. Egyes helyzetekben nem eldönthető, hogy egy elemzési folyamatot milyen irányban érdemes folytatni, illetve jó okkal többféle irányban is folytatható lenne (pl. a *vár* szó előfordulása esetén, a homonímia miatt). Bizonyos szó vagy mondat elolvasásakor tehát még nem állapítható meg, hogy melyik értelmezés a helyes. A mondat vagy a diskurzus további elemeinek az elolvasása a helyzetet általában egyértelműsíti. Az ilyen helyzetek kezelésére egyértelműen a *nemdeterminisztikus* megoldások javasolhatók. A program eljárásai kiszámítják a relációk összes lehetséges értékét, amely értékhalmazból a hívó eljárások kiválasztják a számukra megfelelőket, és csak azok figyelembevételével állítják elő a saját eredményeiket.

Tárgygépmodell. A mi esetünkben a „tárgygépmodell" kifejezés talán tárgymodellként is felfogható, hiszen nem valamilyen konkrét processzorlapka regiszter- és utasításkészletéről van szó, hanem az elsőrendű logika valamelyik részosztályának használatáról és a nyelvi fogalmak elsőrendű logikába történő leképezéséről. Ez lehet maga a *Horn-klózok* részosztálya is, de előfordulhat az is, hogy ennél bővebb osztályt kell választanunk (pl. következményoldali diszjunkcióval, modális operátorokkal kiegészítve). Ha a Horn-klózok szintjén maradunk is, külön szerencsének kell felfognunk, ha az egész a tiszta Prolog következtetési mechanizmusával meghajtható – a feladat egyébként esetleg a Prologétól különböző rezolúciós stratégia alkalmazását is igényelheti [13].

Adatszerkezetek leírása elsőrendű logikában. Az elsőrendű logika alapvetően típustalan. Típusfogalom mégis alkalmazható úgy, hogy az alkalmazható függvénykifejezések körére megkötéseket teszünk. A Prolog-szakirodalomban a fogalom a vál-

tozók lekötési állapotát leíró ún. móddeklarációk kiterjesztéseként vált ismertté [16]. Bár típusos Prolog-megvalósítás létezik [9], teljes, tiszta Prologhoz használható adatszerkezet-megadó nyelvet eddig mégsem rögzítettek, és erre vonatkozó megvalósításról sem számol be senki. A következőkben intuitíven érthető és követhető példákat mutatunk be adatszerkezet-megadásra. A leképezés egyik buktatója a *másodrendű szerkezetek megvalósítása elsőrendű eszközökkel*. Erre két lehetőség adódik:

Reifikáció, vagyis a predikátumszimbólumot adatként kezeljük, így akár változó is felveheti értékül a futás során. Ennek feltétele, hogy a predikátumot ne kelljen predikátumként kiértékelni, vagy ha mégis, akkor a kiértékelés a tételbizonyítási folyamat jól megfogható helyzetében történjék, amikor a predikátumot logikán túli eszközökkel felépítjük, és azon túl már maga is részt vehet a tételbizonyítás folyamatában.

Logikán kívüli eszközök közvetlen alkalmazása: a Prolog-rendszerekben erre a célra szolgálnak a különböző dinamikuskifejezés-felépítő eljárások (pl. a `=../2`). A megoldás hátránya, hogy nem eléggé deklaratív: a lexikális tudás ábrázolása a befoglaló logikai rendszerre (Prolog) nézve specifikus eszközöket használ.

Azonosítás, referensek. *Referenseket* olyan változókkal jelölünk, amelyek kezdetben ismeretlenek, később a referens típusától függő értéket vehetnek fel. *Objektumreferensek* azok, amelyek a világ egy egyéniségét (individuumát, pl. Petőfi Sándor) címzik meg. Ezek a tudásbázis valamely osztályának példányai, az ilyen osztályokat a természetes tulajdonságértékeik nem azonosítják: az adatbázis (pl. telefonkönyv) egy másik Petőfi Sándora még akkor sem egyezik meg a költővel, ha egyébként a többi adatuk is véletlenül megegyezne. Az ettől eltérő adatszerkezeteket *adattípusnak* nevezzük, ezeket az értékeik egyértelműen azonosítják. Efféle típusok például az időreferensek és a címek.

Időreferensek használata időértékekkel lekötődő változóval lehetséges. Az időreferensek leírják a múltban, jelenben és jövőben történő eseményeket, tetszőleges pontossággal, időpontokat és időszakaszokat egyaránt. A `past(date(1-0-0))` időreferens például a ”tavaly ilyenkor” időmeghatározást jelenti nap pontossággal.

Predikátumreferensek az elsőrendű logikában csak *reifikációval* használhatók. A ℜeALIS predikátumreferensei a predikátummintát adják meg, amelyet a `^/2` kapcsolóval kapcsolunk a minta szabad változóihoz. Például az `AG^PAT^TIME^dig(AG,PAT,TIME)` predikátumreferens az „elás” háromváltozójú (ágens, páciens, időpont) alakját jelöli. Megjegyezzük, hogy a predikátumreferenseken keresztül kötjük a mondat logikai alakját az interpretáló belső állapotát jelképező ontológiához.

Az *eseményreferensek* a predikátummintákból az ismert logikai kapcsolók útján létrehozott kifejezéseket jelölik. A predikátumokból alkotott literálisok, illetve az ilyenekből logikai kapcsolókkal képzett kifejezések ebben a helyzetben adatként viselkednek, melyek pontos szerepe (ontológiabővítés, bizonyítás stb.) a programkörnyezetből derül ki.

Következtetéses és relációs tárgymodell. A *relációs szemléletmód* alatt azt értjük, amikor a felismerést (és a generálást is) egyetlen relációban valósítjuk meg. Ennek a relációnak az egyik paramétere a szóban forgó diskurzus belső reprezentációja, a másik paramétere a mondat szöveges alakja. Ez a szemléletmód a Prologot ismerők számára kézenfekvőbbnek tűnhet; hátránya az, hogy a mondatot szekvenciális fájl-

szerűen olvassa, amelyben a többszöri oda-vissza tekerések mindenképpen hatékonyságvesztést és az ábrázolástól idegen megközelítést jelentenek.

A következtetésre alapuló ábrázolásmód a felismerés feladatát hangsúlyozza, és párhuzamba állítja a Horn-klózokon alkalmazott következtetési folyamattal. Vagyis: a Horn-féle következtetési háló forrásai a tényállítások, amelyek a mi esetünkben a beolvasott diskurzus elemei (szavai, morfémái, betűi). Ezeken végzünk felismerési–következtetési műveleteket, amelyek valamilyen végkövetkeztetésben (célállításban) érnek véget. A végkövetkeztetés célszerűen egy olyan állítás lehet, amely paraméterként a teljes diskurzus szerkezetének valamilyen Prolog leképezését adja eredményül. A „lebegő" nyelvtani elemek és referensek, valamint azok összekapcsolódása a következtetéses megoldásban sokkal természetesebb.

A következtetési hálón célszerű rétegeket (vágatokat) értelmezni (I-IV.):

I. A szóalaktani elemzés szintje. Ez a beolvasott karaktersorozaton dolgozik, és ebből előállít egy szó- (morféma-) listát, valamint az egyes elemek diskurzuson, mondaton, szón belüli indexeit. A továbbiakban feltételezzük, hogy ez a lépés már problémamentesen működik, a szógyököket és szóösszetételeket – ahol kellett, azonosította, a hozzá kapcsolódó igekötő-, képző- és ragrendszer felépítését, így a nyelvtani eseteket és egyeztetéseket felderítette. Az elemzés eredményét az alábbihoz hasonló tényállításokban rögzítjük:

```
word(1,noun('Petra',proper,nom,sing-3)).
```

II. A szószerkezet-átmenet szintje. Ez gyakorlatilag a régens–vonzat, az alaptag-adjunktum, valamint az anafora–előzmény nyelvtani viszonyokat deríti fel, a lexikai leírásokban megjelenő *követel / kínál* viszonyokra alapozva.

III. Eventualitások meghatározása a polaritásos hatásláncmodell alapján, következtetésre alapuló ábrázolásban. Az egyes szófajok előfordulása implikálja a szemantikai predikátum fennállását.

```
  sigma(AG^PAT^TIME^INST^CAUS^stabWithInit(
                  AG,PAT,TIME,INST,CAUS)):-
  noun(CAUS,_,nom,AGR), noun(PAT,_,acc,_-3),
  verb(caus('szúr'),_PREFIX,decl,VTIME,AGR),
  noun(INST,common,inst,_-3),
time2verb(TIME,VTIME).
```

IV. Eventualitások meghatározása a hatásláncmodell alapján, relációs ábrázolásban. Az egyes szófajok előfordulása relációban áll a szemantikai predikátummal.

```
  sigma(AG^PAT^TIME^INST^CAUS^stabWithInit(
                  AG,PAT,TIME,INST,CAUS)),
  noun(CAUS,_,nom,AGR), noun(PAT,_,acc,_-3),
  verb(caus('szúr'),_PREFIX,decl,VTIME,AGR),
  [noun(INST,common,inst,_-3)]):-
time2verb(TIME,VTIME).
```

A fenti példában a „Mari kiszúratta az abroncsot egy szöggel" példamondat *szúr* igéjének predikátumkiszámító részletét láthatjuk, melyet az ontológia szintjén a

`stabWithInit/5` predikátummal ábrázolunk. Ez a műveltetéssel és eszközzel történő *kiszúrást* fejezi ki, és az általános `AG` (ágens), `PAT` (páciens), `TIME` (idő) paraméterek mellett még hivatkozik az `INST` (eszköz) és a `CAUS` (kezdeményező) paraméterekre is. A `time2verb` feltétel a `stabWithInit/5` predikátum általános időreferense és az igeidő közötti kapcsolatot számítja ki. A következtetéses modellben a `stabWithInit` predikátum az állítás következményrészében reifikálva, míg az egyes szóalkotóelemekre hivatkozás a feltételrészében szerepel. Ez utóbbiak a relációs modellben az állítás további paramétereiként láthatók.

A bemutatott lexikonrészletből létrejött program egyrészt meglepően egyszerű – lényegében a cikkben említett egyes példaigék (*ás, szúr*) hatásláncmodell-változatainak rögzítését tartalmazza. A tesztsorral való meghajtás eredménye egyfelől az elvárásokat hozta. Másfelől pedig rámutatott: további szemantikai vagy környezeti információk feldolgozása nélkül a helyzet nem egyértelmű, ami a nemdeterminisztikus eredmények túlburjánzásához vezet. Például a fentiekben ontológiai / szemantikai szelekciós információk nélkül nem dönthető el, hogy a *szög* a szúrás eszköze vagy ágense-e: a teszteredményekben – nagyon helyesen – mindkét értelmezést viszontláthatjuk. A két értelmezés között csak egy szemantikus információt is tartalmazó lexikonszerkezet tud egyértelműen választani: az ágensnek önálló cselekedetre képes (pl. az `Agent` fogalomból leszármaztatható) objektumnak kell lennie.

5 Összegzés, értékelés

Miután az 1. szakaszban bemutattuk a ℜeALIS dinamikus diskurzusszemantikai interpretációs modellt [4], majd a 2.-ban a polaritásos hatásláncsaládok lexikai-szemantikai elméletét [2, 3], a 3.-ban pedig az utóbbi beépítését az előbbibe a σ függvény paramétervektora révén, a 4. szakaszban egy implementáció sarokköveit vázoltuk fel, részeredményekről számolva be. Eddig alapelvek és adatformátumok rögzítése és tanulmányprogramok írása történt meg. A helyzetet egyrészt bonyolítja, másrészt megkönnyíti a kutatás korábbi szakaszában létrejött szóalak-elemzési program [5, 6] integrálásának igénye. Egy teljes ℜeALIS-megvalósításról egyelőre korai még a beszámoló, de ilyen irányba haladunk, és ez további közlemények tárgya lesz.

Bibliográfia

1. Alberti, G.: Restrictions on the Degree of Referentiality of Arguments in Hungarian Sentences. In: Acta Linguistica Hungarica Vol. 44 No. 3–4 (1997) 341–362
2. Alberti, G.: A szóképzéssel együttjáró vonzatszerkezet-változások rendszere. In: Nyelvtudományi Közlemények No. 103 (2006) 75–105
3. Alberti, G.: A szóképzéssel együttjáró vonzatszerkezet-változások egy polaritásérzékeny rendszere. In: Fancsaly É. (szerk.): Tanár és tanítvány. Írások Györke József és Hajdú Péter tiszteletére (2002-2007). Studia Linguistica, Dialóg Campus, Bp. – Pécs (2009) 122–145
4. Alberti, G.: ℜeALIS, avagy a szintaxis dekompozíciója. Általános Nyelvészeti Tanulmányok. Megjelenés előtt.

5. Alberti, G., Kleiber J.: The *GeLexi* MT Project. In: Hutchins, J. (szerk.): Proceedings of EAMT 2004 Workshop (Malta). Univ. of Malta, Valletta (2004) 1–10
6. Alberti, G., Kleiber J.: The Grammar of ReALIS and the Implementation of its Dynamic Interpretation. Informatica Vol. 34 No. 2 (2010) 103–110
7. Alberti, G., Károly, M., Kleiber, J.: The ℜeALIS Model of Human Interpreters and Its Application in Comp. Ling. In: Proc. ICSOFT 2010/2. SciTePress Portugal (2010) 468–474
8. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge Univ. Press (2003)
9. de Boer, T. W.: A Beginners' Guide to Visual Prolog. Prolog Development Center A/S, Kopenhagen, Denmark (2009)
http://download.pdc.dk/vip/72/books/deBoer/VisualPrologBeginners.pdf
10. Croft, W.: Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press (2001)
11. Dowty, D. R., Wall, R. E., Peters, S.: Introduction to Montague Semantics. D. Reidel Publishing Company, Dordrecht (1981)
12. van Eijck, J., Kamp, H.: Representing discourse in context. In: vBenthem, J., ter Meulen, A. (szerk.): Handbook of Logic and Language. Elsevier, Amsterdam (1997) 179–237
13. Kilián, I.: Horn clauses. A Two Way Street. Kézirat: Gyűrűfű Műhely Kft.. (korábban publikálva a www.sics.se honlapon) (1996)
14. Kiss, É. K.: Többszörös fókusz a magyar mondatszerkezetben. In: Büky, L., Maleczki, M. (szerk.): A mai magyar nyelv leírásának újabb módszerei II. (1995) 47–66
15. Komlósy, A.: Régensek és vonzatok. Kiefer, F. (szerk.): Strukturális magyar nyelvtan. I. Mondattan. Akadémiai Kiadó, Budapest (1992) 299–527
16. Nakashima, H.: Term Description: A Simple Powerful Extension to Prolog Data Structures. Electrotechnical Laboratory, Umezono, 1-1-4, Ibaraki, Japan (1985)
17. Pléh, Cs.: Topic and subject prominence in Hungarian. In Kiefer, F. (szerk.): Hungarian Linguistics (Linguistic and Literary Studies in Eastern Europe Vol. 4). John Benjamins, Amsterdam (1982)

Személynév-egyértelműsítés a magyar weben

Nagy T. István[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720, Szeged, Árpád tér 2.
nistvan@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: Ebben a cikkben bemutatjuk saját személynév-egyértelműsítő rendszerünket, amely képes egy adott névhez mint keresőkifejezéshez tartozó weboldalakból a különböző személyek és a hozzájuk tartozó honlapok azonosítására. Ezen megközelítés alapvetően az egyes személyekhez automatikusan felismert bibliográfiai jellemzők segítségével rendeli a különböző emberekhez az egy névhez tartozó honlapokat. Tehát a klaszterezés során nem használtuk fel az egyes weboldalak teljes tartalmát. Továbbá reprezentáljuk a magyar személynév-egyértelműsítő korpuszunkat is, melyen kiértékeltük rendszerünket. A kiértékelésre a BCubed metrikákat alkalmaztuk.

1 Bevezetés

Az internetfelhasználók egyik leggyakoribb tevékenysége személyek vagy hozzájuk kapcsolódó információk keresése az interneten. A keresőkben használt keresőkifejezések csaknem 30%-a tartalmaz valamilyen személynevet [1]. Viszont a nevek nagymértékben többértelműek: az amerikai népszámlálási hivatal szerint csaknem 100 millió emberhez alig 90.000 különböző név tartozik [2]. Ugyanez igaz hazánkban is, hiszen a 9 leggyakoribb családnév több mint négymillió ember családneve [3]. Ebből kifolyólag ezen keresőkifejezések eredményei az azonos nevű, de különböző személyekhez tartozó honlapokat tartalmaznak.

A személynevek egyértelműsítése több szempontból is kihívásokkal teli (speciális jelentés-egyértelműsítési) feladat. Egyrészt előfordulhat, hogy az egyes nevek többértelműek, több ezer embernek lehet azonos utó- és/vagy vezetékneve Másrészt bizonyos nevek rendkívül változékonyak, így előfordulhat, hogy egy személyhez tartozó nevet többféleképpen is leírhatunk.

Az egy adott névhez tartozó honlapok különböző személyek szerinti klaszterezésének feladatát a 2007-ben először megrendezésre kerülő Web People Search nyílt nemzetközi verseny tűzte ki céljául [4]. A rendszerek kiértékelése során a szervezők arra a következtetésre jutottak, hogy az egyes honlapok személyekhez való rendelése során igen hasznos jellemzőknek bizonyultak a személyekhez tartozó különböző bibliográfiai attribútumok [5]. Ebben a cikkben bemutatjuk a magyar személynév-

egyértelműsítő korpuszunkat[1] és egy olyan rendszert, amely alapvetően az egyes személyekhez automatikusan felismert bibliográfiai jellemzők segítségével rendeli a különböző emberekhez az egy személyhez tartozó honlapokat. Tehát a klaszterezés során nem használtuk fel az egyes weboldalak teljes tartalmát.

A feladat megoldása során 16 különböző jellemzőt azonosítottunk automatikusan úgymint: *családtag*, *mentor*, *egyéb név*, *iskola*, *díj*, *affiliáció*, *e-mail*, *telefonszám*, *fax*, *weboldal*, *születési dátum* és *hely*, *foglalkozás*, *diplomafokozat, nemzetiség.* Ekkor egy adott oldalt a kinyert jellemzők által leírt vektor reprezentált, melyben az egyes jellemzőket fontosságuk szerint súlyoztuk. Ezután definiáltunk egy hasonlósági mértéket, majd egy csoportba rendeltük a hasonló dokumentumokat.

Az angol és magyar nyelvű személynév-egyértelműsítő rendszerünk kiértékelése azt mutatja, hogy megközelítésünk eredményei szignifikánsan jobbak, mint a klasszikus dokumentumklasszifikációs megközelítéseké.

2 Kapcsolódó munkák

A webtartalom-bányászat célja az interneten elérhető szöveges dokumentumokból valamilyen szempont szerint hasznosnak vélt információk kinyerése. A fejlődés motorja a pénzügyi haszon, hiszen a kibányászhatatlannak vélt, vagy csak nagyon erőforrás-igényesen elérhető információk, összefüggések nagyon sokat érhetnek.

A kezdeti klasszikus webtartalom-bányászati próbálkozások 1998-'99 környékén jelentek meg [4, 5]. Ezek az alapvetően szabályalapú rendszerek vagy kézzel előállított szabályokon, vagy egy manuálisan annotált korpusz felügyelt tanulása során előálló szabályokon alapultak. A következő generációs megközelítések alapvetően gyengén felügyelt tanulási módszerek voltak. Ekkor a különböző rendszerek inputja egy lista volt célinformáció-párokkal. Ezen rendszerek célkitűzése, hogy összegyűjtsék azokat a párokat, amelyek kapcsolódnak egymáshoz. Ilyen párok lehettek például összefüggő entitások, mint ország – főváros [6], híres emberek és kapcsolataik [7], vagy entitás – attribútum párok, mint Nobel díjazottak – év [8]. Ezen rendszerek általában letöltötték azokat a honlapokat, amelyek tartalmazták az aktuális párokat, majd szintaktikai/szemantikai szabályokat tanultak azok mondataiból. Végül egy új weboldalkorpuszon alkalmazták az előzetesen megtanult mintákat, hogy új párokat nyerjenek ki. Ezen megközelítések alapvetően az internet redundanciáját használják ki. Azon a hipotézisen alapulnak, mely szerint az interneten a hasznos információk több helyen is elérhetőek, ezért néhány nagyon pontos szabály segítségével a szükséges információk kinyerhetővé válnak.

A második WePS kampány személynév-egyértelműsítési részfeladatán a beküldött rendszerek [5] többsége használt valamilyen előfeldolgozó lépést, mielőtt az egyes dokumentumokat reprezentálták volna. Majd valamilyen általános klaszterező algoritmust alkalmaztak. Ugyanakkor több csapat is úgy gondolta, hogy klaszterezés szempontjából igen sok információt tartalmazhat a különböző dokumentumokban található tulajdonnevek [5].

[1] A korpusz szabadon elérhető a Creative Commons licenc alatt.

3 Jellemzőalapú személynév-egyértelműsítés

Alapvető hipotézisünk az, hogy az egyes személyeket leíró biográfiai jellemzők hasznosak lehetnek a klaszterező algoritmus számára. Például ha két honlapról is kiderül az illető születési helye és dátuma, és ezek megegyeznek, akkor majdnem biztosak lehetünk abban, hogy ugyanarról a személyről van szó. Ezért 16 különböző jellemzőosztály definiáltunk, és próbáltuk meg ezen osztályokba tartozó jellemzőket automatikusan kinyerni az egyes weboldalakból. Minden egyes dokumentumot az ezen jellemzőkből alkotott vektortérmodell reprezentált. Végül ezt a teret klasztereztük, és azonosítottunk az egyes személyekhez tartozó weboldalakat.

A jellemzők kinyerése során alapvetően a HTML-oldalak szöveges részeire fókuszáltunk, mivel úgy találtunk [9], hogy több oldal tartalmaz releváns információt szöveges részben, mint strukturált formában.

3.1 Előfeldolgozás

A rendszerünk bemenetéül egy személynévhez tartozó a Yahoo! kereső által visszaadott weboldalak szolgáltak. Mivel úgy találtuk, hogy a weboldalakon található hasznos információ nagyrészt azok szöveges részében fordul elő, ezért alapvetően az egyes oldalak szöveges bekezdéseire koncentráltunk. Ezáltal a különböző nyelvfeldolgozó eszközök számára zajos és nehezen feldolgozható elemeket elhagytuk.

A weboldalakon előforduló bekezdések azonosításához a magyarlancot [12] alkalmaztuk minden oldal DOM fájának elemeire. Amennyiben az oldalon található szövegrészlet hosszabb volt 60 karakternél és több mint egy igét tartalmazott, akkor azt bekezdésnek jelöltük. Néhány jellemzőt a saját tulajdonnév-felismerő [13] rendszerünkkel azonosítottunk, amelyet a HVG korpuszon tanítottunk.

3.2 B-Cubed kiértékelési metrika

A klaszterezés kiértékelési metrikájául az B-Cubed mérték [10] kiterjesztett változatát használtuk, követve a WePS 3 verseny hivatalos kiértékelési útmutatóját. Ebben az esetben pontosságot és fedést számolunk, ugyanakkor szükséges a helyesség kiterjesztése azokban az esetben, amikor egy dokumentumot több klaszterbe is besorolunk. Ezért definiáltuk a többszörös pontosságot és fedést:

$$\textit{Többszörös fedés}(e,e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|L(e) \cap L(e')|}$$

$$\textit{Többszörös pontosság}(e,e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|}$$

Ebben az esetben e és e' két különböző elem, míg L(e) az e elemhez tartozó kategóriákat, C(e) pedig az e-hez tartozó klasztereket jelöli. Többszörös pontosságot csak abban az esetben használtunk, amennyiben e és e' klasztereket osztott meg, továbbá

többszörös fedést, amennyiben e és e' kategóriákat osztott meg. Előző értéke akkor volt maximális (1), amennyiben a megosztott kategóriák száma kevesebb vagy egyenlő volt, mint a megosztott klaszterek száma. Ugyanakkor értéke akkor volt minimális (0), ha a két elem nem osztott meg egy kategóriát sem. A többszörös fedés értéke akkor volt maximális, amikor a megosztott klaszterek száma kevesebb vagy egyenlő volt a megosztott kategóriák számával, ha pedig két elem nem osztott meg egy klasztert sem, minimális értéket eredményezett.

3.3 Jellemzőkinyerés

Néhány a WePS 2 versenyen résztvevő jellemzőkinyerő rendszerhez hasonlóan a saját megközelítésünk is két fontos lépésre épült: lehetségesjellemző-kinyerési és jellemzőverifikációs modulra. Tehát először kinyertük a lehetséges jellemzőket a bekezdésekből, majd kiválasztottuk ezek közül a végsőket.

A jellemzőkinyerés részprobléma megoldása során szükségesnek tűnt a jellemzőosztályok kategorizálása. Az azonos logikai osztályba tartozókat csoportosítottuk. Így például azonos csoportba kerültek az *egyéb név*, *családtag* és a *mentor*, hiszen ezek mind egyes személyek nevei. Ugyanakkor egy hierarchikus rendszert definiáltunk az összetartozó jellemzőkön belül. Így egy nevet csak akkor jelöltünk *mentornak*, amennyiben az nem volt sem *családnév* és *egyéb név* sem. A jellemzők különböző csoportosításai az első táblázatban láthatók.

1. táblázat: A jellemzők csoportosítása.

Név	Elérhetőség	Szervezetek
családtag	e-mail	iskola
egyébnév	weboldal	díj
mentor	telefonszáma	affiliáció
	fax	

A továbbiakban kitérünk az egyes jellemzők azonosításának részleteire.

Születési dátum: amennyiben a szótövesítés után egy bekezdés tartalmazta a *születik*, *születési dátum* stb kifejezések bármelyikét, akkor lehetséges dátumokat kerestünk ezen szavak környezetében. Ehhez egy dátumvalidátort alkalmaztunk, amely 9 különböző reguláris kifejezés segítségével próbálja azonosítani a különböző formában megadott dátumokat.

Születési hely: amikor egy adott bekezdés szótövesített változata a *születik*, *szület*, *születni*, *szülőváros* kifejezések bármelyikét tartalmazta, akkor alkalmaztuk a saját, HVG korpusz földrajzi név osztályon tanított tulajdonnév-felismerő rendszerünket, hogy azonosítsuk a lehetséges szülőhelyeket. Végül egy frázist születési helynek jelöltünk, amennyiben azt a szülőhely-validátorunk elfogadta.

Szervezetek (*iskola*, *díj*, *affiliáció*): mivel úgy találtuk, hogy ezen jellemzők egyes szervezetek nevei, ezért ezeket egy csoportba soroltuk. Ugyancsak saját tulajdonnévfelismerő eszközünket alkalmaztuk, amely ebben az esetben a HVG korpusz szervezet osztályán lett tanítva. Amennyiben a tulajdonnév-felismerő rendszerünk talált egy

szervezetet a bekezdésekben, akkor azt először az iskola jellemző szempontjából vizsgáltuk. Amennyiben a kinyert szervezetnév megadott környékén előfordult valamely kulcskifejezés, mint például *diploma*, *oktatás*, *tudomány*, akkor azt lehetséges *iskolának* jelöltük. Egy ilyen kifejezést akkor fogadtunk véglegesen el, amennyiben az iskolavalidátorunk azt elfogadta. Ez abban az esetben történt meg, amennyiben az adott kifejezés minden szava nagybetűvel kezdődött, néhány kötőszót kivéve, mint például az *és,* továbbá tartalmaz néhány kulcskifejezést, úgymint *Iskola*, *Akadémia*, *Egyetem*, *Főiskola* stb. Amennyiben az adott kifejezést elvetettük, a továbbiakban *díj* jellemző szempontjából vizsgálódtunk. Így, amennyiben az aktuális kifejezés olyan kifejezések mellett fordul elő, mint *díj*, *nyer*, *év* stb. akkor azt potenciális *díj* attribútumként kezeltük. Egy ilyen kifejezést csak abban az esetben jelöltünk *díj* jellemzőnek, amennyiben azt egy általunk definiált díjvalidátor elfogadott. Ez akkor történt meg, ha az aktuális frázis minden szava nagybetűvel kezdődött, kivéve néhány kötőszót, úgymint az *és,* továbbá olyan kifejezéseket tartalmaz mint *díj*, *legjobb*, *játékos* stb. Amennyiben a potenciális szervezetnevet sem iskola, sem *díj* jellemzőként nem sikerült azonosítanunk, akkor azt végül *affiliáció*ként jelöltük.

Nevek (*családtag*, *egyéb név*, *mentor*): mivel ezen jellemzők mind valamilyen személyek nevei, ezért ezeket egy csoportba rendeltük. A név típusú attribútumok azonosítására szinten a saját fejlesztésű tulajdonnév-felismerő rendszerünket alkalmaztuk, ám ezúttal a HVG korpusz név osztálycímkéjén tanított modellt alkalmaztuk. A modell által kinyert személynévelemet *családtagnak* jelöltünk, amennyiben az valamilyen rokonságot kifejező szó környezetében fordult elő, mint például, *fia*, *apja* stb. (ezen kifejezések listáját a Wikipédia rokonság[2] szócikkéből gyűjtöttük). Azonban sok esetben ez a feltétel nem teljesült, így ekkor az aktuális potenciális nevet lehetséges *egyébnévként* kezeltük. Úgy gondoltuk, hogy egy adott személy nem ad meg egy másik nevet azonos számú szóval, (ez a megállapítás nem feltétlenül igaz becenevek esetén) ugyanakkor az *egyébnév* mindenképp tartalmazza az eredeti név legalább egy részét. Tehát ha az aktuális név Kovács István volt, Kovács Józsefet nem fogadtuk el egyébnévnek, míg Kovács T. Istvánt igen. Amennyiben egy nevet nem jelöltünk sem *családtag*, sem *egyébnévnek*, akkor azt végül a *mentor* jellemzőosztály szempontjából vizsgáltuk. Abban az esetben, ha az aktuális név néhány kulcskifejezés környékén fordult el, úgymint *edző*, *mentor* stb. akkor azt végül *mentor* osztályba soroltuk.

Titulus: manuálisan létrehoztunk egy 60 elemből álló listát, amely különböző tudományos fokozatokat, diplomákat tartalmaz. Amennyiben az aktuális név adott közelségében a lista egy elemét találtuk, akkor azt *titulus* jellemzőnek jelöltük.

Nemzetiség: összeállítottunk egy 371 elemből álló listát, mely különböző nemzeteket tartalmaz. Ekkor minden nemzetiséget megpróbáltunk kinyerni az oldalról, végül a leggyakoribbat jelöltük nemzetiség attribútumnak.

Amikor az elérhetőség jellemzőket próbáltuk azonosítani, akkor nem csak az oldalak bekezdéseit vizsgáltuk, hanem az egész oldalt, ugyanis úgy találtuk, hogy ezen típusú jellemzők bárhol előfordulhatnak a weboldalakon.

[2] http://hu.wikipedia.org/wiki/Rokonság

Telefonszám: amennyiben egy szövegrészlet tartalmazta a *tel*, *telefon*, stb. kifejezések egyikét, akkor a következő igen megengedő reguláris kifejezéssel kerestünk lehetséges telefonszámokat:

```
(((?[0-9+(][.()0-9s/-]{4,}[0-9])(? d{1,5})?)
```

Amennyiben volt találat, akkor egy általunk definiált validátor segítségével választottuk ki a telefonszámokat.

Fax: a telefonszámhoz hasonlóan jártunk el, mivel a két jellemző meglehetősen hasonló. Ugyanakkor ebben az esetben a *fax* szó környékén vizsgálódtunk.

E-mail: Úgy gondoltuk, hogy ha valaki közzéteszi az e-mail címét, az az esetek többségében egyben link is. Ezért elsősorban az olyan linkeket vizsgáltuk, amelyek a *mailto* tagot tartalmazták. Ezenkívül igen gyakori, hogy az e-mail cím tartalmazza a személy nevét, vagy annak egy darabját. Definiáltunk egy e-mailcím-validátort, amely abban az esetben fogadott egy e-mail címet, ha az tartalmazta a személy nevének karaktertrigramjainak valamelyikét. Ugyanakkor definiáltunk egy stoplistát is, amely olyan szavakat tartalmazott, mint például *webmaster*, *wiki*, *support* stb. Ezenkívül minden elfogadott e-mail címből kinyertük a *domain* címet, amit később az *internetcím* jellemzőnél használtuk fel.

Internetcím: Úgy találtuk, hogy a személyekhez köthető internetoldalak címe tartalmazzák az adott személynevet, vagy annak egy darabját. Ugyanakkor ezen attribútumok is jellemzően link formában fordulnak elő az egyes weboldalakon. Tehát az internetcím-validátorunk az olyan webcímeket fogadta el, amelyek tartalmazták az adott nevet, vagy annak egy részét, esetleg az e-mail címből kinyert domént.

3.4 Weboldalak klaszterezése

Az egy személyhez tartozó honlapok klaszterezése során úgy gondoltuk, hogy csupán a személyes információk alapján képesek vagyok a különböző emberekhez tartozó dokumentumokat osztályozni. Továbbá képesek vagyunk megállapítani, egy adott névhez tartozó dokumentumok hány különböző személyhez tartoznak. Ez az adat a klaszterezés szempontjából különösen fontos, hiszen a klaszterező algoritmusok többségéhez szükséges előre definiálni a klaszterek számát.

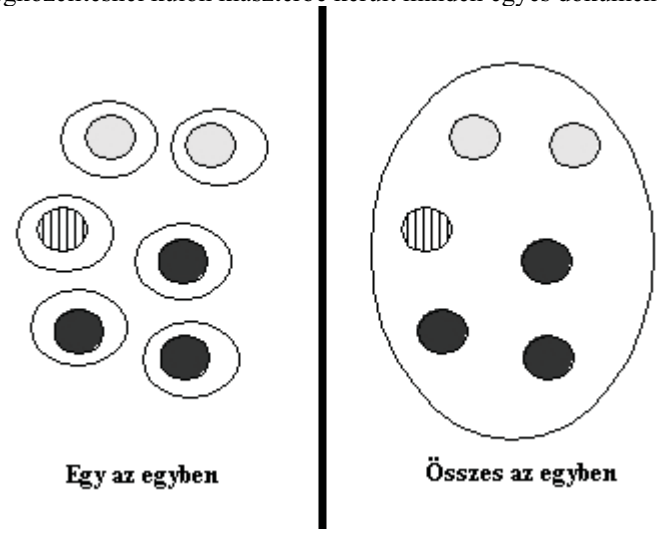
Az egyes jellemzőkre egy súlyozást definiáltunk. A leghasznosabb attribútumoknak az *internetcím*, *e-mail cím*, *telefonszám*, *fax* és az *egyéb név* bizonyult, ezért ezek 3 súlyt kaptak. Továbbá a *születési dátum* 2-es, míg a *születési hely*, *mentor*, *affiliáció*, *nemzetiség*, *családtag*, *iskola* és *díj* 1-es értéket kaptak. Ekkor minden dokumentumot a kinyert jellemzőkből álló vektor reprezentált. Ahhoz, hogy egy hatékony hasonlósági metrikát tudjunk meghatározni, előbb szükséges volt az egyes jellemzők normalizálása, egységes formára hozása. Ezért különböző szabályok és reguláris kifejezések segítségével egységesítettük azokat.

Az alapvetően weboldalakról kinyert személyes jellemzők segítségével végzett klaszterezés során egy alulról fölfelé történő heurisztikát alkalmaztunk. Ebben az esetben először minden dokumentum egy külön klaszterben van, majd a különböző klasztereket addig vonjuk össze iteratívan, amíg a megállási feltételt el nem érjük. Minden lépésben a leghasonlóbb klaszterek kerülnek összevonásra, ahol minden klaszter a centroidjával van reprezentálva, és két centroid közti távolság az őket leíró normalizált, súlyozott vektorok euklideszi távolsága. Az algoritmus számára a megál-

lási feltétel a legnagyobb hasonlósági mérték kevesebb mint 3-as mivolta. Tehát két klasztert abban az esetben nem vontunk össze, amennyiben a kettő közti hasonlósági mérték kisebb volt 3-nál.

Az alapvetően attribútumokat használó megoldás mellett, néhány alapmódszert is kipróbáltunk. Ekkor az egyes nevekhez tartozó dokumentumhalmazokat, a különböző dokumentumokból létrejövő vektortérmodell reprezentált. Ehhez a WEKA Java csomagban [11] található KMeans algoritmust is alkalmaztuk. Ugyanakkor ezen megközelítésnek, mint a klaszterező algoritmusok többségének, szükséges előre definiálni a klaszterek számát. Mivel az adott feladat során ez az érték nem ismert, ezért különböző heurisztikák segítségével próbáltuk meg megbecsülni azt. Az első esetben [Kmeans], az előzőekben már bemutatott, alulról fölfelé történő, jellemzőkön alapuló megközelítés által végeredményül kapott klaszterszámot adtuk meg. Másik esetben [Simple], a kiértékelő korpuszon, az egy névhez tartozó átlagos személyek számát (hét) adtuk meg. Végül a [Perfekt] esetben az annotátorok által meghatározott, adott névhez tartozó személyek számát kapta meg a KMeans algoritmus.

Ezen kívül még két egyszerű alapmegközelítést is adtunk, melyeket az 1. ábrán láthatunk. Az első esetben minden dokumentumot egy klaszterbe tettünk, míg az egy az egyben megközelítésnél külön klaszterbe került minden egyes dokumentum.



1. ábra. A két alapmegközelítés.

4 Kiértékelés

4.1 Korpusz

Rendszerünk kiértékelésére létrehoztunk egy magyar nevekhez tartozó weboldalkorpuszt, manuálisan annotált honlapokkal, amely elérhető a http://www.inf.u-

szeged.hu/rgai/nlp/homepagewsd weboldalon. Hogy eredményeink összevethetőek legyenek más nemzetközi eredményekkel, a tesztkorpuszt a meglévő korpuszokhoz hasonlóan hoztuk létre. A nevek közé több közéleti szereplő került, úgymint Csányi Sándor (OTP vezér és színész), továbbá Magyarországon igen gyakori nevek, mint például a Kovács István vagy a Szabó Zsófia. Ugyanakkor arra törekedtünk, hogy ezen gyakori nevek közt is szerepeljenek híres személyiségek, ahogy az első esetben a bokszoló, míg a másodikban a színésznő. Továbbá Schmitt Pál egy igazán érdekes kihívásnak ígérkezett, hiszen az élet különböző területein tölt be fontos pozíciókat, így a hozzá kapcsolódó weboldalak is igen eltérőek lehetnek.

A dokumentumhalmazban minden névhez a Yahoo![3] kereső által megadott első 100 találat került letöltésre, így a korpusz végül 960 weboldalt tartalmazott. Ezek közül összesen 572 oldalt kötöttek az annotátorok egy adott személyhez, vagyis egy névhez átlagosan 57 oldal kapcsolódott. Ugyanakkor a különböző nevek esetén igen nagy eltérések vannak, hiszen míg Zrínyi Miklós esetében a találatok nagy többsége valamilyen intézményhez köthető, addig például Schmitt Pál esetében az oldalak többsége a konkrét személyhez tartozik. A 10 névhez összesen 120 különböző személyt azonosítottunk, de míg a Kovács István esetében 30 különböző egyén fordult elő, addig a Schmitt Pálhoz tartozó weboldalak alapvetően a köztársasági elnökhöz voltak köthetők.

4.2 Eredmények

A különböző megközelítések eredményei a második táblázatban láthatóak. Az algoritmusokat B-Cubed pontosság, fedés és az ezekből számított F-mértékkel értékeltük ki. A táblázatból kitűnik, hogy az általunk megadott algoritmus érte el a legjobb eredményt az adott korpuszon. Míg a klaszterező eljárások közül az érte el a legjobb eredményt, amikor megadtuk a klaszterek pontos számát. A másik két eljárás más-más pontosság és fedés mellett ért el azonos F-mértéket.

2. táblázat: Eredmények.

Megközelítés	BCubed pontosság	BCubed fedés	F-mérték
Jellemzők	0,59	0,64	0,59
All_In_One	0,43	0,84	0,50
Perfekt	0,59	0,37	0,43
Simple	0,52	0,38	0,36
kMeans	0,69	0,28	0,36
One_In_One	0,93	0,24	0,35

5 Konklúzió

Ebben a cikkben bemutattunk az első magyar nyelvű személynév-egyértelműsítő megközelítésünket, amely hatékonyan volt képes kezelni a problémát. A kiértékelés-

[3] www.yahoo.com

hez létrehoztuk az első magyar nyelvű személynév-egyértelműsítő korpuszt. Rendszerünk a weboldalak folyó szöveges részéből dolgozik és alapvetően a személyek bibliográfiai attribútumai alapján egyértelműsíti a személyneveket. 16 különböző attribútumosztályt definiáltunk, amelyeket automatikus eszközökkel nyerünk ki. A klaszterezés 0,59 B-cubed F-mértéket ér el, ami az angol nyelvre publikált korpuszokkal és algoritmusokkal összevethető eredmény.

Habár eredményeink jónak tekinthetőek, a rendszernek számos továbbfejlesztési iránya van, amelyeket a jövőben meg kívánunk valósítani. Ilyenek például a táblázatos részekből kinyerhető információkkal való kiegészítés, mélyebb szintaktikai információk figyelembevétele a validátoroknál, illetve a bekezdések fő alanyainak azonosítása (a hibák egy része annak a hipotézisnek a következménye, hogy az egy oldalon talált minden bibliográfiai adat az oldal tulajdonosáé).

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Bibliográfia

1. Ide, N., Veronis, J.: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics Vol. 24 No. 1 (1998) 1–40
2. Guha, V., Garg, A.: Disambiguating People in Search. In: Proceedings of the 13th World Wide Web Conference (WWW 2004). ACM Press (2004)
3. A leggyakoribb magyar családnevek: http://www.chem.elte.hu/departments/elmkem/baranyai/nevek.htm
4. Sekine, S., Artiles, J.: WePS 2 Evaluation Campaign: Overview of the Web People Search Attribute Extraction Task. In: 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference (2009)
5. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics, Prague, Czech Republic (2007) 64–69
6. Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderl, S., Weld, D. S., Yates, E. Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence Vol. 165 (2005) 91–134
7. Cheng, X., Adolphs, P., Xu, F., Uszkoreit, H., Li, H.: Gossip galore – a selflearning agent for exchanging pop trivia. In: Proceedings of the Demonstrations Session at EACL 2009. Association for Computational Linguistics, Athens, Greece (2009) 13–16
8. Li, H., Xu, F., Uszkoreit, H.: A seeddriven bottom-up machine learning framework for 8 extracting relations of various complexity. In: Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic (2007)
9. Nagy, I., Farkas, R., Jelasity, M.: Researcher affiliation extraction from homepages. In: Proceedings of the NLPIR4DL ACL Workshop (2009) 1–9

10. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 17th international conference on computational linguistics. ACL (1998)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Vol. 11, No. 1(2009)
12. Zsibrita J., Nagy I., Farkas R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 394–395
13. Szarvas, Gy., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: The Ninth International Conference on Discovery Science 2006. LNAI 4265 (2006) 267–278

A Magyar WordNet felhasználhatósága lexikális jelentés-egyértelműsítésben*

Kuti Judit[1], Darja Fišer[2]

[1]MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
1068, Bp., Benczúr u. 33.
kutij@nytud.hu
[2]University of Ljubljana, Faculty of Arts, Department of Translation Studies
Aškerčeva 2, SI - 1000 Ljubljana
darja.fiser@guest.arnes.si

Kivonat: Tanulmányunkban bemutatunk egy vizsgálatot, amellyel a magyar WordNet használhatóságát teszteltük a gépi fordítás során alkalmazható lexikális jelentés-egyértelműsítésben. A vizsgálat során lefordítottuk angolra egy magyar szöveg tartalmas szavait a MetaMorpho gépi fordítórendszerrel, valamint a Magyar WordNettel való jelentés-egyértelműsítésen keresztül megfeleltettük ezeket a Princeton WordNet synseteinek. Egy angol nyelvű referenciafordításhoz képest automatikusan értékeltük ki a kapott fordításokat. A MetaMorpho gépi fordítórendszer magyar–angol nyelvpárra jelenlegi állapotában jobb fordításokat ad, mint a WordNet által javasolt lexikális fordítások; teljesítményét tehát a jelen vizsgálat alapján úgy tűnik, a HuWN nem javítaná.

1 Bevezető

A jelentés-egyértelműsítés mint a nyelvtechnológia egyik központi feladata számos alkalmazásban kap fontos szerepet: a legfontosabbak ezek közül a gépi fordítás, az információkivonatolás, illetve az információkinyerés. A feladat, komplexitásából adódóan, egyelőre nem tekinthető megoldottnak – sem magyarra, sem más nyelvekre (l. [1]). Általánosan a jelentés-egyértelműsítés folyamatát két alapvető lépésre bontjuk: (i) valamilyen jelentéstár kiválasztása, illetve létrehozása, valamint (ii) a jelentéstárban szereplő jelentések hozzárendelése a kívánt szóalakokhoz valamilyen algoritmus segítségével. Nemzetközi szinten az egyik legelterjedtebb jelentéstár a WordNet (PWN – l. [5]).[1] A WordNet mint általános adatbázistípus az angol nyelvű Princeton

* Jelen tanulmány a "A Magyar és Szlovén WordNet összehasonlító kiértékelése gépi fordításban" c. projekt keretén belül született, melyet a Magyar-Szlovén Kormányközi TéT együttműködés támogat 2009-2010-ben.

[1] A legkülönbözőbb jelentés-egyértelműsítő részfeladatokhoz (célszó jelentés-egyértelműsítése, automatikus kulcsszókinyerés, szemantikai szerepek címkézése stb.) nagy százalékban ennek az adatbázisnak különböző verzióit használják mint jelentéstárat. A Senseval versenyeken használt jelentés-egyértelműsített korpuszok több mint fele valamilyen WordNet-típussal lett annotálva.

WordNetre épülő lexikális hálót takar, amelynek alapegysége a fogalom / jelentés (szakszóval *synset*), nem pedig a tradicionális szótárak alapegysége, a szó / lexéma. A wordnetek egy adott nyelv lexikalizálódott jelentéseit az egymáshoz való jelentéstani viszonyaik által alkotott hálóban helyezik el, a viszonyokat a háló éleiként, a jelentéseket ezen élek találkozási pontjaiként, csomópontjaiként szemléltetve.

A WordNetek közismert gyengesége a poliszém szavak jelentéseinek túl finom megkülönböztetése, ami a gyakorlatban nagyon megnehezíti egy átlagos beszélő számára egy szóalak valamely WordNet-beli fogalomnak való megfeleltetését. A WordNet-beli fogalmak elkülönítése utólag már gyakran nem tűnik motiváltnak, a jelentésegységek sokszor átfedésben vannak egymással. Párhuzamos WordNetekben – mint pl. a HuWN és a PWN, ahol a magyar és angol nyelvű fogalmak egyedi azonosítójukon keresztül meg vannak feleltetve egymásnak – gyakori az a jelenség, hogy egy mindkét nyelven poliszém szó más-más aljelentéssel több párhuzamosított synsetet is "elfoglal" önmagában, anélkül, hogy a jelentések közötti különbségtétel oka nyilvánvaló volna. Így látszólag mindkét WordNetben duplázva, triplázva szerepel egy adott szót tartalmazó synset – fordítási szempontból mindenképp redundáns módon. Az alábbiak jól példázzák ezt az esetet: a Princeton Wordnet 2.0-s verziójában a *give* ige több mint 40 synsetben szerepel (nem kollokációban, hanem önálló igeként), s ebből több mint 20 esetben egymagában szerepel a synsetben, egyéb szinonima nélkül. Ezek közül a synsetek közül többnek is olyan magyar megfelelője van, amelyben az *ad* ige szintén önmagában, szinonima nélkül szerepel. Két ilyen synsetpár jelentését alább idézzük (a definíciót illetve egy példamondatot emelünk ki):

- give:27 def.: estimate the duration or outcome of something
  He gave the patient three months to live.
  ad:16 def.: Időtartamot megbecsül.
  Az orvosok három hónapot adtak a betegnek.
- give:40 def.: allow to have or take
  I give you two minutes to respond.
  ad:12 def.: Valamennyi időt kiszab, illetve engedélyez vmire.
  Öt percet adok neked arra, hogy elkészülj.

A WordNetnek ez a tulajdonsága megnehezíti az egyébként is köztudottan nehéz jelentésannotációs feladatot[2], és rontja annak az esélyét, hogy a jelentésannotációs feladatot végző humán annotátorok közötti egyetértés megüssön egy, a gépi jelentésegyértelműsítéshez elfogadható referenciamértéket.[3]

[2] "Wordsense tagging is one of the hardest annotation tasks." ([3])

[3] Ilyen jellegű kísérletre a tavalyi évben került sor, amikor megvizsgáltuk (l. [7] és [8]), hogy – többek között – a Magyar WordNet igei része mennyire alkalmas arra, hogy egy szövegben a legpoliszémebb igék előfordulásait humán annotátorok egyértelműen beannotálják jelentésekkel. A vizsgálat kiértékelése az annotátorok közötti egyetértés mértékét vette figyelembe. A Magyar WordNet (csakúgy, mint a többi vizsgált adatbázis) ezen a vizsgálaton gyenge eredményt ért el, azaz az annotátorok közötti egyetértés mértéke nem ütötte meg azt a szintet, amelyet általánosan az egyértelmű jelentésannotáció feltételeként szabni szoktak.

Nem minden jelentés-egyértelműsítési feladathoz szükséges azonban, hogy a kiértékelést egy humán annotátorok által jelentésannotált tesztkorpuszhoz képest végezzük. A fordítás szempontjából ugyanis legitimálható módon nincs jelentősége, hogy a jelentéstár "megfelelő" vagy "nem megfelelő" jelentése vezetett-e a helyes fordításhoz. A fenti esetet példáként véve mindegy, hogy az *ad* igét tartalmazó két synset közül melyiken keresztül érjük el a *give* angol megfelelőt. A gépi fordítás olyan speciális, jelentés-egyértelműsítést igénylő feladat, ahol a forrásnyelvi szó célnyelvi fordításának megfelelő volta elegendő kiértékelési szempont. A kiértékelés itt tehát hasonlítható a rossz matematikatanuló módszeréhez: mindegy, hogy hogyan jutunk el a végeredményhez, csak helyes legyen. Tanulmányunkban tehát nem közvetlenül a HuWN-nel végzett jelentés-egyértelműsítés kiértékelése zajlik egy humán annotátorok által jelentésannotált tesztkorpuszhoz képest, hanem annak kiértékelése, hogy a párhuzamos WordNetek alapján kapott lexikális szintű fordítás hogy viszonyul a gépi fordítás során kapott fordítási eredményhez.

2 Célkitűzés és módszertan

Esettanulmányunkban azt vizsgáltuk, hogy a Magyar WordNet mint jelentés-egyértelműsítő rendszer és a vele összekötött Princeton WordNet mint angol szótár tud-e javítani lexikális szinten egy adott a magyar–angol irányú gépi fordítórendszer teljesítményén, illetve hogy szófajonként van-e releváns különbség az eredményekben.

2.1 A felhasznált erőforrások

Kísérletünkhöz a magyar–angol nyelvpárra elérhető gépi fordító szolgáltatások közül a legjobb teljesítményt nyújtó[4] MetaMorpho rendszert választottuk. A MetaMorpho gépi fordítórendszer szabályalapú rendszer, de a transzfer és közvetítőnyelves módszerekkel szemben kizárólag direkt megfogalmazásokból áll. Ezek a direkt megfeleltetések azonban nem direkt módon, hanem az elkülönülő generáló fázisban érvényesülnek. A minták egységesen szolgálnak a nyelvtan és szótár leírására is. A fordító gerincét igei vonzatkeretminták adják, amelyeknek az illesztése kulcsfontosságú a fordítandó mondat szintaktikai elemzése szempontjából. A fordítórendszerben a szófaji egyértelműsítést a forrásnyelvi mondat névszói szerkezeteinek és az igei vonzatkeretminták illesztése biztosítja. Az igei vonzatkeretminták illesztése egyben poliszém igék jelentései közötti jelentés-egyértelműsítést is végez.

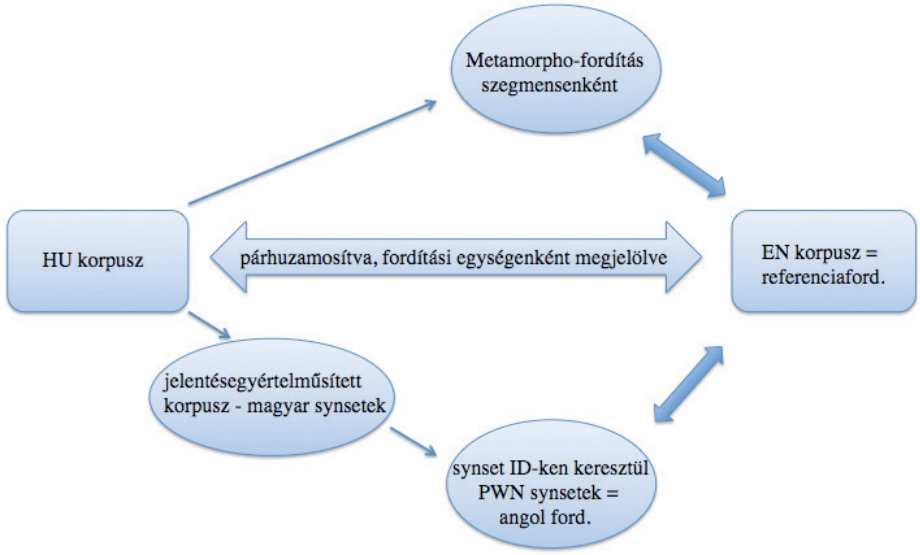
A Magyar WordNet (HuWN) jelenlegi állapotában kb. 37.000 fogalmat tartalmaz, melyeknek nagy része főnév (mintegy 28.500 fogalom), a maradék 8.500 fogalom szófaji megoszlása pedig a következő: 4100 melléknév, 3400 igei, 1000 határozószó. A Magyar WordNet csaknem minden synsete vagy egyedi azonosítóján keresztül, vagy egy ún. nyelvközi relációval meg van feleltetve az angol nyelvű Princeton WordNet 2.0-s verziójának.

[4] A teljesítményt előzetes felmérések alapján becsültük meg.

A vizsgálathoz egy EU-s rövidhíreket tartalmazó párhuzamos korpuszt használtunk, amely mind magyarul, mind angolul kb. 50.000 szövegszó hosszúságú. A korpusz a http://ec.europa.eu/news/ weboldalról származik.[5] Egy-egy rövidhír átlagosan 10 mondatot tartalmaz. A korpusz öt domént ölel fel (mezőgazdaság, pénzügy, kultúra, gazdaság, munkaügy), azaz az általános szókincset hivatott lefedni.

2.2 A munkafolyamat

A vizsgálat a következő fő lépésekből áll: (1) a magyar nyelvű korpusz főneveinek, igéinek és mellékneveinek jelentés-egyértelműsítése a HuWN felhasználásával, majd a beazonosított jelentések (synsetek) angol megfelelőinek kikeresése az angol nyelvű WordNetből, (2) a magyar nyelvű korpusz lefordítása a MetaMorpho fordítóval, majd a fordítás lemmatizálása, (3) az 1. és 2. lépésekből nyert angol fordítások helyességének automatikus kiértékelése a párhuzamos korpusz angol mondataihoz képest, lemmaszinten.[6] A vizsgálat főbb lépéseit az alábbi ábra szemlélteti:



1. ábra. A vizsgálat főbb lépései.

[5] Ezúton szeretnék köszönetet mondani Héja Enikőnek az általa gyűjtött korpusz felhasználhatóvá tételéért, valamint a jelen tanulmányhoz fűzött hasznos megjegyzéseiért.

[6] A vizsgálat elvégzéséhez a következő előfeldolgozó lépésekre volt szükség: (1) a Magyar WordNet XML formátumából a jelentés-egyértelműsítő szoftvernek megfelelő bemeneti fájlokat készíteni, (2) a magyar nyelvű korpuszt a jelentés-egyértelműsítő szoftver által megkövetelt bemeneti formátumra hozni, (3) a rendelkezésünkre álló magyar és angol korpuszt lemmatizálni, (4) a korpuszokat párhuzamosítani (szószintű megfeleltetésre nem volt szükség).

Az ábrán vastag ferde nyilak jelölik a kiértékelés automatikusan végezhető részét: (a fenti felsorolásban a (3)): mind a MetaMorpho gépi fordítórendszer által nyújtott lexikális fordításokat, mind a magyar és angol párhuzamos WordNeteken keresztül kapott lexikális fordításokat összevetettük fordítási egységenként (szegmensenként) a referenciafordításként használt, lemmatizált angol korpusszal.

A magyar korpusz szavainak jelentés-egyértelműsítését a HuWN gráf felhasználásával a Baszk Egyetem által fejlesztett, ingyenesen elérhető, nyelvfüggetlen, UKB nevű eszközzel végeztük (l. [2]), amely a PageRank algoritmust [4] használja fel a jelentés-egyértelműsítés során. Az UKB az általa használt tudásbázist (jelen esetben a WordNetet) gráfként kezeli, melyben a csomópontokat a relatív szerkezeti jelentőségükhöz mérten súlyozza. Az egyes csomópontok súlya a hozzájuk vezető relációktól függ: ha *i* és *j* csomópont között van kapcsolat, akkor *j* súlya *i* súlyának arányában megnő. A program kimenete az adott szóra a szóhoz tartozó csomópontok (itt WordNet synsetek) közül a legnagyobb súlyú. Agirre és Soroa [2] újítása többek között abban áll, hogy az egész tudásbázis gráfját felhasználják, nem csak egy algráfot vonnak be a jelentés-egyértelműsítésbe.

A magyar nyelvű jelentés-egyértelműsítés során a korpusz főneveit, igéit és mellékneveit egyértelműsítettük, azaz az UKB minden ilyen szóhoz hozzárendelt egy-egy WordNet-synsetet. A WordNet-synsetek azonosítóján keresztül eljutottunk a PWN megfelelő fogalmához, azaz a kiindulási szavunk WordNet által javasolt angol nyelvű fordításához. A korpuszt a MetaMorpho fordítóval is lefordítottuk. A kapott gépi fordítást lemmatizáltuk, és kiválogattuk belőle a főneveket, mellékneveket és igéket.

A MetaMorpho fordítót a teljes korpuszszövegen futtattuk le, nem pedig lemmatizált alakokon, mert a fordító mintaillesztése szempontjából szükséges, hogy teljes igei szerkezeteket felismerhessen, és ne csak lemmákat fordító szótárként funkcionáljon a program (pl. egyben felismerjen olyan kollokációkat, mint az *érvénybe lép*). Ilyen szintű kollokációfelismerésre a WordNet a jelen egyértelműsítő algoritmus használata mellett nem volt alkalmas (bár pl. az említett kollokációt tartalmazza), mert az UKB által megkívánt korpuszbemenet lemmatizált alakokat kívánt meg, azaz az esetleges többszavas kifejezéseket eleve elválasztva kezelte, és próbálta egyértelműsíteni.

A hírkorpusz szövegében előforduló főneveknek jelentős hányada tulajdonnév (Named Entity). Ezek, a hírek aktuálpolitikai jellegéből adódóan túlnyomó többséggel olyan esetleges, szótárban nem szereplő nevek, amelyek, ha a kiértékelésnél figyelembe vennénk őket, torzítanák a két fordítási módszer összevethetőségét.[7] Ezért ezeket mind a WordNet, mind a MetaMorpho fordításaiból kiszűrtük.

A gépi fordítórendszer, illetve a jelentés-egyértelműsítő algoritmus jellegét tekintetbe véve előzetes elvárásunk az, hogy az igék esetében a MetaMorpho rendszer fog jobb fordítást nyújtani – lévén, hogy a fentiekben említett igei kollokációfelismerésre a WordNet számára a jelen vizsgálatban nem volt esély. Főnevek esetében feltételezhető, hogy a WordNeten keresztül kapott fordítások nagyobb arányban lesznek jobbak, mint a MetaMorpho rendszeréi, mivel az UKB egy adott szó egyértelműsítésekor a szó szövegkörnyezetében előforduló, vele valamilyen szemantikai viszonyban lévő

[7] A MetaMorpho rendszer ugyanis, ha nem ismer fel egy szót, visszadja azt kimenetként – azaz tulajdonnevek esetében automatikusan, valódi fordítási eljárás nélkül helyes kimenetet produkál.

szavakat vizsgálja, és veti össze a HuWN gráffal, és a HuWN főnévi gráfja kellően nagy ahhoz, hogy sikeresnek tételezzük fel ezt a műveletet. A melléknevek esetében kb. egyforma teljesítményre számítunk a két fordítórendszer részéről.

3 Eredmények és kiértékelés

Az automatikus kiértékelő lépésben a két módszerrel kapott fordításokat összevetettük a lemmatizált angol korpusszal, amit emberi fordításként, referenciafordításként kezeltünk. Mind a két gépi fordítási kimenetben (itt tág értelemben véve a WordNet által kínált lexikális fordításokat is gépi fordításnak nevezzük) megőriztük azokat a szegmenshatárokat, amelyek a magyar és angol korpusz párhuzamosításakor mint fordításiegység-határok születtek. Így az összehasonlítás szegmensenként történt. Amennyiben egy adott szegmens egy lexikális egységének volt megfelelője az angol korpusz megfelelő szegmensében, azt automatikusan jó fordításnak könyveltük el.[8] Az ilyen találat hiánya azonban nem tekinthető automatikusan rossz fordítás indikátorának – hiszen egy szövegnek többfajta helyes fordítása is létezhet –, pusztán kézi kiértékelést tesz szükségessé.

A vizsgálatot az a kérdés vezérelte, hogy a HuWN felhasználása a lexikális jelentés-egyértelműsítésben javíthat-e lexikális szinten a magyar-angol irányú gépi fordítás minőségén. A fordítás eredményét pontosság (precision) alapján mértük, azaz a jól fordított szavak és az összes lefordított szó arányára voltunk kíváncsiak, s ezt hasonlítottuk össze a két módszerrel kapott fordítások esetében. Jelen vizsgálat során a fedés (recall) mérése nem volt alkalmazható, a következők miatt. A WordNet által kapott fordítások esetében meg lehet határozni, hogy a korpusz összes főnévi, melléknévi és igei szavának mekkora százalékára készült fordítás. Ugyanez azonban nem állapítható meg a MetaMorphóval készült fordítás esetében. A MetaMorpho rendszer ugyanis nem szavakat, hanem mondatokat fordított, s ily módon nem feleltethetők meg egymásnak egyértelműen a magyar szöveg és az angol fordítás szavai. A MetaMorphóval készült angol fordítás összesen jóval több főnevet, melléknevet, igét tartalmaz, mint a magyar korpusz – míg a WordNettel készült fordításra ez nem igaz, hiszen ott lexikális elemeket fordítottunk. Az arányok összehasonlítása tehát értelmetlen lenne a két esetben.

A kiértékeléskor tokenek arányát vettünk figyelembe, bár ez implikálja, hogy egy gyakoribb előfordulású szó nagyobb súllyal szerepel a kapott pontossági értékben. Ennek ellenére, mivel szövegek fordítási minőségéről van szó, ennek a súlyozásnak van létjogosultsága.

A WordNeten keresztül kapott fordítások esetében az esetleges többszavas kifejezések helyességének automatikus ellenőrzésére két út is kínálkozott: a referenciafordítással való összehasonlításkor egyrészt tekinthettük illeszkedésnek, ha az adott többszavas tagjai közvetlenül egymás mellett jelentek meg, de megengedőbb esetben

[8] E mögött az az előfeltevés húzódott meg, hogy amennyiben ugyanaz a szó szerepel az emberi fordítás és egy „gépi" fordítás ugyanazon fordítási egységében, jogunk van feltételezni, hogy ugyanannak a szónak a fordításáról van szó, nem pedig puszta véletlenről.

azt is, ha a tagok bárhol szerepeltek a megfelelő referenciaszegmensben.[9] Alább bemutatjuk mindkét úton kapott eredményeket.

1. táblázat: Az automatikus kiértékelés eredményei: a két fordítási módszer pontossága.

	HuWN pontossága a többszavas kif-ek pontos illesztésével	HuWN pontossága a többszavas kif-ek laza illesztésével	MetaMorpho pontossága
főnév	31,69%	31,81%	32,24%
melléknév	28,13%	28,27%	32,96%
ige	15,22%	15,28%	20,12%
össz.:	**28,12%**	**28,22%**	**28,97%**

A fentiek fényében elmondhatjuk, hogy az eddig elvégzett automatikus kiértékelés alapján nagyságrendileg mindkét fordítási módszer hasonló eredményt nyújtott. A MetaMorpho rendszer mind összesítésben, mind szófajokra bontva jobban teljesített, mint a két párhuzamos WordNet mint lexikális fordító. Előzetes elvárásunk, miszerint a MetaMorpho az igei jelentés-egyértelműsítő mechanizmusának köszönhetően az igék esetében jobb fordítást nyújt majd, mint a WordNeteken keresztül kapott fordítások, annyiban is beigazolódott, hogy a két rendszer pontossága közötti nagyságrendi különbség az igék esetében a legnagyobb.[10] Azon hipotézisünk, miszerint a főnevek esetében a WordNeteken keresztül kapott fordítások bizonyulhatnak jobbnak, nem igazolódott be, bár nyilvánvaló, hogy nagyságrendileg a főnevek esetében közelíti meg egymást leginkább a két módszer pontossága.

4 További munkálatok

Ahhoz, hogy a jelenlegi kiértékelésnél megbízhatóbb eredményt kapjunk, természetesen szükséges az automatikusan nem kiértékelhető fordítások (legalább egy részének) kézi kiértékelése. Érdekes lenne azt is megvizsgálni, hogy a két fordítórendszer hibaeseteiben van-e felismerhető tendencia: hasonló esetekben adnak-e rossz fordítást, vagy komplementer esetekben. További kutatás tárgya lehetne egyrészt az ellentétes irányú nyelvpáron (angol–magyar) lefuttatott hasonló kísérlet ugyanezzel a két fordí-

[9] A MetaMorpho által nyújtott fordítást lemmatizálás után tudtuk csak összevetni a referenciafordítással, így ebben a lépésben sajnos mindenképp külön tagokra bontódtak fel az esetleges többszavas kifejezések. Tagjaikat azonban külön-külön természetesen meg lehetett találni az angol korpusz megfelelő mondatában.

[10] A Szlovén WordNet és egy szlovén-angol gépi fordító program (Presis) viszonylatában Fišer és Vintar [6] hasonló kísérletet végzett, amelyen valamivel jobb eredményt el a WordNet által nyújtott fordítás. Ez valószínűleg annak tudható be, hogy a Presis fordítórendszerben semmiféle jelentésegyértelműsítés nincs beépítve.

tórendszerrel, valamint más, magyar–angol / angol–magyar nyelvpárra elérhető fordítók és a magyar–angol párhuzamosított WordNet teljesítményének összehasonlítása.

Bibliográfia

1. Agirre, E., Edmonds, Ph.: Word sense disambiguation. Algorithms and Applications. (Text, Speech and Language Technology). Springer-Verlag New York, Inc., Secaucus, NJ (2007)
2. Agirre, E., Soroa, A.: Personalizing PageRank for Word Sense Disambiguation. In: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece (2009)
3. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics Vol. 34 No. 4 (2008) 555–596
4. Brin S., Page L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems Vol. 30 No. 1-7 (1998)
5. Fellbaum, C.: WordNet An Electronic Lexical Database. MIT Press (1998)
6. Fišer D., Vintar, Š.: Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajanju. In: Proceedings of the 13th International Multiconference Information Society - IS 2010 (2010)
7. Héja, E., Kuti, J., Sass, B.: Jelentésegyértelműsítés - egyértelmű jelentésítés? In: Tanács A., Szauter D., Vincze V. (szerk.): MSZNY2009, VI. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2009) 348–352
8. Kuti, J., Héja, E., Sass, B.: Sense disambiguation - "Ambiguous sensation"? Evaluating sense inventories for verbal WSD in Hungarian. In: Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages (2010)

A metaforikus nyelvhasználat korpuszalapú elemzése

Babarczy Anna[1], Bencze Ildikó[1], Fekete István[1],
Simon Eszter[1,2]

[1] BME Kognitív Tudományi Tanszék,
1111 Budapest, Stoczek u. 2.
{babarczy,ibencze,ifekete,esimon}@cogsci.bme.hu
[2] MTA Nyelvtudományi Intézet,
1068 Budapest, Benczúr u. 33.

Kivonat: Jelen tanulmány a konkrét–absztrakt (vagy metaforikus) nyelvhasználat korpuszalapú elemzésének segítségével arra keresi a választ, hogy a fogalmi metafora hipotézis milyen mértékben járulhat hozzá a metaforikus jelentések értelmezéséhez természetes nyelvi szövegekben. A kutatás céljaira egy 300 ezer szavas magyar nyelvű korpuszt hoztunk létre különböző szövegtípusokból. [3] és [2] metaforaindexéből 12 ismert fogalmi metaforát választottunk ki, melyek mindegyikéhez két kifejezéslistát állítottunk össze – az egyik a forrástartományra, a másik a céltartományra jellemző kifejezéseket foglalta magában. A metaforák automatikus azonosításához Martin módszerét [17] alkalmaztuk, vagyis olyan mondatokat kerestünk, amelyekben mindkét tartomány kifejezései szerepeltek egyazon mondaton belül. A hipotézis alapján azt feltételeztük, hogy ha egy mondat tartalmaz egy forrás- és egy céltartományi kifejezést is, akkor az metaforikus lesz. Az eredmények azt mutatják, hogy egy forrás–cél tartománypároson belül nem bármilyen asszociáció vezet metaforikus értelmezéshez, és a valóban metaforicitásra utaló relációk mibenléte leginkább az adott szöveg nyelvi tulajdonságain múlik.

1 Absztrakt tudás: a kognitív metaforaelmélet és a statisztikai megközelítés

A metaforamegértés kérdése szorosan kapcsolódik az absztrakt tudás reprezentációjának és elsajátításának problémaköréhez. Az elvont tudás reprezentációjáról két radikálisan különböző, de elméletileg összeegyeztethető megközelítés uralkodik a kognitív tudományban: (i) a testesültség (*embodiment*) hipotézis, amely szerint absztrakt tudásunk fogalmi metaforákra vezethető vissza; és (ii) a statisztikai tanulás hipotézis, amely szerint a nyelv statisztikai tulajdonságai segítségével sajátítjuk el az elvont fogalmakat.

A testesültség nézet a kognitív nyelvészet egyik alaptétele [6, 12, 13, 14, 15]. Kognitív nyelvészeti megközelítésben az absztrakt fogalmak mentális reprezentációja fizikai, elsősorban az ember által közvetlenül tapasztalható jelenségeken alapul. A nyelvi folyamatokra vonatkozóan ez azt jelenti, hogy az elvont fogalmakat (a céltartományt) kódoló nyelv produkciója és feldolgozása konkrét fogalmak (a forrástarto-

mány) metaforikus kiterjesztésén keresztül valósulhat meg, amihez ún. fogalmi metaforák, azaz konkrét—absztrakt fogalmi megfeleltetések állnak a nyelvhasználó rendelkezésére [12, 14]. Az *időt pazarol* vagy *védi az álláspontját* kifejezések hátterében például az AZ IDŐ PÉNZ, illetve az A VITA CSATA fogalmi metaforák állnak: a forrástartományokat a pénzről és a csatáról alkotott konkrét tapasztalatunk képezi, amelyeket aztán az 'idő' és a 'vita' elvont fogalmak strukturálására használunk fel. A sokak által vitatott [pl. 5, 18] kognitív nyelvészeti álláspont erős verziója szerint az ember képtelen absztrakt fogalmakról konkrét terminusok nélkül gondolkodni.

Az absztrakt tudás mibenlétének másik magyarázata, a statisztikai tanulás elmélete szerint a nyelv statisztikai tulajdonságai segítségével sajátítjuk el és strukturáljuk absztrakt fogalmainkat [2, 16]. A nézet azt hirdeti, hogy a nyelvhasználó a számára ismert nyelvi szimbólumok disztribúciós tulajdonságaiból kiindulva értelmez és használ újabb nyelvi szimbólumokat, beleértve az elvont fogalmakat kódoló nyelvet is.

A két megközelítés elméletileg megfér egymás mellett, hiszen elképzelhető, hogy absztrakt tudásunk mindkét forrást felhasználva alakul ki. A testesültség és a statisztikai elmélet közti különbséget tulajdonképpen arra a kérdésre vezethetjük vissza, hogy lehetséges-e szimbólumlehorgonyzás (jelentés kialakulása) kizárólag nyelvi szimbólumokra épülve. A metaforamegértésre nézve pedig az a lényegi kérdés, hogy függetlenek-e az absztrakt fogalmak a konkrét fogalmaktól a nyelvhasználat során. A két szemlélet a különféle kognitív rendszerek és modalitások szerepét és súlyát vitatja az absztrakt tudásunk reprezentációjának kérdése kapcsán.

Jelen tanulmány a konkrét–absztrakt (vagy metaforikus) nyelvhasználat korpuszalapú elemzésének segítségével arra keresi a választ, hogy a fogalmi metafora hipotézis milyen mértékben járulhat hozzá a nem szó szerinti (vagy metaforikus) jelentések értelmezéséhez természetes nyelvi szövegekben. [12] és [14] metaforaindexéből 12, a nemzetközi szakirodalomban ismert fogalmi metaforát választottunk ki, amelyeket különböző szövegtípusokon teszteltünk. A hipotézis szerint egyrészt azt feltételeztük, hogy ha egy adott mondat tartalmaz egy elvont céltartományhoz tartozó fogalmat kódoló kifejezést, akkor a megfelelő forrástartományhoz tartozó kifejezést is tartalmaznia kell [17]. Másrészt, ha egy mondat tartalmaz egy forrás- és egy céltartományi kifejezést is, akkor a mondat nagy valószínűséggel metaforikus lesz.

2 Korábbi eredmények a testesültség hipotézis mellett és ellen

2.1 Pszicholingvisztikai kísérletek

Gibbs és Matlock [8] a testesültség hipotézist, pontosabban a metaforikus szimuláció (*metaphorical simulation*) jelenségét vizsgálja. Eszerint a metaforikus kifejezések értelmezése során egyfajta fizikai mozgásszimulációt végzünk, vagyis elképzeljük a metaforikusan használt szó által leírt konkrét cselekvést vagy eseményt. Kísérleti bizonyítékaink vannak arra, hogy például a szenzomotoros tapasztalat nagyban befolyásolja az időről szóló metaforikus kifejezések értelmezését [1], akárcsak a párkapcsolatokat utazásként leíró szövegek megértését [8]. A szerzők értelmezése szerint a testi szimuláció, azaz a szó szerinti jelentés mentális aktiválódása segíti a figuratív nyelv feldolgozását.

Ezek az eredmények azt a feltételezést támasztják alá, miszerint a metaforák megértésekor valóban az absztrakt céltartománynak a konkrét forrástartományra való leképezése történik meg. Ugyancsak ezt a feltételezést erősítik meg egyes lexikális döntési feladatok, melyek során azt figyelték meg, hogy az A DÜH EGY TARTÁLYBAN LÉVŐ FELHEVÍTETT FOLYADÉK (ANGER IS A HEATED FLUID IN A CONTAINER) vagy az AZ OPTIMIZMUS FÉNY (OPTIMISM IS LIGHT) konceptuális metaforákat tartalmazó kifejezések után a kísérleti alanyok gyorsabban döntöttek arról, hogy a *hőség*, illetve a *fény* lexémák valódi szavak-e, mint az ilyen metaforákat nem tartalmazó kifejezések után [7].

Akadnak azonban olyan vizsgálatok is, amelyek az eddig elmondottakkal ellentétes következtetésre jutottak. Keysar és munkatársai [10] olyan szövegek megértésének feldolgozási idejét mérték, amelyek az A SZERELEM EGY BETEG PÁCIENS (LOVE IS A PATIENT), az A VITA UTAZÁS (ARGUMENT IS JOURNEY) és az A GONDOLATOK EMBEREK (IDEAS ARE PEOPLE) fogalmi metaforákat tartalmazták. A kísérlet olyan újszerű metaforikus kifejezések megértését tesztelte, amelyeket vagy a már említett metaforatípus konvencionális példái előztek meg, vagy pedig nem metaforikus mondatok. Az eredmények szerint az első esetben a megértés nem volt gyorsabb, mint a másodikban, ami arra utal, hogy a konvencionális kifejezések feldolgozása során nem aktiválódtak a megfelelő fogalmi leképezések. AZ AZ IDŐ TÉR (TIME IS SPACE) fogalmi metafora vizsgálatát célzó kísérletek eredményei szintén nem adnak egyértelmű választ arra a kérdésre, hogy a téri sémák feltétlenül szükségesek-e az időről való gondolkodáshoz [21].

2.2 Korpuszelemzések eredményei

A korpusznyelvészeti módszereket segítségül hívó kutatók az elméleti megközelítések sokféleségének és a pszicholingvisztikai kísérletek ellentmondásos eredményeinek problémáját általában abban látják, hogy azok egyrészt túlságosan a metaforák fogalmi természetével vannak elfoglalva, és így figyelmen kívül hagyják a nyelvi tényezőket, másrészt nem természetes nyelvi adatokat használnak a kísérletek lebonyolításához, hanem nyelvi intuíciókon alapuló kitalált példákat, amelyek félrevezetőek lehetnek.

Stefanowitsch [20] az érzelmekkel kapcsolatos metaforák korpuszalapú elemzése során azt találta, hogy az ún. metaforikus sablon módszer (*metaphorical pattern method*), amely a metaforák céltartományára jellemző szavak korpuszokban való vizsgálatát jelenti, jóval hasznosabb lehet az elméleti kutatók által használt introspekciónál – két okból is: az egyik, hogy ezzel a módszerrel olyan metaforatípusokat is fel lehet lelni, amelyekről eddig nem esett szó a szakirodalomban, a másik pedig, hogy a gyakorisági mutatókat figyelembe véve meg lehet határozni, hogy az egyes céltartományokat mely leképezések jellemzik leginkább. A fogalmi metafora hipotézis szerint például a BOLDOGSÁG céltartományt a következő forrástartományok strukturálják: FENT, FÉNY, MELEGSÉG, TERMÉSZETI ERŐ stb. Stefanowitsch az általa használt módszerrel további forrástartomány-típusokat határozott meg, amelyek szintén a BOLDOGSÁG absztrakt kategóriát írják le: FOLYADÉK, ÖSSZETÖRHETŐ TÁRGY, BETEGSÉG, AGRESSZÍV ÁLLATI VISELKEDÉS, ORGANIZMUS stb.

A nyelvi metaforák grammatikai viselkedésének vizsgálata is olyan fontos részletekre világít rá, amelyeket a fogalmi metafora hipotézisben figyelmen kívül hagynak. Deignan [4] elemzéseiből kiderül, hogy a különböző szavak, kifejezések többnyire más-más grammatikai jellemzőkkel, illetve logikai relációkkal rendelkeznek a szó szerinti és a metaforikus használatban. Az AZ EMBERI VISELKEDÉS ÁLLATI VISELKEDÉS fogalmi metafora esetén például azok a szavak, amelyek a forrástartományban szerepelnek, és entitásokat jelölnek, metaforikus használatukban többnyire igeként vagy melléknévként fordulnak elő. A szerző egyéb metaforatípusok vizsgálata alapján számos példával mutatja meg, hogy metaforikus használatban a szavak jóval kevesebb grammatikai szabadsággal rendelkeznek, mint amikor szó szerinti jelentésükben jelennek meg. Ez azt jelenti, hogy a forrástartományban lévő entitások közti logikai reláció nem egyszerűen megismétlődik a céltartományban, ahogyan azt a kognitív metaforaelmélet jósolná, hanem át is alakul: a szavak metaforikus jelentésükben önálló életet kezdenek élni.

Természetesen olyan elemzések is léteznek, amelyek alátámasztják a fogalmi metaforákon alapuló megközelítést. Martin [17] a metaforákat megelőző kontextusokat vizsgálva azt találta, hogy azok a kontextusok jósolják be legmegbízhatóbban a célmetaforát, amelyek ugyanolyan típusú metaforikus kifejezéseket tartalmaznak, a legkevésbé pedig azok, amelyekben a forrástartomány szavai szó szerinti jelentésükben fordulnak elő. A szerző szerint ez az eredmény azt a korábbi kísérletet erősíti meg, amelyben azt találták, hogy a metaforikus kontextus felgyorsítja a célmetafora megértését, a forrástartomány szavainak szó szerinti jelentésben való használata pedig gátolja azt.

A fentebb bemutatott korpusznyelvészeti elemzések alapján a fogalmi metafora hipotézist és a pszicholingvisztikai kísérleteket célzó egyik legfontosabb kritika abban áll, hogy nem fektetnek elég hangsúlyt a metaforikus nyelvhasználat nyelvi jellemzőire. Ezek az adatok azonban, mint kiderült, igen fontosak, hiszen rámutatnak, hogy olyan egyéb tényezők is szerepet játszanak a figuratív nyelvhasználatban, mint a gyakoriság, a kollokáció, a nyelvi sablonok, a grammatikai formák, továbbá a nyelvi és szövegtípusbeli változatosságok.

3 A korpuszépítés

3.1 A korpuszelemzés módszertani kérdései

A metaforák korpuszalapú vizsgálata során több nehézséggel állunk szemben. Egyrészt a korpusz kiválasztása önmagában is meghatározó jelentőségű lehet, másrészt pedig a metaforikus kifejezések szövegekben való azonosítása sem problémamentes. Ez utóbbi azért okoz nehézséget, mert a kognitív szakirodalomban tárgyalt fogalmi leképezések általában nincsenek sajátos nyelvi formákhoz kötve, és így nem könnyű meghatározni azokat a nyelvi jegyeket, amelyek leginkább jellemezhetik az egyes tartományokat. Az egyik lehetséges módszer így a kézi keresés, amely során a kutatók saját nyelvi intuícióikra támaszkodva próbálják összegyűjteni egy adott korpuszból a szerintük metaforikusnak ítélt kifejezéseket. Mivel ez az eljárás meglehetősen idő- és munkaigényes, legalább részben automatizált módszerekkel is érdemes pró-

bálkozni. Ilyen módszer a forrástartomány szókincsére való rákeresés (pl. Deignan elemzései). Ebben az esetben összegyűjtjük egy adott metaforatípus forrástartományára potenciálisan jellemző szavakat, majd megnézzük, hogy milyen arányban fordulnak elő ezek metaforikus értelemben. Egy harmadik módszer a céltartomány szókincsére való rákeresés (pl. Stefanowitsch elemzései), amely talán azért lehet sikeresebb, mint az előző kettő, mert azokban a metaforikus mondatokban, amelyek tartalmaznak egy céltartományi kifejezést, általában egy forrástartományi kifejezés is megjelenik, s így nagyobb az esély az ún. metaforikus sablonok fellelésére. Végül negyedik módszerként olyan mondatokra is rákereshetünk, amelyek egy adott metaforának mind a forrás-, mind pedig a céltartományára jellemző szavakat is tartalmazzák (pl. Martin módszere). Ennek az eljárásnak az a hátránya, hogy így csak előre meghatározott metaforikus leképezéseket tudunk tesztelni, és a Stefanowitsch-féle módszerrel szemben az új metaforatípusok fellelése eleve kizárt. Ezzel szemben nagy előnye, hogy gyorsabban megy az annotálás, így nagyobb szövegeken is alkalmazható.

Természetesen mindegyik esetben szükség van egyrészt megfelelő szólisták összeállítására, másrészt pedig annak explicit meghatározására, hogy mi számít metaforikus kifejezésnek, és mi nem.

Az eddigi korpusznyelvészeti kutatások nagyrészt a metaforikus kifejezések nyelvi jellemzőire voltak kíváncsiak, ezért általában az első három elemzési módszer valamelyikét alkalmazták. Ezzel szemben jelen tanulmány elsősorban arra keresi a választ, hogy a fogalmi metaforáknak szövegekben való automatikus megtalálása menynyire sikeres a testesültség hipotézisét alapul véve. A megfelelő korpusz és elemzési módszer kiválasztására nézve ez a következőket jelentette:

- többféle fogalmi metaforát tesztelni;
- olyan korpuszt vizsgálni, amely többféle szövegtípusból áll;
- olyan példákat találni, amelyek mind a forrás-, mind pedig a céltartomány jellemző szavait tartalmazzák;
- kimerítő listát összeállítani mindkét tartományra vonatkozóan.

Ennek megfelelően [12] és [14] metaforaindexéből 12 széles körben ismert fogalmi metaforát választottunk ki, melyek közül az egyiknek mindkét irányú megvalósulását külön vizsgáltuk (a több fent van/a kevesebb lent van), így tulajdonképpen 13-féle annotáció lehetséges (a példák az általunk annotált szövegekből származnak):

1. A VÁLTOZÁS MOZGÁS (CHANGE IS MOTION): ***jön*** *a* ***hideg****; rohamléptekkel* ***közeledik*** *a* ***szünidő****; mélységes* ***szomorúság járta át*** *a lelkem*
2. AZ IRÁNYÍTÁS FENT VAN (CONTROL IS UP): ***magas rangú*** *katonatiszt; az amerikai parti* ***őrség*** *és a haditengerészet járőrei* ***felügyelik*** *a houstoni csatornát*
3. A TÖBB FENT VAN (MORE IS UP): ***magasabbra*** *kúszik az* ***átlaghőmérséklet****;* ***mértéken felül*** *bosszantott az ismeretlen idétlen tréfája*
4. A KEVESEBB LENT VAN (LESS IS DOWN): *mély* ***hangját lehalkítva*** *folytatta;* ***leszállították igényüket*** *kétszáz rúpiáról egy fémpumpára;* ***lelohad*** *a* ***szerelem***
5. A HALADÁS ELŐRE MOZGÁS (PROGRESS IS MOTION FORWARD): *a műszaki* ***haladás*** *[...]* ***előrevihet*** *bennünket ezen az úton;* ***rendbe jönnek*** *a dolgok*

6. AZ ERŐFORRÁSOK ÉTELEK (RESOURCES ARE FOOD): *rengeteg **áramot fogyaszt**; finom artériahálózat **táplálja vérrel** a sebészek beható vizsgálata alatt álló régiót*
7. AZ ELME GÉPEZET (THE MIND IS A MACHINE): *fokozni akarják **szellemi kapacitásukat**; **kattant** valami Mihail Alekszandrovics **fejében***
8. AZ IDŐ PÉNZ (TIME IS MONEY): *nem **pazarolom** az **időmet**; mennyi **időbe kerül** a kivitelezés*
9. A DÜH HŐSÉG (ANGER IS HEAT): *a **vita hevében** elfelejtettem bemutatkozni; a **lobbanékony** helytartó milyen különös formában **torolja meg***
10. A KONFLIKTUS TŰZ (CONFLICT IS FIRE): *le akartad **rombolni** a templomot, s erre **tüzelted** a népet; **kitört** a **háború***
11. AZ ELMÉLETEK ÉPÜLETEK (THEORIES ARE BUILDINGS): *a **genetika alapjai**; az öreg előbb **megdöntötte** mind az öt **bizonyítékot**, és aztán (...) ő maga **felállított** egy hatodikat*
12. AZ ALKOTÁS ÉPÍTÉS (CREATION IS BUILDING): *alapjaiban kell **átalakítanunk** az **életünket**; így **formálhattak jogot** a meghódított területekre*
13. A POLITIKA HÁBORÚ (POLITICS IS WAR): *a velejéig korrupt **kormány** és rendőrség **kirabolja** a népet; csakis az **ellenséges propaganda** állíthatja*

Mivel széles körben használt fogalmi metaforákat választottunk a testesültség hipotézisének korpuszalapú vizsgálatára, minél reprezentatívabb korpuszt kellett építenünk, amiben mindegyik választott metaforatípus elég sokszor fordul elő. A projekt eredeti célkitűzései között szerepelt, hogy a metaforákat többnyelvű párhuzamos korpuszon vizsgáljuk, ezért olyan szövegeket kellett szereznünk, melyek mind a 4 előirányzott nyelven (magyar, angol, spanyol, olasz) elérhetők és szabadon felhasználhatók kutatási célokra. Jogtiszta szövegeket gyűjteni mind a 4 nyelven meglehetősen nehéz, sokszor kivitelezhetetlen feladatnak bizonyult, így végül magyar nyelvű korpuszunk csak 3 szövegtípust tartalmaz: regények, *National Geographic* cikkek és filmfeliratok az alábbi arányban:

1. táblázat: A korpusz összetétele.

Szövegtípus	**Szövegszavak száma**
National Geographic cikkek	68 997
Filmfeliratok	32 148
Regények	208 384
Összesen	309 529

Mivel a szövegek különböző formátumokban kerültek a birtokunkba, először is egységesíteni kellett őket: minden dokumentumot UTF-8 karakterkódolású sima szöveggé alakítottunk. A korpuszt szövegszavakra és mondatokra bontottuk, majd minden szövegszóhoz egyértelmű morfológiai elemzést rendeltünk a HunPos morfológiai egyértelműsítő [9] segítségével.

3.2 A gold standard korpusz

A cél tehát olyan korpusz létrehozása volt, amelyben a metaforikus mondatok meg vannak jelölve azzal a címkével, amelyik fogalmi metaforához tartoznak a kiválasztott 13 közül. Feltételezésünk szerint az a mondat metaforikus, amelyben egyaránt megtalálható egy fogalmi metafora forrás- és céltartományához tartozó szó. Módszereink teszteléséhez szükségünk volt egy *gold standard* korpuszra, melyben a metaforikus mondatok kézzel be vannak jelölve.

Erre a célra építettünk egy minikorpuszt a teljes korpusz 10 százalékából. A minikorpusz tehát kb. 30000 szövegszóból áll, és a teljes korpuszt arányosan reprezentálja. A minikorpuszt 3 részre osztottuk; mindegyik részben 2 annotátor kézzel bejelölte az általa metaforikusnak ítélt mondatokat. A metaforikusság meghatározásához [19] kritériumait vettük alapul, néhány ezek közül: az állandósult kifejezéseket, „halott metaforákat" vagy azokat, amelyek csak etimológiai szempontból számítanak metaforáknak, nem vettük figyelembe (pl. a *depresszió* nem számít metaforikusnak); az igekötők számítanak (a *le* vagy *fel* mint FENT, illetve LENT forrástartományok); az allegóriák nem; ha egy metaforának az ellentettjét találtuk, akkor azt nem vettük bele az adott metaforatípusba. Ezenkívül az összes vizsgált típusnál külön-külön röviden össze is foglaltuk a fontosabb útmutatásokat. Például az A TÖBB FENT VAN fogalmi metaforánál a következő útmutatót használtuk: „Minden olyan mennyiséget jelentő kifejezést annotálunk, amit vertikális skálán képzelünk el, pl. *ár, bér, hőmérséklet*. Minden olyat annotálunk, amiben szerepel a *csúcs* szó: *csúcstermelés, csúcstechnológia* stb. Az olyan kifejezések, amelyek arról szólnak, hogy valamiből sok van, és az egy nagy kupacot alkot – pl. *hegyekben áll, tornyosul* –, nem metaforák, nem annotáljuk."

Az annotátorok közötti egyetértés megállapításához a legegyszerűbb egyetértési mértéket használtuk, ami azt számolja, hogy az esetek hány százalékában ítéltek azonosan az annotátorok. Mivel rendkívül alacsony értékeket kaptunk – ami arra utal, hogy a metaforikusság definíciója eleve kérdéses, nehezen meghatározható –, úgy döntöttünk, hogy az annotátorok által metaforikusnak ítélt mondatok unióját vesszük. Így 155 metaforikusnak annotált mondat szerepel a *gold standard* minikorpuszunkban.

3.3 Az automatikus azonosításhoz használt szólisták összeállítása

A metaforák automatikus azonosításához Martin módszerét [17] alkalmaztuk, vagyis olyan mondatokat kerestünk, amelyekben mindkét tartomány kifejezései szerepeltek egyazon mondaton belül. A hipotézis alapján azt feltételeztük, hogy ha egy mondat tartalmaz egy forrás- és egy céltartományi kifejezést is, akkor az jó eséllyel metaforikus lesz. Ehhez szükségünk volt forrás- és céltartományi szavakat tartalmazó szólistákra. Illusztrációként álljon itt egy minta az AZ ELME GÉPEZET fogalmi metaforához tartozó listákból:

2. táblázat: AZ AZ ELME GÉPEZET fogalmi metaforához tartozó forrás- és céltartományi szólisták részlete.

the_mind_is_a_machine_source	**the_mind_is_a_machine_target**
erő	képzelet
kapacitás	szellemi
élesít	memória
működik	agykéreg
feldolgoz	információ
aktiválódik	agyterület
végrehajt	homloklebeny
létrehoz	fej

A forrás- és a céltartomány szólistáinak összeállítását három különböző módszerrel végeztük: a) asszociációs kísérlet alapján, b) szinonimaszótár alapján és c) referenciakorpusz alapján.

Az első módszer esetében a pszicholingvisztikai vizsgálatok körében bevett asszociációs kísérletet választottuk. 138 egyetemi diák végezte el a kísérletet, ami a következőképpen zajlott: a kiválasztott fogalmi metaforák hívószavai megjelentek a képernyőn, majd a kísérleti személynek egy perc állt a rendelkezésére, hogy olyan szavakat írjon, amelyek a tesztszóról eszébe jutnak. Például az A VÁLTOZÁS MOZGÁS metafora esetében a *változás* szó jelent meg a képernyőn mint forrástartományi, és a *mozgás* szó mint céltartományi tesztszó. Az így kapott listákat normalizáltuk: kiszűrtük a többszavas kifejezéseket, a tulajdonneveket és az ellentéteket, feloldottuk a rövidítéseket, majd töveltük a szavakat a Hunmorph morfológiai elemző [22] segítségével.

A második módszer során az asszociációs kísérletből nyert szólistákat kibővítettük a szavak szinonimáival a *Magyar szókincstár* [11] alapján. Ennek hatására a listák mérete természetesen sokszorosára nőtt, annak ellenére, hogy a szinonimák közül a népnyelvi, szleng és ritkán használt szavakat kihagytuk.

A harmadik módszer során [17] alapján tudatosan válogattunk össze szavakat mindegyik forrás- és céltartományhoz az előzőleg kézzel annotált *gold standard* minikorpuszból. Ebből következőleg ezt a módszert a későbbiekben a korpusznak egy másik 10 százalékán teszteltük.

Mindhárom szólista esetében a következő lépésben az eredeti és a morfológiailag egyértelműsített szövegekből, valamint a szólistákból XML-fájlokat állítottunk elő, amelyekben az eredeti szövegek az egyes szólistáknak megfelelően annotációkkal vannak ellátva, azaz a szólisták alapján feltételezett fogalmi metaforák jelölve vannak. Az XML-fájlok minden további korpusznyelvészeti feldolgozó lépését a GATE alkalmazás segítségével végeztük, amely egy könnyen kezelhető grafikus felülettel ellátott szövegfeldolgozó szoftvercsomag [3]. Az automatikusan annotált szöveget kézzel ellenőriztük, és korrigáltuk a szavak többértelműségéből adódó hibákat, azaz kitöröltük azokat a címkéket, amelyek a szólistán szereplő szóval megegyező alakú, de más jelentésű és/vagy szófajú szót jelöltek, pl. az A DÜH HŐSÉG fogalmi metafora esetében az *ég* és *nap* szavaknak az 'égbolt' és '24 óra' jelentésű előfordulásait.

3.4 Eredmények

A három módszer eredményeit az általánosan alkalmazott pontosság és fedés alapján értékeltük ki. A pontosság ebben az esetben azt mutatja meg, hogy az automatikus felismerő rendszer által metaforikusnak ítélt mondatoknak mekkora hányada ténylegesen metaforikus. A fedés értékéből pedig azt tudhatjuk meg, hogy az emberi elemzők által metaforikusnak ítélt mondatok közül hányat talált meg a rendszer. Az F-mérték pedig ezek súlyozott harmonikus közepe, vagyis a hatékonyság végső mérőszáma.

3. táblázat: A három módszer eredményei.

Módszer	Fedés	Pontosság	F-mérték
Asszociáció	6/155 (3,8%)	6/80 (7,5%)	5,65%
Szótár	28/155 (18,06%)	28/617 (4,5%)	11,28%
Korpusz	41/131 (31,29%)	41/74 (55,4%)	43,34%

Az eredményekből azt láthatjuk, hogy az asszociációs módszerrel lényegesen kevesebb olyan mondatot találtunk, amely forrás- és céltartományi kifejezést is tartalmaz, mint a másik két módszerrel. Pontosság tekintetében az asszociációs kísérlet valamivel jobb eredményre vezetett, mint a szótáralapú módszer, de mindkettő messze elmarad a korpuszalapú annotáció eredményétől. Ez utóbbi módszer bizonyult a legeredményesebbnek a fedés tekintetében is, vagyis a metaforikusság gépi azonosításában nem célravezető az asszociációs módszeren alapuló pszicholingvisztikai megközelítés, ellentétben a célzott kézi válogatással.

3.5 Problémás esetek

Az eddigiekből is tisztán látszik, hogy nem könnyű feladat egy mondatról eldönteni, hogy metaforikus-e vagy sem. Általános tapasztalat, hogy ha emberi erővel nehéz megtalálni egy szövegben bizonyos elemeket, akkor azok automatikus azonosítása sem fog jó eredményt hozni. Összességében ki kell mondanunk, hogy a testesültség elméletén alapuló feltételezésünk, miszerint egy metaforikus mondatban meg kell jelennie mindkét tartományi kifejezésnek, nem helytálló. Ezt erősítik a korpusz kézi annotálása során gyűjtött példák is, melyek metaforikusak ugyan, de nem szerepel bennük mindkét tartományhoz tartozó kifejezés, vagyis a fedés szempontjából problémásak.

Bizonyos mondatokban csak forrástartományi szót találunk: *Aztán egy nap **lelépett*** (A VÁLTOZÁS MOZGÁS). Itt csak egy mozgást kifejező szó van a mondatban, míg a változásra expliciten nem utal semmi, mégis pontosan tudjuk, hogy nem a járdáról való lelépésről van szó, hanem az esemény szereplőinek életében bekövetkezett változásról. Más esetekben a céltartományi kifejezés szerepel ugyan a szövegben, de nem a célmondatban, hanem az azt megelőző szövegkörnyezetben: *Van Toch kapitány majd megfullad a felháborodástól [...] arca bíborszínt öltött [...] feje **elkékült*** (A DÜH HŐSÉG).

Olyan esetek is léteznek, amikor legkevésbé sem a mondaton múlik annak metaforikussága, hanem csupán egy szón, amely magába foglalja a forrás- és céltartományi jelentést is: ***előléptetés*** (A HALADÁS ELŐRE MOZGÁS).

Továbbá olyan mondatból is sok van, amelyben szerepel ugyan mindkét tartományi kifejezés, mégsem metaforikus: *Mérnökök és **vezetők** tanakodnak kisebb csoportokban a 23 emelet **magas** fúrótorony tövében* (AZ IRÁNYÍTÁS FENT VAN). Ez utóbbi mondatok felelősek az alacsony pontossági értékekért.

4 Tanulságok, összegzés

Mivel kutatásunk célja elsősorban a fogalmi metaforák szövegekben való automatikus azonosítására irányult, nem tértünk ki a talált példák grammatikai elemzésére és a szövegek típusainak a különböző metaforákkal való összefüggéseire sem. Első ránézésre azonban úgy tűnik, vizsgálatunk eredményei megerősítik az előzőekben bemutatott korpusznyelvészeti elemzések eredményeit, főként a kollokációkat és a metaforikus kifejezések nyelvi formáját illetően. Erre utal az is, hogy míg az asszociációs kísérlet segítségével összeállított listák bejósló ereje nagyon gyenge volt, addig a tartományokra jellemző szavak szövegből való célzott összeválogatása hozta a legjobb eredményt, ami tehát azt jelenti, hogy nem bármilyen asszociáció vezet metaforához, hanem csak bizonyos szavak, kifejezések együttes előfordulása. Például a *pazarol* és *idő* vagy a *gerjeszt* és *harag* szavak egy mondaton belül szinte mindig metaforát eredményeznek. Ezenkívül a grammatikai forma fontosságát érintő adatokat is említhetünk: az AZ ERŐFORRÁSOK ÉTELEK fogalmi metaforánál a referenciakorpusz alapján a forrástartományt leginkább igék jellemzik (pl. *fogyaszt*, *felfal*, *táplál*); ezzel szemben az asszociációs módszerrel összegyűjtött szavak többsége főnév (pl. *edény, fagylalt, reggeli*). Ez megint csak azt a Deignan [4] által kapott eredményt támasztja alá, miszerint a metaforikus kifejezések jelentős hányadában a forrástartományt képviselő szavak többnyire igeként vagy melléknévként jelennek meg, aminek az lehet a magyarázata, hogy a metaforikus beszédben általában az absztrakt entitásokat próbáljuk leírni, és így a konkrét forrástartományból leginkább viselkedést, tulajdonságokat, cselekvést leíró szavakat veszünk át. Az asszociációs kísérlet egyik gyenge pontja tehát az lehetett, hogy bármilyen szót figyelembe vettünk, függetlenül annak szófajától.

Természetesen ezeknek a feltevéseknek az alátámasztásához a talált metaforák teljesebb elemzésére van szükség. Ugyanakkor az eredeti célkitűzést követve ugyanezeknek a szövegeknek az angol, spanyol és olasz nyelvű változatát is érdemes lesz a jövőben megvizsgálni, és az eredményeket összevetni a magyar adatokkal, hiszen abból újabb következtetéseket vonhatunk le a nyelvi tényezőkre és a fogalmi metaforák természetére vonatkozóan.

Bibliográfia

1. Boroditsky, L., Ramscar, M.: The roles of body and mind in abstract thought. In: Psychological Science Vol. 13 (2002) 185–188
2. Burgess, C., Lund, K.: Representing abstract words and emotional connotation in high-dimensional memory space. In: Proceedings of the Cognitive Science Society, Lawrence Erlbaum Associates, Hillsdale, NJ (1997) 61–66
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia (2002)
4. Deignan, A.: Metaphor and corpus linguistics. John Benjamins, Amsterdam/Philadelphia (2005)
5. Gentner, D., Bowdle, B., Wolff, Ph., Boronat, C.: Metaphor is like analogy. In: Gentner, D., Holyoak, K. J., Kokinov, B. N. (szerk.): The analogical mind: Perspectives from cognitive science. MIT Press, Cambridge MA (2001) 199–253
6. Gibbs, R. W.: Embodiment and cognitive science. Cambridge University Press, New York (2006)
7. Gibbs, R. W.: Experimental tests of figurative meaning construction. In: Radden, G., Köpcke, K-M., Berg, Th., Siemund, P. (szerk.): Aspects of meaning construction. John Benjamins, Amsterdam/Philadelphia (2007) 19–33
8. Gibbs, R. W., Matlock, T.: Metaphor, imagination and simulation. Psycholinguistic evidence. In: Gibbs, R. W. (szerk.): The Cambridge Handbook of Metaphor and Thought. Cambridge University Press, Cambridge (2008) 161–176
9. Halácsy, P., Kornai, A., Oravecz, Cs.: HunPos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic (2007) 209–212
10. Keysar, B., Shen, Y., Glucksberg, S., Horton, W. S.: Conventional language: How metaphorical is it? Journal of Memory and Language Vol. 43 (2000) 576–593
11. Kiss G.: Magyar szókincstár. Tinta, Budapest (2007)
12. Kövecses, Z.: Metaphor. A Practical Introduction. University Press, Oxford (2002)
13. Kövecses, Z.: A metafora. Gyakorlati bevezetés a kognitív metaforaelméletbe. Typotex, Budapest (2005)
14. Lakoff, G., Johnson, M.: Metaphors we live by. University of Chicago Press, Chicago (1980)
15. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York, NY (1999)
16. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. In: Psychological Review Vol. 104 No. 2 (1997) 211–240
17. Martin, J.H.: A corpus-based analysis of context effects on metaphor comprehension. In: Stefanowitsch, A., Gries, S.Th. (szerk.): Corpus-based approaches to metaphor and metonymy, Mouton de Gruyter, Berlin/New York (2006) 214–236
18. Murphy, G. L.: On metaphoric representation. Cognition Vol. 60 (1996) 173–204
19. Pragglejaz Group: MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol Vol. 22 No. 1 (2007) 1–39
20. Stefanowitsch, A.: Words and their metaphors: a corpus-based approach. In: Stefanowitsch, A., Gries, S. Th. (szerk.): Corpus-based approaches to metaphor and metonymy. Mouton de Gruyter, Berlin/New York (2006) 63–105

21. Szamarasz, V. Z.: Az idő téri metaforái: a metaforák szerepe a feldolgozásban. Világosság Vol. 47 (2006) 99–109
22. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software (2005) 77–85

IV. (Szemantikus) keresés

MASZEKER: projekt szemantikus keresőtechnológia kidolgozására

Szőts Miklós[1], Csirik János[2], Gergely Tamás[1], Karvalics László[3]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7
{szots, gergely}@all.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
csirik@inf.u-szeged.hu
[3] Szegedi Tudományegyetem, Könyvtár- és Humán Információtudományi Tanszék,
Szeged, Egyetem u. 2.
zkl@hung.u-szeged.hu

Kivonat: Egy merész nyelvészeti projektről számolunk be, a MASZEKER szemantikus keresést megcélzó projektről, amelyen az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem közösen dolgozik. A cél olyan technológia kidolgozása, amely a jól formált keresőkifejezés jelentésreprezentációját illeszti a szövegekre olyan egyezést keresve, amely kifejezheti a keresőkifejezés jelentését. Két felhasználási területre, mégpedig a szabadalmi keresésre, valamint néprajzi keresésre prototípus rendszert kívánunk fejleszteni. A technológiát nyelvfüggetlennek tervezzük, természetesen egyes komponenseinek nyelvfüggőnek kell lenniük. Angol és magyar nyelvű változatot fogunk fejleszteni. Magát a keresést végző rendszert kiegészítik az archívumot feldolgozó modulok (tematikus klaszterezés, témafüggő szinonimagenerálás).

1 Bevezetés

Annak ellenére, hogy a Google látszólag „egyeduralkodóvá" vált a keresőrendszerek piacán (vagy tán épp ezért) folyamatosan „forró terület" a nagyobb tudású (vagy akár új elvű) keresők fejlesztése. Ezért az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtár- és Humán Információtudományi Tanszéke közös projektet (TECH_08_A2/2-2008-0092) indított az NKTH támogatásával.

A tervezett projekt célja egy olyan, új elveken alapuló integrált keresőrendszer, a MASZEKER kifejlesztése, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül a keresést végző felhasználó szemantikai kompetenciáját az eddigieknél nagyobb mértékben kiaknázva teszi lehetővé a természetes nyelvi dokumentumtárakban (szövegekben) történő valóban *tartalmi* keresést. Egyszerűen szólva: a felhasználó jól formált frázisokkal, mondatokkal specifikálhatja, milyen tartalmú dokumentumokat keres.

A projekt során kifejlesztett technológia magja *nyelvfüggetlen*, a rendszer prototípusát pedig magyar és angol nyelvű szabadalmi leírások, illetve néprajzi anyagok feldolgozására fejlesztjük ki.

2 State of art

A bevezetőben említett „forró terület" látképéből minket a szemantikai keresők érdekelnek. Természetesen – mint annyi szakszó az informatikában – a „szemantikai" is a lehető legkülönbözőbben értelmezhető. Sokan a szavak, szóösszetételek szintjén értelmezik: szavak közti jelentésösszefüggések feltárásával egészítik ki a kulcsszó szerinti keresést. Ilyen a már elterjedt látens szemantika algoritmusa[1] (l. [5]). Elterjedőben van a keresők valamilyen ontológiához, tezauruszhoz való kapcsolása, ilyen alapon működik a magyar fejlesztésű, de nemzetközi hírnevet szerző HealthMash kereső is (l. http://www.weblib.com/products/healthmash). A MEDLINE-on működő KLEIO kereső (ismertetőt találhatunk [2]-ben) szintén ontológiákhoz van kapcsolva, de a névelemfelismerés (NER) technikáját is használja. A keresőkifejezésben megengedi, hogy a kulcsszavakhoz a felhasználó megadja annak besorolását, pl. *PROTEIN*:cat. Már ezzel is jelentősen javítja a keresés recallját, amint az idézett példa is illusztrál. Mi azonban szemantikai keresés alatt olyan folyamatot értünk, amely összefüggő szövegrészek jelentése alapján ítél valamely dokumentumot relevánsnak.

A szemantikus keresők két nagy osztályba sorolhatóak (l. [1]): lehetnek statikusak vagy dinamikusak. A statikus keresők előre elkészítik a keresett honlapok, dokumentumok szemantikus reprezentációját, és felindexelik azokat; míg a dinamikusak a keresőkifejezés jelentésreprezentációját a keresés alatt elemzett szövegrészekre illesztik. Másik általános osztályozási szempont az, hogy témafüggetlenek vagy egy területre specializáltak. Csak néhányat sorolunk itt fel, egy teljesebb áttekintés letölthető a www.maszeker.hu oldalról.

A HAKIA (l. [8]) általános célú, ontológiai szemantikára (l. [9]) alapozott, statikus keresőrendszer. Honlapok szövegei jelentésreprezentációjának alapján előre elkészíti a lehetséges kérdésekre adható válaszokat, amelyek közül az adekvátat a keresés közben csak ki kell választania. Inkább a tudáskinyerés területéhez tartozik, de a szemantikus keresés általában könnyen átfogalmazható tudáskinyerésre. A HAKIA egy erre a célra kifejlesztett, 8 500 fogalmat tartalmazó ontológiára támaszkodik. Ehhez csatlakozik egy kb. 100 000 szójelentést és több mint 1 000 000 szót tartalmazó szótár.

A Cognition (l. [3]) egy átfogó NLP framework, amely egy témafüggetlen keresőmotort is tartalmaz; szintén statikus rendszer. Több, egy-egy területre vagy dokumentumhalmazra specializált alkalmazása van, pl. a Wikipédiára, illetve a MEDLINE abstracts-ra is kifejlesztettek egy-egy speciális keresőt. Ontológiája 7 500 fogalmat tartalmaz, amelyekhez 536 000 szójelentés kapcsolódik.

A Powerset a Cognitionhoz hasonló rendszer. Sok információnk nincs róla, mivel a Microsoft megvette, és beépítette a fejlesztés alatt lévő keresőjébe (l. [10]).

[1] Részletes ismertetése letölthető a www.maszeker.hu honlapról.

Az UpTake (l. [14]) egy utazási információkat szolgáltató kereső, amely több mint 5 000 honlapot indexelt fel. Jellegzetessége, hogy a felhasználóval folytatott párbeszédet támogat, azaz az általánosabb kéréstől a specifikusabb felé mozoghat a felhasználó. Azt tervezik, hogy a rendszer alapjául szolgáló ontológiát tanulóalgoritmusokkal bővítik.

A GoWeb (l. [4]) az élettudományokra specializált kereső. Természetes nyelvű kifejezést fogad el inputként, s egy tradicionális, kulcsszó szerinti keresés eredményeit veti alá szemantikus elemzésnek. Háttere a Gene és a MeSH ontológia. Az eredményhez ezeknek az ontológiáknak releváns részleteit is megmutatja. E leírásból is kitűnik, hogy a GoWeb dinamikus kereső.

A MEDIE (l. [2], [7]) a már említett KLEIO-hoz hasonlóan a MEDLINE-on keres; azonban a KLEIO-hoz képest jelentős előlépés, hogy már szintaktikus és szemantikus elemzést alkalmaz az események kinyerésére. Egyelőre csak *alany-ige-tárgy* alakú kereső kifejezéseket kezel. [2] beszámol további kutatási irányokról, amelyek hasonlóak a mieinkhez.

3 A MASZEKER kereső felépítése

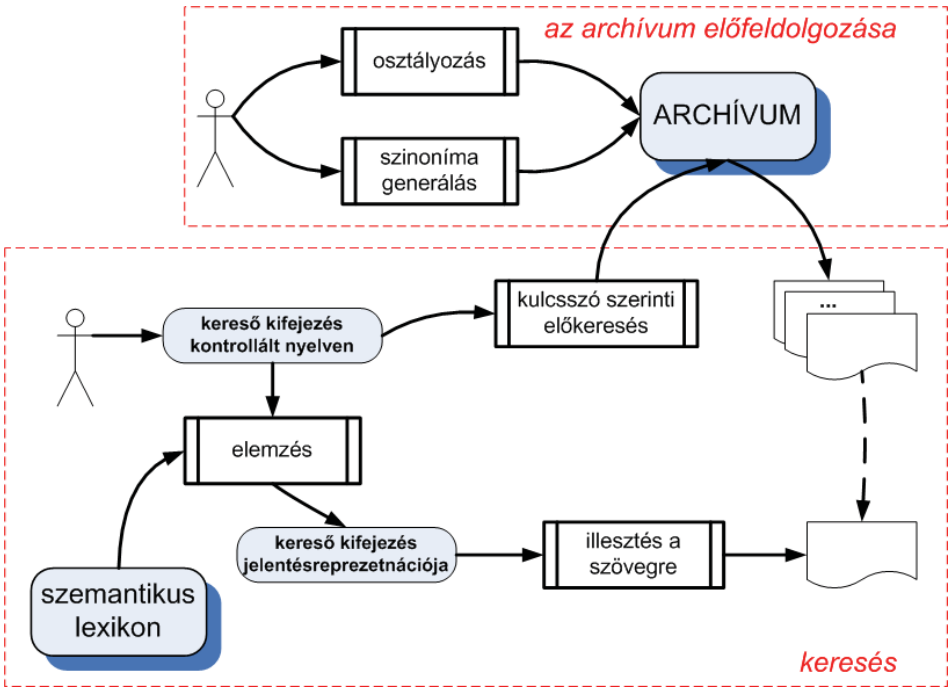
Ha a fent vázolt „tájképbe" illesztjük koncepciónkat, a következőképpen foglalhatjuk össze:

- általában nagyméretű ontológiákra épülnek a szemantikus keresők – mi egy kisméretű általános csúcsontológiát és ehhez csatlakozó, ugyancsak kisméretű tárgykörfüggő felső ontológiákat kívánunk használni;
- ennek megfelelően, – bár általános technológiai vázat építünk, – témakörökre kiélezett, tehát vertikális rendszereket kívánunk létrehozni;
- dinamikus keresőt tervezünk, bár bizonyos esetekben nem zárkózunk el az előzetes szemantikai feldolgozástól és felindexeléstől sem.

A rendszer áttekintő architektúrája az 1. ábrán látható.

Az ábrának megfelelően a releváns dokumentumok keresése a következő lépésekből áll:

1. a felhasználó egy kontrollált nyelven adja meg a keresőkifejezést,
2. a szintaktikus és szemantikus elemzés előállítja a keresőkifejezés jelentésreprezentációját,
3. a szavak szerinti keresés először az archívumot,
4. azokra a szövegszegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a keresőkifejezés jelentésreprezentációját.



1. ábra. A MASZEKER rendszer áttekintő architektúrája.

3.1 Elemzés

A szintaktikus elemzésre egy robusztus algoritmust dolgoztunk ki, amely azokat a részeket, amelyekkel nem tud megbirkózni, átugorja. A keresőkifejezés megadására szolgáló kontrollált nyelvet azonban pontosan elemzi.

A szintaktikus elemzés két lépésben történik. Egy előfeldolgozás kijelöl bizonyos pontokat a szövegben, pl. a felsorolás elemeinek kezdetét. Ezután egy dependencia-nyelvtanon alapuló elemző fut végig a szövegen. A szabadalmi szövegekben sok kvantitatív jelző fordul elő, a legváltozatosabb formában (például: *aspirin crystals 20-60 mesh in size* vagy *about 3-10% by weight of a polymeric mixture*). Ezekre külön CFG nyelvtant dolgoztunk ki.

A szintaktikus és a szemantikus elemzés párhuzamosan történik. Ennek több oka van, a legfontosabb az, hogy a szemantikus elemzés a szintaktikus elemzés bizonytalanságait segít kezelni, azaz visszahat a szintaktikus elemzésre, sőt a POS-tagger ítéleteit is változtathatja. Ugyanis beleütköztünk olyan hibás szófaj-meghatározásba, amely eltorzítja a szintaktikus elemzést. Főleg az angol nyelvben sok az olyan szó, amely egyaránt szerepel igeként és főnévként, például az *extract* szó.

A jelentésreprezentáció kialakítását davidsoni alapokon [11] kezdtük el, azaz az igék és az eseményszerűségeket jelentő főnevek jelentését reifikáljuk: maga az ese-

mény egy token lesz, és a szereplőket kötik hozzá szereprelációk. Logikailag azt jelenti, hogy a többargumentumú relációkat áttranszformáljuk kétargumentumúakra.

A davidsoni közelítés több szempontból is kedvező. A szemantikus lexikon szempontjából célszerűbb az eseményjelentésű szavakból kiindulni, amikor a jelentéskapcsolatokat leírjuk. Rugalmasságot ad: bármikor újabb dependenssel lehet bővíteni a leírást, mivel nem kell a relációjelentések argumentumszámát meghatározni. Illeszkedik a dependenciaalapú szintaktikus elemzés eredményére – valójában az összefüggés fordított: a dependenciaalapon működő szintaktikus elemző algoritmust választottuk az eseményalapú szemantikus szerkezethez. Robusztus is: ha nem áll rendelkezésre elegendő információ, a részleges jelentésreprezentáció automatikusan előáll.

Látható, hogy a szereprelációk megfelelnek a tematikus szerepeknek [11]. A különbség annyi, hogy nem kívánunk általános tematikus szerepkészletet átvenni vagy alkotni, hanem témakörönként és kontextusokként definiálunk szereprelációkat (l. a szemantikus lexikonról szóló szekcióban az erről szóló részt). Néhány nyelvi jelenségre külön kidolgoztunk reprezentációs formalizmust, például a tagadásra, a „one of …” jellegű kifejezésekre, a tulajdonságok kifejezésére.

Az igénypont szakaszban a legnagyobb problémát a koordinációk, ill. a felsorolások detektálása jelenti, többször találkozunk egymásba ágyazott felsorolásokkal is. Jelenleg olyan algoritmuson dolgozunk, amelyek a koordinált frázisok hasonlósága alapján rendeli egymás mellé a megfelelő frázisokat. Nemcsak morfológiai, szintaktikai ismérveket veszünk figyelembe, hanem szemantikusakat is. Például tipikusak azok a felsorolások, amelyek valamely szabadalmazandó gyógyhatású készítmény összetételét adják meg, ilyenkor anyagmennyiségek vannak megadva.

A szintaktikus elemzés nemcsak párhuzamosan működik a szemantikussal, hanem párhuzamosan is fejlesztjük. Ezzel elkerüljük, hogy olyan problémába ütközzünk, mint amilyenről [2] beszámol, tudniillik, hogy a MEDIA esetében az elkészült HSPG nyelvtanhoz problémás hozzáilleszteni egy szereprelációkra alapozott jelentésreprezentációt.

3.2 Szemantikus lexikon

Ennek megfelelően a szemantikus lexikonunkban is a szintaktikus és szemantikus információk párhuzamosan lesznek elrendezve, például a vonzatkeretekkel együtt a megfelelő tematikus szerepek. A szemantikus lexikon kulcsfontosságú az elemzéshez. Mint írtuk, nem óriás ontológiát akarunk építeni vagy kölcsönözni. E helyett alkalmazunk egy általános csúcsontológiát (lényegében a DOLCE-ből [6] kölcsönözve), és ehhez kapcsolódnak témakörönként és kontextusokként szigetszerű ontológiák. Az ontológiák osztályai alatt szinonimahalmazok lesznek. Így egy háromrétegű lexikont kapunk, ahol a nyelvi elemek képezik a nagy tömegű információt, a felettük lévő ontológia pedig definiálja azokat az osztályokat, amelyekbe a szinonimahalmazok tartoznak, illetve meghatározza azokat a relációkat, amelyek szerepelhetnek a jelentésreprezentációban.

[2] beszámol arról, hogy a japán fejlesztésű MEDIE továbbfejlesztése is a szereprelációk bevonásával történik, azonban ők egy általános szerepreláció-készletet kívánnak alkalmazni. Mi célszerűbbnek találjuk több, de egyszerűbb szerepreláció-

készletet alkalmazni. Például a *kezel/treat* igéhez nemcsak más vonzatok társulnak, ha gyógyászati készítmények alkalmazásának témakörében használjuk (*treating a patient with a disease* vagy *treating a disease in a patient*[2]), vagy az előállításukban (*treating something with a material*), hanem más szereprelációk is. A *with* prepozíció az első esetben egy „kedvezőtlen állapot" szerepet játszó fogalmat kapcsol az eseményhez, a második esetben pedig „eszköz"-t. A példából az is látszik, hogy gyakorlati, alkalmazási szempontból szabadon eltérünk a nyelvészetben használt tematikus szerepektől, – ez is azt teszi lehetővé, hogy a szemantikus lexikon szerkezetét a második réteg kontextusok szerint is tagolja.

A szinonima fogalmát tágabban értelmezzük, mint szokásos: nem a kifejezések felcserélhetősége az ismérv, hanem az, hogy azonos szituációt/objektumot írnak-e le. Például a *kap* és *ad* szinonim lesz, a vonzatkeret különbözőségét a szereprelációk egyenlítik ki. Ebből következően a szavakat a vonzataikkal együtt kell szerepeltetni; a párhuzamos szintaktikai elemzés miatt a vonzathoz a nekik aktuálisan megfelelő szereprelációkat is hozzá kell rendelni. Sőt, amikor a vonzatok csak bizonyos osztályból kerülhetnek ki, ezeket is.

Nemcsak a szinonim kifejezések lesznek illeszthetőek, hanem azok is, amelyek valamilyen módon implikálják a jelentésreprezentációban szereplőt. Ilyen implikációs viszony a *fajtája* reláció (például az *ékszer* szóhoz illeszthető a *gyűrű*), de nem csak ez. Ilyen a *szükségszerűen következik* reláció is – például, ha a kereső kifejezésben az *érintkezik* ige szerepel, az *irritál* illeszthető hozzá. Természetesen tagadás esetén a szükségszerű következményen alapuló implikációs viszonyok megfordulnak. Tehát a szinonimahalmazok mind a fajtája, mind a szükségszerű következmény relációk szerint rendezve vannak.

3.3 Keresés

A kulcsszó szerinti keresés eredményéül kapott dokumentumokon folyik a szemantikus keresés. Kijelöltetnek azok a szövegszakaszok, amelyekben kulcsszavak szerepelnek, és ezekre kísérli meg rendszerünk a keresőkifejezés jelentésreprezentációjának illesztését.

A keresőkifejezés jelentésreprezentációjának illesztése elvileg háromféle módon hajtható végre:

- generálható a kijelölt szövegszegmens jelentésreprezentációja, és hasonlóságot keresünk a keresőkifejezés jelentésreprezentációjával;
- a szövegszegmenst csak szintaktikusan elemezzük, és a szemantikus lexikon segítségével az algoritmus azt állapítja meg, hogy a szöveg kifejezései és a köztük lévő szemantikus reláció illenek-e a jelentésreprezentációra;
- a szövegszegmens elemzését a keresőkifejezés jelentésreprezentációja vezérli egy rekurzív algoritmussal.

[2] Tisztán pragmatikus okokból a fenti frázisokban a *with* és *in* prepozíciókkal jelzett vonzatokat az igéhez kötjük, nem a főnevekhez.

Az első megoldás nyilvánvalóan pazarló. A harmadik változatot választjuk, bár lehetséges, hogy a szabadalmak igénypont szakasz közti keresés esetén a második változatot célszerű használni.

A találatokat relevancia-sorrendbe rendezzük pontosságuk szerint. Négy nagy osztályt szándékozunk megkülönböztetni:

- teljes találat,
- részleges találat,
- csak kulcsszó szerinti találat,
- ellentmondásos.

3.4 Az archívum feldolgozása

Mint az 1. ábra mutatja, a tulajdonképpeni keresési feladatot – annak megkönnyebbítése érdekében – kiegészítettük az archívum feldolgozásával. Ez két tevékenységet takar: a dokumentumok tematikus klaszterezését és osztályozását és a szakterületekre jellemző szinonimaosztályok generálását.

Több klaszterezési algoritmust kipróbáltunk. Választásunk a Cluto g1p módszerre esett, amely kísérleteinkben meglehetősen pontosnak bizonyult. A kapott eredmények: precision 89,4%, recall 99,1%, f-measure 94%.

A szinonimagenerálás során a mondatokból kiválogatott minták összehasonlítása alapján (kölcsönösinformáció-nyereség) keresünk "szemantikusan" hasonló főneveket. Igaznak bizonyult az a feltevés, hogy sokszor nem szinonim szavakat talál meg az algoritmus, hanem antonímákat, illetve olyan klasztereket, amelyekben hasonló szerepű fogalmak vannak (pl. egyesülés, bomlás, vegyülés, feloldódás). Az azonban a mi esetünkben nem baj, ha a szokottnál lazább szinonimafogalommal dolgozunk. A kísérletezés még kezdeti fázisban van, később dől el, hogyan vezérelhetjük a tanulást, illetve milyen mértékben van szükség emberi kontrollra.

4 A felhasználási területekről

A projekt két felhasználási területet vállalt fel: a szabadalmi keresést és a néprajzi információkeresést. Többé-kevésbé vakon választottuk ezt a két területet, azaz nem jól átgondolt szakmai érvek döntöttek. Azonban sikerült két olyan területet találni, amelyek a lehető legnagyobb mértékben különböznek egymástól[3]. Míg a szabadalmi keresés nagy múlttal, általánosan használt keresőrendszerrel, technológiával rendelkezik, tematikailag nagyon részletesen osztályozottak a dokumentumok, addig a néprajzi területen alapvető eszközök hiányoznak – elsősorban Magyarországon. Míg a szabadalmak legfontosabb része, az igénypont szakasz, félformális szövegnek tekinthető, a néprajzi gyűjtések feldolgozásához a szöveg normalizálásával kell kezdeni (l. [12]). Ugyanakkor a néprajz és a számítástudomány közös területe lett a narrációk kutatása, azaz a néprajzi szövegekre alkalmazható formális rendszerek kutatása. A

[3] Mind a szabadalmi keresésre, mind a néprajzi témájúra vonatkozó helyzetfelmérés, ill. követelményfeltárás letölthető a www.maszeker.hu honlapról.

célok is különbözőek: a szabadalmi kutatásban a szabadalmi bejelentéshez hasonló tartalmú dokumentumot kell keresni[4], a néprajzban motívumok, típusok szerint kell keresni. Igaz, ez utóbbiak definiálása is kutatási feladat.

A fenti különbségekből adottan az általunk fejlesztett technológia különböző módon lesz hasznosítva e két területen.

- A szabadalmi keresés területén az Európai Szabadalmi Hivatal (EPO) által rendelkezésre bocsátott speciális kereső programot (EPOQUENet) használnak. Ez természetesen kulcsszavak szerint keres. A szemantikus keresést végző modul az EPOQUNet találataiból alkotott archívumon fog működni. Képes lesz az igénypont szakaszt teljesen feldolgozni. Arra nem vállalkozunk, hogy következtetésekkel megállapítsuk a talált dokumentum viszonyát a benyújtotthoz[5], – ez mérhetetlen nagyságú és komplexitású világtudást kívánna meg. Azonban súlyt fektetünk arra, hogy a szabadalmak szövegét, ill. találatainkat strukturáltan jelenítsük meg, hogy a keresőt segítse annak áttekintésében.
- A néprajznál viszont maga a korpusz összeállítása is feladat, jelenleg magyar nyelvű hiedelem-, táltosszöveg és mesegyűjteményünk van, amely nyelvészeti feldolgozása megtörtént (l. [12] [13]). A néprajzos kutatóknak már az is nagy eredménynek számít, hogy kollokációkereső programot tudnak futtatni az anyagon (motívumkeresés). Most úgy látjuk, hogy a néprajzi keresésnél a legfontosabb annak feldolgozása lesz, hogy az egyes motívumok milyen hierarchiát alkotnak (pl. a *segítő* lehet *segítő állat*, vagy még specifikusabban *segítő kutya*), és az, hogy milyen megfogalmazásokból lehet következtetni ezek előfordulására. Például a *varázstárgyat ad* jelentésű frázisok alanya *segítő*.

A szemantikus keresés technológiájának kidolgozásánál a szabadalmi keresésre koncentrálunk, a néprajzi keresésnél a már kifejlesztett technológiát alkalmazzuk. Viszont a tematikus osztályozó modult a néprajzi anyagokon teszteljük, és szerepe a néprajzi információkeresésben lesz.

5 A projekt állása

Ebben az évben egy 0. prototípus kerül megvalósításra, amely a fontos funkciókat végrehajtja, de még az algoritmusok finomhangolása nem történik meg – azaz számos ritkábban előforduló nyelvi fordulattal nem fog megbirkózni. Hasonlóképpen a szemantikus lexikon sem a végleges szerkezetben fog rendelkezésre állni, s csak korlátozott tartalommal.

A 0. prototípus kifejlesztése nemcsak azt a célt szolgálja, hogy az algoritmusainkat teszteljük és finomítsuk, hanem azt is, hogy a jövendő felhasználókkal – jelen esetben a szabadalmi hivatal munkatársaival és a néprajzi korpuszokat feldolgozó munkatársakkal egyeztessük a keresés működését, a keresőkifejezés megadási módjait és az eredmény bemutatását. Ugyanis nemcsak magát a keresés technológiáját dolgozzuk ki, hanem olyan felhasználói interfész felületeket, amelyek a szemantikus kereséshez

[4] Nagyon elnagyolt leírás, vannak különböző, de lényegileg ehhez hasonló keresési feladatok is.

[5] A szabadalmak elbírálásánál ennek több fokozatát definiálták.

illenek. Különösen izgalmas probléma megmutatni az egyes találatoknál azt, hogyan illik a dokumentum szövege a találatra. Erre a funkcióra a szöveg grafikus megjelenítését tervezzük.

A jövő évre tervezünk egy fejlettebb prototípus változatot, amely már teljes fegyvertárral mutatja be a kifejlesztésre kerülő technológiát.

Bibliográfia

1. Abolhassani, H., Esmaili K. S.: A categorization scheme for semantic web search engines. In: 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06) (2006)
2. Ananiadou, S., Thompson, P., Nawaz, R.: Improving Search through Event-based Biomedical Text Mining. In: Darányi, S., Lendvai, P. (szerk.): Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts (2010) 42–54
3. Dahlgren, K.: Technical overview of Cognition's semantic NLP (as applied to search). Technical report, Cognition Technologies, Inc. (2007) http://www.cognition.com/pdfs/Cognition_Semantic_NLP_for_Search_Overview.pdf
4. Dietze, H., Schroeder, M.: GoWeb: A semantic search engine for the life science web. In: Burger, A., Paschke, A., Romano, A., Splendiani, A. (szerk.): Proceedings of the Intl. Workshop Semantic Web Applications and Tools for the Life Sciences SWAT4LS. Edinburgh (2008)
5. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.(szerk.): Handbook of Latent Semantic Analysis. University of Colorado Institute of Cognitive Science Series, Psychology Press (2007)
6. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: WonderWeb Deliverable D18: Ontology Library (2001)
7. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka Y., Yosida K., Ninomiya T., Tsujii J.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In: Annual Meeting - Association for Computational Linguistics (2006) 1017–1024
8. Nirenburg, S.: Homer, the author of the Iliad and the computational linguistic turn. In: Words and Intelligence II. Springer (2007)
9. Nirenburg, S., Raskin, V.: Ontological Semantics. The MIT Press (2004)
10. Montalbano, E.: Microsoft testing Kumo search engine internally. NetworkWorld, March 3, 2009. WWW document. http://www.networkworld.com/news/2009/030309-microsoft-testing-kumo-search-engine.html (accessed March 27, 2009)
11. Parsons, T.: Events in the Semantics of English: A Study in Subatomic Semantics. MIT Press, Cambridge (1990)
12. Szauter D., Vincze V., Almási A., Alexin Z., Kiss M.: Morfoszintaktikailag annotált néprajzi korpusz. In: Tanács, A., Szauter, D., Vincze, V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2009)
13. Szőts, M., Darányi, S., Alexin, Z., Vincze, V., Almási, A.: Semantic Processing of a Hungarian Ethnographic Corpus. In: Darányi, S., Lendvai, P. (szerk.): Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts (2010) 112–115
14. UpTake under the hood—the Interview. Alt-SearchEngines, May 14, 2008. WWW document. http://www.altsearchengines.com/2008/05/14/uptake-under-thehood-exclusive-interview/ (accessed March 27, 2008)

Nyelvészeti problémák a szabadalmak feldolgozásában

Vincze Veronika[1], Nagy Ágoston[1], Klausz Ágnes[1], Almási Attila[1], Kiss Márton[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{vinczev, nagyagoston, aklausz, mkiss}@inf.u-szeged.hu,
vizipal@gmail.com

Kivonat: A szabadalmak számos olyan sajátossággal bírnak, amelyek azok nyelvi elemzését – az általános tématerületű szövegekhez képest – jelentősen megnehezítik. Szintaktikailag bonyolult felépítésű szerkezetek, beágyazott mondatok, összetételek és felsorolások szép számmal találhatók bennük, igen sok bennük a visszautalás (anafora), és az elliptikus tagmondatok, vonatkozó mellékmondatok és utómódosítók használata is jellemző. A szabadalmak szókincse is jellegzetes: a terminus technicusokon kívül bizonyos szófordulatok jelenléte is tipikusnak mondható. Mindezen jellemzőkből adódó problémák kezelésére különféle szabályalapú módszereket dolgoztunk ki, melyeket az előadásban ismertetünk.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg, ugyanakkor a készülő rendszer könnyen adaptálható lesz más területekre is. Mivel a szabadalmak rendkívül sok tudományterületet fednek le, melyek mindegyike sajátos jellemzőkkel bír (mind stilisztikai, mind terminológiai szempontból, mind pedig a szabadalmak felépítését tekintve), a projekt keretein belül egy adott osztályozási jelzettel ellátott szabadalmak feldolgozására összpontosítunk, nevezetesen az A61K (orvostudományi) osztályra.

A szabadalmak számos olyan sajátossággal bírnak, amelyek azok nyelvi elemzését – az általános tématerületű szövegekhez képest – jelentősen megnehezítik. Az előadásban e sajátosságokat, az ezekből adódó problémákat és a rájuk adott megoldásokat ismertetjük.

2 A szabadalmak felépítése

A szabadalmak egységes szerkezettel bírnak. A címlap tartalmazza az úgynevezett bibliográfiai adatokat, amelyben megtalálható többek között a szabadalom iktatási száma, a benyújtás időpontja, a szerzők és a feltalálók neve. Az első oldalon szerepel még a találmány néhány soros összefoglalója, amelyet ábrákkal is ki lehet egészíteni.

Itt található a cím is, amely meghatározza a találmány tárgyát, majd a leíró részben annak pontos jellemzőit fejtik ki a szerzők különös tekintettel a találmánnyal megoldandó feladatra, az alkalmazási területekre, példákkal, ábrákkal, táblázatokkal szemléltetve. Az igénypontok pedig a szabadalmak oltalmi körét határozzák meg, azaz azt, hogy mit szeretnének a feltalálók levédetni.

A találmányt az úgynevezett főigénypont azonosítja a legáltalánosabban. A főigénypontban megtalálható a találmánynak a célul kitűzött feladat megoldásához elengedhetetlenül szükséges minden jellemzője (l. [7]). Emiatt a továbbiakban elsődlegesen a főigénypontok nyelvi feldolgozására összpontosítunk.

A főigénypont szerkezete eléggé kötött. Ez már abból is adódik, hogy a főigénypont hossza csak egy mondat lehet: a legtöbb problémának ez a forrása, mert mindent ebbe az egy mondatba próbálnak beletömöríteni. A főigénypont mindig azzal kezdődik, hogy milyen kategóriába tartózik a levédetni kívánt szabadalom, például módszer, eljárás, eszköz, összetétel. Eztán következik ezek kifejtése: milyen lépésből/anyagokból áll a főigénypont elején említett dolog, és ezeket az alpontokat rekurzívan továbbfejtik.

3 A szabadalmak nyelvi jellemzői

Mint már említettük, a szabadalmak terminológiai és stilisztikai szempontból is eltérnek az általános doménből vett szövegektől. Mind a magyar, mind az angol szabadalmakra jellemző, hogy nyelvezetük tömör, lényegre törő. Szintaktikailag bonyolult felépítésű szerkezetek, beágyazott mondatok, összetételek és felsorolások szép számmal találhatók bennük. A megfogalmazásban pontosságra törekednek a szerzők, igyekeznek kimerítő leírást adni a találmányról, ugyanakkor megfigyelhető az a tendencia is, hogy – az esetleges későbbi jogviták elkerülése végett – bizonyos általánosító stratégiákat alkalmaznak, így lehetővé válik a jellemzők és az alkalmazási területek bővítése, illetve a későbbiekben esetleg relevánssá váló esetek hozzáadása („beleértése" a szabadalomba) [7]. Ilyen nyelvi stratégiára hozunk néhány példát:

- a kimerítőnek látszó felsorolások végén szereplő *stb*.;
- a felsorolások előtt szereplő *pl.* vagy *például*;
- megengedő *vagy* használata;
- általános jelentéstartalmú határozók használata (*rendszerint*, *általában*).

E stratégiák némileg párhuzamot mutatnak a bizonytalanságot jelölő kifejezésekkel (angol terminológiával élve a hedge, illetve weasel kifejezésekkel [2]), míg azonban például a Wikipédia szócikkein belül ezen általánosító, kétértelmű és nem kimerítő leírást adó kifejezések használata nemkívánatosnak minősül, addig a szabadalmak nyelvezetében a fenti okok miatt ez teljességgel megszokott stratégia.

Mivel a főigénypontnak tartalmaznia kell minden szükséges, a szabadalom lényegét érintő információt, továbbá a hagyományoknak megfelelően a főigénypont egyetlen mondatból áll, ezért nem várható el, hogy a főigénypontot egy egyszerű, könnyen feldolgozható mondat alkossa [7]. Szintaktikai szempontból jellemezve a mondatokat elmondhatjuk, hogy igen hosszú, többszörösen összetett mondatok alkotják a szaba-

dalmak szövegét – egy-egy főigénypont (azaz egy mondat) akár több oldal hosszúságú is lehet. Ebből adódóan igen sok bennük a visszautalás (anafora), és az elliptikus tagmondatok, felsorolások, vonatkozó mellékmondatok és utómódosítók használata is jellemző. A mondatok pontos szintaktikai elemzését a fentiek mellett az is nehezíti, hogy a központozás nem túl következetes. A fentiek miatt [7] szerint a szabadalmak külön nyelvtannal (szintaxissal) bírnak, mely nem esik egybe a(z angol) nyelvtannal.

A szabadalmak szókincse is jellegzetes: a terminus technicusokon kívül bizonyos szófordulatok (*azzal jellemezve*) jelenléte is tipikusnak mondható, melyek nem feltétlenül találhatók meg egy általános célú szótárban, így ezeket külön fel kell venni, illetve a kezelésükre külön szabályokat kell írni. A szabadalmak értelmezését az is megnehezítheti, hogy – mivel a leírt találmány új – a találmány leírására használt szavak is új értelmezésben használtatnak a szabadalomban [7].

4 Nyelvi problémák a szabadalmakban

A szabadalmak nyelvi sajátosságaiból adódó, az általános doménre felkészített nyelvi elemzők számára [5] problémát jelentő esetek a következők:

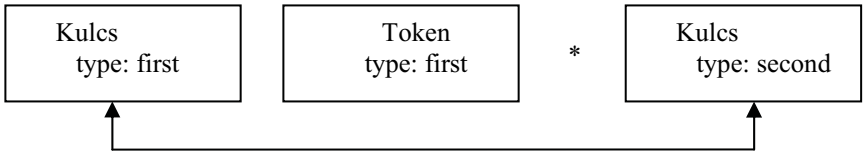
- rendkívül hosszú mondatok (kulcsok és utómódosítók)
- adjunktumok
- sajátos fordulatok
- összetételek
- felsorolások
- kvantitatív szerkezetek
- kémiai névelemek

A fenti problémák kezelésére különféle szabályalapú módszereket dolgoztunk ki, melyeket az alábbiakban ismertetünk részletesen.

4.1 Kulcsok

Egy szabadalom főigénypontja általában egy többszörösen összetett, nagyon nehezen elemezhető mondat sok alá- és mellérendeléssel. Ezeknek a nem ritkán több mint száz szavas mondatoknak a gépi elemzése a jelenlegi elemzők segítségével nem lehetséges. Olyan megoldást kellett találnunk, amely segítségével e mondatokat olyan elemi mondattöredékekre tudjuk bontani, melyek elemezhetőek gépi algoritmusok segítségével. Ezért az utómódosítók, valamint a mellékmondatok kezdetét **kulcs**okkal jeleztük.

Kulcs alatt általánosan a feldolgozott szövegnek azokat a szakaszait értjük, ahol a módosító-módosított főnév viszony jelenléte *pusztán formai alapon* felismerhető. A kulcsok egy első és egy második részből épülnek fel.



1. ábra. A kulcsok felépítése.

- **Egyszerű kulcs:** Az egymást követő kulcsok jelölésére szolgál abban az esetben, ha a kulcs első részéhez nem kapcsolódik távoli második típusú kulcs. Például: *substance which*, *group consisting*.
- **Összetett kulcs:** Összetett kulcsról beszélünk, ha a kulcs első és második tagja nem közvetlenül követi egymást, vagy a kulcs első részéhez több második rész is tartozik. Például: *the **process comprising** the steps of deforming the films (18) to form a multiplicity of recesses (16), **filling** the recesses*.
- **Beágyazott kulcs:** Minden olyan esetben alkalmazandó, ahol nem érvényesíthető a következő szabály: „Összetett kulcs második részét mindig az előtte álló összetett kulcs első részéhez kell kötni”. A beágyazott kulcsok egymással sorfolytonosan balról jobbra, kettesével kötendők össze és feldolgozásuk megelőzi az összetett kulcsét. Például: *A **method** for the treatment of systemic infection **diseases**, such as pneumonia, tuberculosis, peritonitis, endocarditis, pyelonephritis, meningitis or septicemia, **caused** by bacterial or protozoal infection, **comprising**:*.

A kulcsokat két osztályba soroljuk felismerhetőségük alapján:

1. Egymást követő kulcs, ezen kulcspárok egésze (első és második részük is) egyből felismerhető. A következő esetekben tekinthető kulcsnak két egymást követő token (a lenti felsorolásban a Stanford szófaji egyértelműsítő [5] jelölésrendszerét használjuk):

   - N + postModifier
   - N + to + VB/VBP
   - N + JJ + Prep
   - N + (WDT|WP|WP$)

2. Csak az elemzés későbbi részében felismerhető kulcspárok, ezen kulcspároknál csak a kulcs második része ismerhető föl pusztán formai jelek alapján. E kulcsok első része az elemzés későbbi részében ismerhető föl, illetve keresendő meg. A következő esetekben tekinthető kulcsnak (kulcs második részének) egy token:

   - **whose**

- **which,** ha előtte **,** vagy **;** van, vagy **and** tokenek állnak
- Minden VBN szófaji kóddal rendelkező token, ha megelőzi egy **,** vagy **;**
- A következő szavak: **comprising|having|consisting|being|including ,** ha megelőzi őket egy **,** vagy **;** vagy az **and**

4.2 Adjunktumok

A köznyelvhez képest szerencsére igen kevés az adjunktumok száma a szabadalmak igen kötött nyelvezetének köszönhetően (csak azt mondják, ami feltétlenül szükséges, azt viszont pontosan). Néhány esetben azonban különös figyelmet igényelt az adjunktumok kezelése.

Az *optionally* gyakorlatilag *vagy*-szerű logikai operátorként viselkedik (valami vagy megtörténik, vagy nem), ezért a szemantikai elemzés során erre hangsúlyt kell fektetni. Egy példa:

*C.sub.6-C.sub.10-arylthio which is **optionally** substituted by nitro, amino, C.sub.1-C.sub.6-alkyl or C.sub.1-C.sub.4-alkoxy*

A példában a *C.sub.6-C.sub.10-arylthio* helyett állhat vagy *nitro*, vagy *amino*, vagy *C.sub.1-C.sub.6-alkyl* vagy *C.sub.1-C.sub.4-alkoxy*.

Egy másik lehetséges problémaforrás, hogy a szabad határozó néha az ige és a vonzata között helyezkedik el:

*consisting **essentially** of a purified mineral composition and optional excipients*

Ez a vonzatkeret illesztése miatt okozhat problémákat, de néhány szabály segítségével áthidalható, szemantikai szinten pedig az ilyen módon az igéhez kapcsolódó legtöbb határozó jelentése elhanyagolható a mondat szempontjából.

A PP-bővítmények (*during a sport activity, without a tabletting excipient*…) vagy az előtte levő NP részei (ill. a főnévi fej bővítményei), vagy pedig az igéhez kapcsolódnak. Ennek eldöntése igen nehéz, sokszor még az ember számára sem egyértelmű. A főnevekhez készítendő vonzatkerettárat kellett ilyen esetekben segítségül hívni (ha a főnévi fejhez egy adott prepozíciót tartalmazó PP kapcsolódik, akkor a főnév bővítményeként kezeljük, ha nem, akkor az igéhez tartozóként), vö. [4].

Bizonyos, jelzőket módosító határozószavak (*pharmaceutically, substantially, dermatologically, therapeutically*…) gyakran kollokációszerűen viselkednek:

*a **dermatologically acceptable** carrier*
*a **therapeutically effective** amount of a compound of Formula I*
*a **pharmaceutically acceptable** salt thereof*

Ezeket egységként vettük fel a szótárban.

4.3 Sajátos fordulatok

A szabadalmak szókincsének jellegzetes elemei bizonyos szófordulatok (*said*, *a plurality of*, *azzal jellemezve…*), melyek nem feltétlenül találhatók meg egy általános célú szótárban, így ezeket külön fel kell venni, illetve a kezelésükre külön szabályokat kell írni. Például a fenti *said* jelző anaforikusan utal vissza egy, a szabadalmi igénypont szövegében már korábban megemlített entitásra, így anaforikus elemként érdemes kezelni.

Az *a plurality of* típusú szerkezetek szemantikailag átlátszóak, noha szintaktikailag a *plurality* számít a kifejezés fejének, szemantikai szinten az *of* prepozíció bővítménye játszik csak fontos szerepet:

A vitamin supplement to temporarily enhance the abilities of a individual during a sport activity comprising a plurality of B family vitamins and one or more other vitamins, minerals, and/or natural ingredients.

Ebből következően a mondat szemantikai reprezentációjában az *a plurality of* nyelvi kifejezés nem is szerepel.

Az *azzal jellemezve* típusú szófordulatokat külön elemként szerepeltetjük a szótárban.

4.4 Összetételek

Az elemzés során problémát okozhatnak a halmozott NP-szerkezetek, ezen belül is különösen az előmódosítók. Mint fentebb említettük, a szabadalmi szövegekre kifejezetten jellemző a tömörség, az informativitásra való törekvés, ami – többek között – a rendkívül hosszú mondatokban, szószerkezetekben nyilvánulhat meg. Ráadásul az angol nyelvben a főnévi előmódosítók számának csupán az érthetőségi korlátok szabnak határt. A több, közvetlenül egymás után álló főnév a gépi elemzés során elsősorban szegmentálási problémát jelenthet.

Többek között az N + ADJ + N szerkezetű magNP-k okozhatnak ilyen problémát, mivel a szerkezeti elemzésük többféleképpen történhet. Alapvetően kétféle variáció állhat fenn: a középső elem, azaz a melléknévi alak vagy az előtte álló főnévhez kapcsolódhat szorosabban, vagy az utána állóhoz. Az utóbbi esetben az N + ADJ szerkezetű NP-nek az első főnév az előmódosítója: [N + [ADJ + N]. A gépi elemző általában ezt a szegmentálási variációt használja alapértelmezésként.

Azonban vannak esetek, amikor az N + ADJ + N szerkezet mellékneve – bár szintén az *utána* álló főnév előmódosítója – az *előtte* álló főnévhez szorosabban kapcsolódik, mivel a vele alkotott jelzői módosító feje. (Itt az első főnév az előmódosító előmódosítója):

[[N + ADJ] + N], pl. [[*silicone conditioning*] *oil*].

Ilyen esetekben a szintaktikai elemzés során a melléknév *után* kell részekre bontani az NP-t. (Amennyiben névelő áll a második főnév előtt, egyértelmű, hogy a melléknevet az *előtte* álló főnévhez kell kapcsolni.)

A szóban forgó melléknévi alakok lehetnek *–ing* végződésűek, illetve *past participle* alakúak. Az előbbiek többnyire tárgyas igéből képzett folyamatos melléknévi igenevek, pl. *containing*, vagy tárgyas igéből képzett melléknevek, pl. *(pH-)responsive, (bio-)absorbable,* de lehetnek egyszerű melléknevek is, pl. *(sodium-)free.* A *past participle* alakúak szintén tárgyas igéből képzettek: *(diabetes-)associated, (lipoprotein receptor-)related.*

A fentebbieken kívül kétértelműek lehetnek még az ADJ + ADJ + N szerkezetű szóösszetételek is, amelyeket [ADJ + [ADJ + N]] szerkezetként (pl. *substituted lower alkyl, inorganic metal oxide*) és [[ADJ + ADJ] + N] szerkezetként (*vascular-related diseases*) is lehet értelmezni.

A melléknevet tartalmazó előmódosítókban az első elem lehet számosságra utaló elem is, ami szintén azt a problémát veti fel, hogy hova kapcsoljuk az utána álló melléknevet abban az esetben, ha nincs kötőjel az elemek között, pl. *penta-substituted C1-C12 alkyl, three- to seven-membered alkylene bridge.*

4.5 Felsorolások

Mivel a szabadalmak főigénypontjai egymondatosak lehetnek csak, ezért a szerzők abba az egy mondatba próbálnak mindent belesűríteni. Ez a felsorolások kezelésének tekintetében is sok bonyodalmat okoz. A felsorolásokat formailag viszonylag könnyű felismerni, mert elemeit vessző, pontosvessző vagy kötőszó választja el (habár sok esetben ez hiányzik). A szintaktikai elemzés szempontjából viszont gyakran nehéz eldönteni, hogy a felsorolást elválasztó elemek után található szó vagy szócsoport minek a bővítménye. Ez amiatt történhet meg, hogy a főösszetevők felsorolása mellett párhuzamosan történik meg az azokban található alösszetevők leírása, amelyek szintén tovább bonthatók. Esetenként így akár 3-4 szint mélységű is lehet egy-egy felsorolás. Általában a vesszővel azonos szinten lévő elemeket sorolunk fel, a pontosvesszőt pedig legalább egy szinttel megy feljebb – de a ”legalább egy” és az ”azonos szinten” sajnos nem elég pontos támpont egy parser létrehozása szempontjából, mert kivételek is lehetnek. Erre példa az alábbi szabadalomrészlet:

R1 and R2 are each selected independently from the group consisting of hydrogen, hydroxyl, amino, ..., alkoxy of 1-6 carbon atoms, alkylthio, aryloxy, ...

A fenti példában az tapasztalható, hogy a *consisting* vonzata a *hydrogen, hydroxyl, amino, alkoxy of 1-6 carbon atoms, aryloxy* stb. Ez számunkra teljesen evidens, de a felsorolásokkal kapcsolatban felállított szabályok szerint a parser logikusan az *alkylthio* és az azt követő felsoroláselemeket az *alkoxy* szóhoz köti, pedig valószínűleg azok is a *consisting* szóhoz tartoznak. Az *atoms* utáni vessző tehát nem azonos szintet, hanem egy szinttel feljebb való ugrást feltételez. A problémán itt még az sem segítene, ha minden, felsorolásban található elem előtt megismételjük a prepozíciót, mert itt mindkét esetben az *of* lenne az.

A felsorolások végén található *and* vagy *or* kötőszó pedig azt jelenti, hogy az adott felsorolás utolsó eleme fogja követni. Ez sok esetben igaz, de találtunk egy többszörösen mellérendelt mondatkezdetet is:

A means for allaying drunkenness, preventing and removing alcohol intoxication and hangover syndrome and ***a method*** *for allaying drunkenness, preventing and removing alcohol intoxication and hangover syndrome by using this means, comprising:*

A fenti példában a *removing* utáni felsorolás okoz problémát: a *preventing* és *removing* tárgyas vonzata az *alcohol intoxication* és a *hangover syndrome*. Azonban ezekhez még hozzá van kötve szintén az *and* kötőszóval a *method* is, amely az elemző számára természetesen ugyanolyan, mint az *alcohol intoxication*, így azokhoz köti testvérként. Itt semmi sem jelzi a feljebb ugrást, ami ráadásul kétszintű: nem a *means for* vonzata a *method*, hanem a gyökérhez köthető a *means* mellé.

4.6 Kvantitatív szerkezetek

A biokémiai szabadalmakban fontos szerepük van a mennyiségjelzőknek, amelyek feladata, hogy a főigénypontokban minél pontosabban leírják egy kémiai összetétel összetevőinek pontos mennyiségét. Mivel a főigénypontok a mérvadóak a szabadalmaztatás során, a szerzők nemcsak az előbb említett pontosságra törekednek, hanem arra is, hogy hasonló összetételt se lehessen alkalmazni, így gyakran használnak olyan szerkezeteket, amelyek az összetevők mennyiségét a *körülbelül* előtaggal módosítják. Így a főigénypontokban egyszerre jelenik meg a pontosság igénye, és a mennyiségmegjelölések kis mértékű elhomályosítása (vö. 3. fejezet).

A szabadalmak mennyiségei rögzített szerkezettel rendelkeznek: általában *-tól/-ig* tartományt fejeznek ki, például *from about 1 gram to about 5 grams of Arginine*. Az ilyen típusú mennyiségjelzők szintaktikai szempontból nem okoznak problémát: általában mindegyik egy megadott mintára illeszkedik, így azok kinyerése viszonylag könnyen megoldható. Szemantikai szempontból viszont az ilyen típusú szerkezetek problémát okozhatnak. Ha egy szabadalmi keresőbe beírjuk, hogy olyan összetételeket keresünk, amelyben *0,5 gramm Arginine* található, akkor az beleesik-e a fent említett példába, azaz *a kb. 1 grammtól kb. 5 grammig terjedő* tartományba? A *körülbelül* szónak így meg kell adni egy viszonylag széles tartományt, amelybe biztosan belefér a keresett elem, de felesleges találatokat nem ad. Ennek a problémának a megoldása további fejlesztések eredményeképpen várható.

A mennyiségjelzős szerkezetek esetében a felismerési problémát az okozza legtöbbször, hogy a mennyiséget kifejező tag túl messzire kerül a hozzá tartozó főnévtől, így azok összekötése nehézzé válik. Vannak olyan esetek, amikor csak a *be* ige ragozott alakjai kerülnek be a mennyiségjelző és a hozzá tartozó főnév közé:

the weight ratio of xanthan to guar gum [being] from 1:3 to 1:10
the weight ratio of crystals to carrier [is] 2-99%

Ezen esetekben a *be* elhagyásával a mennyiségjelző könnyen összeköthető. Azonban vannak olyan esetek, ahol a mennyiségjelzők és a hozzájuk tartozó főnevek nagyon messzire elkerülnek egymástól. Az alábbi két példa is ezt szemlélteti:

the sodium bicarbonate being incorporated in the toothpaste in an amount of at least 60% by weight

the ratio of the components is as follows (wt. %): TBL natural minerals 33-62 vegetable stock 34-61 water the balance.

Az első esetben a *legalább 60 tömeg%* a nátrium-bikarbonátra vonatkozik, de közéjük beékelődik még az, hogy ez az arány miben található, nevezetesen a fogkrémben. A második egy elég extrém példa, és szerencsére ritka is. Itt a mértékegység zárójelben kikerül előre, és egy felsorolásban következik utána az összetevők listája, majd azok mennyisége (már mértékegység nélkül). A természetes ásványok tömegszázaléka 33-62, a zöldségéé 34-61, a többi pedig víz. A felsorolásoknál tovább nehezíti a dolgot, hogy ebben az esetben sincs vessző a felsorolások tagjai között.

Gyakori probléma még, hogy a szöveges formátum nem mindig megfelelő: például táblázatokból egyszerű szövegek keletkeznek, a sorok és oszlopok összemosódásával. Ezekben az esetekben a mennyiségeket még nehezebb összekapcsolni a főnévvel. Erre példa az alábbi táblázat, amelynek szöveges változatát alatta közöljük:

particle size	percentage
5 μm or more and less than 100 μm	5 to 30%
100 μm or more and less than 300 μm	10 to 40%
300 μm or more and less than 500 μm	10 to 50%
500 μm or more and less than 1000 μm	balance

particle size percentage 5 μm or more and less than 100 μm 5 to 30% 100 μm or more and less than 300 μm 10 to 40% 300 μm or more and less than 500 μm 10 to 50% 500 μm or more and less than 1000 μm balance

Ebben a példában a részecskeméréthez tartoznak az alatta lévő elemek, és a százalékhoz az abban az oszlopban található mértékek, a folyó szövegben viszont ezt nehéz összepárosítani.

A kvantitatív szerkezetek felismerésében egy másik nagyobb problémát a létező mértékegységek nagy száma jelenti. További probléma, hogy a mértékegységek gyak-ran rövidített alakjukban szerepelnek, melyek igen gyakran csak 1-2 karakterből áll-nak, ami többértelműségekhez vezethet (pl. az mg betűsor – kis- és nagybetűket nem megkülönböztetve – lehet a magnézium vegyjele is és milligramm is, a C pedig lehet Celsius-fok és a szén vegyjele is, vö. [1, 6]).

4.7 A névelemek annotációja során felmerült problémák

A szabadalmak annotálásakor olyan névelemeket kerestünk, amelyek a kémia területéhez tartoznak, és amelyekre a felhasználó nagy valószínűséggel rákereshet. Három kategóriát vettünk fel: 1) kémiai elemek (nitrogén, oxigén), elemcsoportok (halogének, alkáli földfémek), vegyületek (Na_2O, CaO) és egyéb olyan kifejezések, amelyek

az annotáló számára elég specifikusak voltak ahhoz, hogy ebbe a halmazba kerülhessenek; 2) egyéb, biokémiai szempontból fontos kifejezések: pl. általános anyagnevek (ginzeng, cukor, só stb.), vegyületfajták (szénhidrogének) és egyéb olyan kifejezések, amelyek kémiai szempontból keresőkifejezések lehetnek; 3) konkrét betegségek (Alzheimer-kór, tuberkolózis), betegségcsoportok (gyulladásos betegségek, immunhiányos betegségek) és tünetek (másnaposság).

A kifejlesztett NER modul futásának eredménye a következőkre irányította a figyelmet:

1. A program bizonyos esetekben nem különíti el a névelemek főnévi és jelzői használatát, amire példa az *antibiotic* szó, mely az angolban főnévként és melléknévként is szerepelhet, és a szabadalmakban is kétféleképpen fordul elő (vö. *an* ***antibiotic*** *medication – a total amount of* ***antibiotic*** *and antihistamine*). Az annotálás során a főnévi szerepben lévő elemeket jelöltük.

2. Az annotálás első körében úgy jártunk el, hogy csak azokat az elemeket vettük fel NE-nek, amelyek valamely képlettel (egyértelműen) azonosíthatók voltak. Így fordult elő pl. az anyagnevek esetében, hogy egy adott alakban előforduló szót egyszer NE-nek jelöltünk, más esetben viszont nem. Erre a legjobb példa az *alcohol* szó, mely egyes szabadalmakban valamilyen kémiai szempontból jól beazonosítható vegyület részét képezi (*cetylstearyl alcohol*), máskor viszont csupán mint szeszesital szerepel (pl. az *alcohol intoxication*ben).

A szabadalmakban való keresés és az annotálási elvek nagyobb fokú összehangolása érdekében a jelölési elveket módosítottuk, két kémiainévelem-kategóriát vettünk fel (lásd fentebb), s így az *alcohol*t már minden esetben jelöltük.

3. Többször előfordult, hogy a program – pl. a szabadalmakban előforduló helyesírási hibák miatt – nem megfelelően szegmentált bizonyos elemeket (pl. *…alkarylamino, fluoro, chloro,* ***bromo iodo*** *and trifluoromethyl…*), ezért két, egyébként különálló NE-t egynek tekintett. Ezekben az esetekben a jelölést a valós tartalomból kiindulva (és a nyelvhelyességnek megfelelően) végeztük el.

4. Szófaji problémák:

a) A program minden olyan elemet, amely a szótárjában NE-ként szerepel, alkalmas jelöltnek tekint és kiemel. Pl. a ***water****-soluble*, ***sodium****-free*, ***wax****-like* (vízoldékony, nátriummentes, viasszerű) kifejezések a magyarban egyértelműen nem számítanak névelemnek, második tagjuk pedig az úgynevezett HALFLEX melléknevek közé tartozik [8]. A program úgy jár el, hogy ha talál NE-t, és az kötőjellel kapcsolódik egy másik elemhez, akkor az NE határát kiterjeszti, és annak részeként kezeli a kapcsolódó elemet is, ami ezekben az esetekben nem megfelelő eljárás. A kézi annotálás során ezeket az elemeket nem jelöltük.

b) Egy másik esete annak, hogy a program NE-ként jelöl meg bizonyos, egyébként nem jelölendő elemeket pl. a *carboxylic* és az *enantiomeric* jelzők, amelyekben szerepel egy-egy, a szótárprogramba felvett NE, a *carboxyl* vagy az *enantiomer*, de ami-

att, hogy a program kiterjesztéses elven működik, a teljes kifejezést NE-nek jelöli. Az annotálás során ezeket az elemeket nem jelöltük.

c) Harmadik példa a nem megfelelő jelölésre az *O-glycosidically*. A szótárprogram a nagy *O*-t NE-ként kezeli, és mivel az a) ponthoz hasonlóan, kötőjellel kapcsolódik az utána következő taghoz, a kettőt egy NE-nek veszi, ami szintén nem megfelelő, mivel a teljes kifejezés egy határozószó. A kifejezés itt sem lett megjelölve.

5 A korpusz

A nyelvészeti problémák feltárásához és a kidolgozott algoritmusok és módszerek ellenőrzéséhez nélkülözhetetlen volt összeállítanunk és kézzel annotálnunk egy korpuszt. A korpusz 313 szabadalmat tartalmaz az IPC osztályozási rendszer A61K besorolású szabadalmai közül. Mivel a kutatás jelen fázisában a szabadalmak fő igénypontjait tanulmányozzuk így ezekben jelöltük be kézzel az alábbiakat: 1) kvantitatív szerkezetek mintái; 2) perdurant jelentésű kifejezések; 3) kulcsok; 4) kémiai névelemek és 5) felsorolások és felsorolásjelzők.

A korpuszon az annotálás Microsoft Wordben történt, majd e dokumentumokat konvertáltuk TXT-be és az annotációkat pedig UIMA-ba [3]. Így könnyen elemezhettük és felhasználhattuk a kézzel jelölt korpuszt.

6 Eredmények

A kulcsok felismerésére létrehozott program működésének kidolgozásához, valamint a program ellenőrzésére egy 60 szabadalomból álló korpuszban jelöltük be kézzel a kulcsokat. A mintakorpusszal összehasonlítva a kulcsok azonosítására kidolgozott eljárást az alábbi mérőszámokat kaptuk.

1. táblázat: A kulcsok felismerésének eredményei.

	Pontosság	**Fedés**	**F-mérték**
Kulcsok megszorítás nélkül (teljes kulcs):	75.47%	75.59%	75.53%
Csak a kulcs első része:	70.61%	71.09%	70.85%
Csak a kulcs második része:	78.27 %	78.042 %	78.16%

A fenti értékekből is látszik, hogy az algoritmus a kulcsok első felének detektálásakor hibázik többet, míg a kulcsok második felét valamivel jobban képes detektálni. A kapott értékek növelése egy bizonyos szintig megoldható további szabályok bevezetésével. További eredményeink: a kémiai névelemek felismerésében 95,25%-os F-mértéket, míg a magNP-k azonosításában 92,59%-os F-mértéket értünk el.

7 Összegzés

A tanulmányban bemutattuk a szabadalmak nyelvi sajátosságait és az azokból fakadó elemzési problémákat. Utóbbiakra számos szabályalapú megoldást dolgoztunk ki, melyek segítségével az elemző algoritmusunk mind pontosság, mind fedés terén (azaz F-mértéket tekintve is) számottevő javulást mutatott. A jövőben az algoritmus további tökéletesítése, illetve a most még nem megoldott problémák (pl. felsorolások) kielégítő kezelése a célunk.

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Agatonovic, M., Aswani, N., Bontcheva, K., Cunningham, H., Heitz, T., Li, Y., Roberts, I., Tablan, V.: Large-scale, Parallel Automatic Patent Annotation. In: Proceedings of 1st International CIKM Workshop on Patent Information Retrieval - PaIR'08. Napa Valley, California, USA (2008)
2. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Uppsala (2010) 1–12
3. Kiss M., Nagy Á.: Egy nyelvészeti UIMA folyamat a kézi annotálástól az eredmények megjelenítéséig. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 362–364
4. Klausz Á., Vincze V., Nagy Á., Almási A.: Vonzatkeretek vizsgálata orvostudományi tárgyú, angol nyelvű szabadalmi szövegeken. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 180–189
5. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
6. Nyilas S., Németh G., Almási A.: Szótáralapú kémiai NE-felismerő rendszer. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 379–383
7. Osenga, K.: Linguistics and patent claim construction. Rutgers Law Journal Vol. 38, No. 61 (2006) 61–108
8. Vincze V., Lucza M., Csendes D., Kiss G: Szótárazási dilemmák a MetaMorpho magyar-angol fordítóprogram névszói adatbázisának építésében. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 180–189

Vonzatkeretek vizsgálata orvostudományi tárgyú, angol nyelvű szabadalmi szövegeken

Klausz Ágnes, Vincze Veronika, Nagy Ágoston, Almási Attila

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{aklausz, vinczev, nagyagoston}@inf.u-szeged.hu,
vizipal@gmail.com

Kivonat: Orvostudományi tárgyú, angol nyelvű szabadalmi szövegekben előforduló igék s főnevek vonzatkereteit vizsgáltuk. Az előfordulási gyakoriságuk alapján összeállítottunk egy kifejezetten az orvostudományi tárgyú szabadalmi szövegekre jellemző vonzatkerettárat, amely hasznosítható a hasonló tárgyú szövegekre alkalmazandó szintaktikai és szemantikai elemzők építésében.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg. A rendszer kialakításához a szabadalmi szövegek sajátságai miatt a meglevő nyelvi elemzők testre szabása szükséges, ezért célunk volt egy olyan igei és főnévi vonzatkerettár kialakítása, melyet a későbbiek során egyéb, orvostudományi tárgyú (szabadalmi) szövegek elemzéséhez is fel tudunk használni mind szintaktikai, mind szemantikai szinten.

Megvizsgáltuk, hogy a különféle igék a különféle vonzatkereteikkel milyen gyakran fordulnak elő ezen orvostudományi szakszövegekben. Az eredményt egy általános célú szótárban (az online Google Dictionaryben [2]) található igékkel és vonzatkereteikkel hasonlítottuk össze. Arra voltunk kíváncsiak, hogy a szótárban található igék és vonzatkereteik mennyire fedik le a 60 szabadalomból álló mintakorpuszunkban szereplőket, azaz egy általános célú szótár vonzatkeretei mennyire alkalmazhatók egy speciális tematikájú szövegre.

2 Igék (és vonzatkereteik) kigyűjtése a szabadalmakból, illetve egy általános célú szótárból

Ebben a részben az igék és a vonzatkeretek kigyűjtésének lépéseit ismertetjük a szabadalmi szövegekből és a rendelkezésre álló szótárállományból.

2.1 A gépileg beazonosított igék kézi ellenőrzése

Első lépésként – a Stanford elemzőt [6] használva – gépileg beazonosítottuk a szabadalmi szövegekben az igéket, majd az igének minősített elemeket kézzel is ellenőriztük, amire több okból is szükség volt. Egyrészt a POS-tagger időnként olyan szóalakokat is igének jelölt, amelyeknek egyik lehetséges szófaji kódja valóban ige, azonban az adott szövegkörnyezetben más szófajú szóként fordultak elő. Másrészt pedig arra is volt példa, hogy a gépileg megtalált szavak ugyan igei alakban fordultak elő, azonban főnévi vagy melléknévi szerepük volt. Főnévi szerepben az ún. *gerund*ként (magyarra *-ás*, *-és* végű főnévként fordítandó, angolban *-ing* végződésű igealakként) fordultak elő, s állhattak alanyként, tárgyként (pl. *a method comprising administering a pharmaceutical composition*), és esetenként határozóként is (pl. *a method for inhibiting thrombosis*, *capable of reducing lung volume*). Melléknévként szerepelhettek a főnévi szerkezet előmódosítójaként – egyrészt *-ing*-es alakban (folyamatos melléknévi igenévként, pl. *a protecting group*), másrészt az ige 3. alakjának formájában (past participle, pl. *protected amino group*, *alkylene-substituted amino*).

A fentebb említett különböző elemek esetében el kellett döntenünk, hogy igeként kezeljük-e őket. Figyelembe véve a szemantikai és szintaktikai sajátosságaikat, különböző módokon jártunk el. Mivel a gerundnak például egyaránt van főnévi és igei jellege is, szóba jöhetett az igeként történő kezelése. És emellett is döntöttünk, hiszen a gerund alakok automatikusan öröklik annak az igének a vonzatkereteit, amelyekből képezve lettek, tehát egy igei vonzatkerettár építése szempontjából releváns információkat hordoznak.

Azonban azokat a particípium alakokat, amelyek előmódosító funkciójú melléknév szerepét töltötték be (*a protecting group, protected amino group*), nem vettük fel az igei vonzatkerettárunkba. Ugyanis – bár ezek is öröklik az ige eredeti vonzatait – ezen szószerkezetek esetében a szintaktikai viszonyt kifejező prepozíció a felszínen nem jelenik meg (pl. a *treat with heat* szerkezet *heat treated*-ként jelenik meg), és az elmaradó prepozíció kezelése problémákat vethet fel az elemző számára. Másrészt pedig a melléknév és az azt megelőző tárgy gyakran kötőjellel van egymáshoz kapcsolva (*electron-withdrawing groups*), vagyis ezekben az esetekben már összetett szónak, vagyis egyetlen lexikai elemnek is lehet tekinteni őket.

A kézi ellenőrzés során egyéb esetek is voltak, melyekben nem volt evidens, hogy egy adott szóalakot igeként célszerű-e kezelni vagy sem. Ilyenek voltak bizonyos utómódosítók igéből képzett elemei (pl. *a method comprising administering a pharmaceutical composition*), többszavas kifejezések igei elemei (pl. *as follows*), az alany és állítmány nélküli mellékmondatok, azaz melléknévi szószerkezetek *-ing*-es alakja (pl. *when treating…*), a szenvedő szerkezet maradványaként álló, s alanykomplementumként funkcionáló igei 3. alakok (past participle) (pl. *when administered to*). Ezekben az esetekben egyedi elbírálást alkalmaztunk. Vagyis ha úgy ítéltük meg, hogy ezen kifejezések szignifikánsan magas számban fordulnak elő a szabadalmi szövegekben, akkor felvettük őket a vonzatkerettárunkba. Így jártunk el például a külön szótári tételt is alkotó, lexikalizálódott elemekkel kapcsolatban, (pl. *as follows*, *provided that*, *according to*), amelyek leggyakrabban kötőszóként vagy elöljárószóként funkcionálnak.

2.2 Vonzatkeretek kigyűjtése

Az igék kézi ellenőrzése után a vonzatkeretek kigyűjtése következett – szintén kézileg, (ezt gépileg – a szabadalmi szövegekre testreszabott nagy pontosságú szintaktikai elemző híján – nem lehetett megoldani). A vonzatkeret fogalmát – praktikussági okokból – tágan értelmeztük: az ige *kötelező* vonzatainak összességén kívül az egyéb, szorosan összetartozó elemekből álló kifejezéseket is idevettünk (amelyeket alább részletesebben tárgyalunk), és felvettük a kerettárunkba, hiszen célunk volt egy, a szintaktikai és szemantikai elemzéshez gyakorlatban jól használható eszköz kialakítása.

A vonzatkerettárunk összeállításakor elsősorban természetesen az ige kötelező bővítményeire fókuszáltunk. Az igéknek a tranzitív és nem tranzitív alakjait egyetlen igének és egy elemnek tekintettük a vonzatkerettárunkban, annak ellenére, hogy különböző a vonzatkeretük, pl. a *substitute* ige lehet tárgyas és tárgyatlan is. Tárgyas formájában a vonzatkerete: *V N,* vagy *V N for N;* tárgyatlan formájában: *V*, vagy *V for N.* Ezeket a vonzatkereteket tehát mind felvettük a *substitute* igéhez.

Mivel minden angol igének, így a szabadalmakban szereplő összes igének is van (nyelvtani) alanya, ezt a vonzatot default elemnek tekintettük, s nem vettük fel egyetlen ige vonzatkeretéhez sem.

Kérdést vetett fel, hogy a (közel) azonos jelentéssel bíró és formailag is csak minimálisan eltérő alakú prepozíciókat (pl. *combine together/together with, depend on/upon*) különálló vonzatkeretként célszerű-e kezelni. Mivel az automatikus szintaktikai elemzés nem szemantikai jellemzőkből indul ki, úgy döntöttünk, hogy különálló vonzatkeretekként kezeljük őket.

Hasonló kérdéskörbe tartozó problémát vetett fel a *from* prepozíció esetenkénti megjelenési formája: a *remove*, ill. a *vaporize* vonzataként néhány esetben *therefrom*-ként jelent meg (*drying said plasticized granules to remove substantially all the solvent* ***therefrom***), ami a *from there* szerkezet módosult formája. A *therefrom* megjelenési alakot nem vettük fel külön vonzatkeretként, mivel a *from that* szinonimájaként kezelendő. Az elemző algoritmus implementálásakor emiatt a *therefrom* és *thereof* szóalakokra fokozott figyelmet kell fordítani, mert a tapasztalatok alapján a Stanford parser tévesen főnévnek tekinti e szóalakokat, valójában pedig határozószavak, és nyelvtani szerepüket tekintve PP-k, a *there* alkotóelem pedig anaforikusan utal vissza egy korábbi összetevőre.

A kötelező bővítményeken kívül olyan szókapcsolatokat is felvettünk a vonzatkerettárunkba, amelyek ugyan nem kötelező vonzatai az igének, azonban megítélésünk szerint kiemelkedően jellemzőek a szabadalmakra. Ilyenek voltak bizonyos szabad határozók (pl. célhatározói *to infinitivus* alakok).

A vonzatkeretek kigyűjtése után megszámoltuk, hogy az adott ige az egyes vonzatkereteivel hányszor fordul elő és meghatároztuk, hogy ez az előfordulási szám gyakorinak számít-e a többi vonzatkeret előfordulásához képest viszonyítva (pl. megnéztük, hogy a *consist* ige hányszor fordul elő összesen, és ebből hányszor fordul elő *in* + főnév vonzattal). Erre azért volt szükség, mert a különböző igék összességében nem azonos gyakorisággal (és nem azonos számú vonzatkerettel) fordultak elő a korpuszban, így nem tudtunk meghatározni egy általános érvényű küszöbértéket, amely felett gyakorinak minősítünk egy adott vonzatkeretet.

2.3 Igék és vonzatkereteik kigyűjtése a Google Dictionaryből

Következő lépésként a szabadalmakból kigyűjtött igéket a Google Dictionaryből is kigyűjtöttük, vonzatkereteikkel együtt. (A vonzatkerettárunkban is a szótár jelöléseit követtük, mely szerint – az általános használattól eltérően – a V-ed az ige 3. alakját (past participle) jelöli). Az internetes szótár nem volt egészen következetes a vonzatkereteket illetően, hiszen olyan szerkezeteket is különálló vonzatnak vett, amelyek valójában ugyanannak a vonzatnak különböző (szabályszerűen képezhető) alakjai. Pl. az ige + főnév (V N) vonzatkeret és az '*-ing*'-es alak + főnév (V-ing N) különböző vonzatkeretként fordul elő, holott ez a kettő valójában ugyanaz a vonzat (hiszen az *-ing*-es alak automatikusan képezhető az elsőből). Így a második képletet (V-ing N) redundánsnak tekintettük, s ezért nem vettük fel külön vonzatkeretként a kerettárba.

A szenvedő szerkezetet jelölő vonzatkeret ('be' V-ed) szintén redundáns elemként jelent meg a Google Dictionary vonzatkerettárában (hiszen ez is automatikusan előállítható az alapértelmezettnek tekintett aktív igei szerkezetekből), azonban – az *-ing*-es alakokkal ellentétben – ezeket különálló vonzatkeretnek tekintettük, mert a passzív igei alak eléggé szabadalomspecifikus; ezenkívül bizonyos esetekben a 'be' V-ed 'by' alakot is felvettük jelentéstani okokból, pl. *characterized by, substituted by.*

3 Az igei vonzatok két halmazának összevetése, orvostudományi szakszövegekre alkalmazható vonzatkerettár összeállítása

A következő fázisban összevetettük a szabadalmak igei vonzatkereteit a Google Dictionaryből nyert vonzatkeretekkel, és megvizsgáltuk, hogy mennyire feleltethetők meg egymásnak. Mint az várható volt, a kettő nem volt tökéletes fedésben. Háromféle eset fordult elő: a szabadalmakban szereplő igék

a) a korpuszban szereplő vonzatukkal együtt megtalálhatók voltak a Google Dictionaryben is (pl. *adhere*, *impregnate*, *regard*). (Némely esetben ugyan a szabadalmakban szereplő amerikai angol helyesírású szó helyett a brit angol helyesírású verziót találtuk meg (pl. *analyse* vs. *analyze*), de ezeket természetesen találatnak tekintettük.)

b) szerepeltek ugyan a Google Dictionaryben, azonban a korpuszban előforduló vonzatkeretük(kel) nem. Ezekben az esetekben be kellett illesztenünk a kerettárba egy-egy új vonzatkeretet (pl. a *bind* ige 'to' + főnév vonzatkeretét); s voltak esetek, amikor több új vonzatkeretet is fel kellett vennünk a listára (pl. a *combine* ige esetében ötöt).

c) egyáltalán nem szerepeltek a Google Dictionaryben. Ezek túlnyomó többsége orvosi/kémiai terminus technicus volt, pl. *acidify*, *benzofuse*, *coprecipitate*. (Azonban olyan általánosabb jelentésű igékkel is találkoztunk a korpuszban, melyeknek a szótárból (igeként) történő hiányzása némileg meglepő volt: pl. a *potentiate* ige hiánya, ill. a *passage* szóalak kizárólag főnévként történő szereplé-

se). Ezeket az igéket természetesen egy az egyben felvettük az igei listára a korpuszban szereplő vonzatukkal.

Mivel a b) és c) pontban leírt esetekre számos példa előfordult, evidenssé vált az – amit sejteni lehetett előre is –, hogy az orvostudományi szabadalmi szövegeknek megvan a saját szakszókincsük, és bizonyos nyelvtani fordulatok is elsődlegesen rájuk jellemzők és nem a köznyelvre, vagyis általános célú szótárt nem lehet megfelelően alkalmazni orvostudományi szabadalmi szövegekre. (Ez nyilván jelentős információ a szintaktikai (és szemantikai) elemző kialakításához).

Ennek fényében tehát a Google Dictionaryből nyert vonzatkerettárat jelentős mértékben ki kellett egészítenünk a szabadalmi szövegekből kigyűjtött vonzatokkal, s ezáltal kialakítottunk egy, specifikusan az orvostudományi szabadalmakra alkalmazható vonzatkerettárat.

4 Eredmények

Az elkészült vonzatkerettár 220 igét tartalmaz, melyeknek összesen 1498 vonzatkeretük lett felvéve (ebből 93 nem szerepelt a Google Dictionaryben, ezeket a szabadalmak szövege alapján illesztettük be).

A köznyelvi szóhasználathoz hasonlóan a szabadalmi szövegekben is a legtöbb ige egy vonzatkereten belül egy vagy két vonzattal rendelkezik, s a háromvonzatos ige (pl. *inject N through N to N*) ritka.

Ha viszont a vonzat*keretek* számát vizsgáljuk, a következőket találjuk. A szabadalmakban előforduló igék Google Dictionaryben szereplő megfelelőit tekintve a legtöbb vonzatkerettel rendelkezők a következők: *come* (24), *make* (24), *take* (22), *stand* (21), *leave* (20). Viszont a nagyszámú vonzatkeretek ellenére újabbakkal kellett kiegészítenünk ezen igéknek a szabadalmakra testre szabott vonzatkeretlistáját, hiszen a referenciaszótárunk vonzatkeretei csak elenyésző mértékben fedték le a szabadalmakban szereplő vonzatkereteket: a legtöbbjük esetében csak egy vagy két olyan vonzatkeretet találtunk a Google Dictionaryben, amely a szabadalmakban találhatóval megegyezett. Azonban arra is volt példa, hogy a referenciaszótárunkban található nagyszámú vonzatkeretből egyik sem egyezett meg a szabadalmakban találhatókkal. Például a *take* a Google Dictionaryben huszonkétféle vonzatkerettel szerepel, de a szabadalmakban csak egy 23. vonzatkerettel fordul elő: (*be taken together (with) N*). Tehát – többek között – a fentebbi igék esetében egy vagy két vonzatkerettel ki kellett egészítenünk a Google Dictionaryből nyert – és egyébként gazdag – vonzatkeretlistát.

Ezzel szemben a szabadalmi szövegekben az igék jóval kevesebb ténylegesen előforduló vonzatkeretét figyelhettük meg. Hat vonzatkerettel két ige rendelkezik: az *add* és a *combine*. Öt vonzatkerettel a *comprise* és a *form* igék, néggyel a *define*, *select* és *determine*, három vonzatkerettel 11 ige, két vonzatkerettel 46 ige, 1 vonzatkerettel (amely általában egyetlen tárgyi vonzatot tartalmaz és így az ige tranzitív voltára utal) 172 ige rendelkezik.

1. táblázat: A legtöbb vonzatkerettel rendelkező igék vonzatai.

add:	*be V-ed to N*	**combine**:	*V N with N*
	V N		*V with N*
	V to N		*V together*
	V N to N		*be V-ed together with N*
	be V-ed in N		*be V-ed with N*
	V to N N		*be V-ed*

A (szabadalmi szövegekben) a legtöbb vonzatkerettel rendelkező, fentebb említett *add* és *combine* ige a Google Dictionaryben is viszonylag nagy számú vonzatkerettel rendelkezett (9, illetve 6), azonban mivel ezek nem vágtak egybe a szabadalmakban előforduló vonzatkeretekkel, a vonzatkeretlistánkat ki kellett egészíteni (az *add* ige vonzatkereteit kettővel, a *combine* igéét pedig öttel).

A szótárba összesen 16 darab új, vonzatkeretes igét kellett felvenni: ezek olyan szavak voltak, amelyek – többségükben kémiai, illetve orvosi szakszavak lévén – nem voltak megtalálhatók a Google Dictionaryben. Ilyen volt például az *admix* (*admix N with N*), *solubilize* (*solubilize N*) vagy *anellate* (*be anellated with N*).

39 ige esetében fordult elő, hogy a szabadalmakban a Google által hozzájuk rendelt vonzatkereteik nem szerepeltek, de valamilyen más, azaz új vonzatkerettel viszont igen. Ilyen igékre példa a *prescribe*, amely a szabadalmakban *prescribe to N N*, vagy az *engineer*, amely *be engineered to N* alakban fordult csak elő. A köznyelvben leggyakoribbnak tekinthető igék, például a *take* esetében is ez volt a helyzet, amint már fentebb utaltunk erre.

A legtöbb új vonzatkeretet a *combine* kapta, egészen pontosan ötöt, pl. a *be combined together with N* alakot. Ezen kívül három új vonzatkeretet kellett felvenni a *define* (pl. *be defined as*), *determine*, *rack* és *select* igékhez. A többi igét legfeljebb kettő új vonzatkerettel kellett kibővíteni.

5 Megfigyelések a vonzatkerettáron

5.1 Kompozicionalitás

A korpuszban előforduló igéket és vonzatkereteiket érdemes például a kompozicionalitás szempontjából megvizsgálni. (Minden igei vonzatkeretet kigyűjtöttünk függetlenül attól, hogy azok az igével kompozicionális szerkezetet alkotnak-e.) Az itt előforduló vonzatkeretek legtöbbször kompozicionális szerkezetet alkotnak az igével (vagyis az összetétel jelentését egyértelműen meghatározza az összetevőinek (az igének és vonzatának) jelentése és az összetétel módja), pl. *dilute with N, be added to N, impart from N to N.*

Azonban előfordultak nem kompozicionális szerkezetek is: például a *stand for* elöljárós ige ‘jelent’, ‘helyettesít’ értelemben nem kompozicionális: *R2 and R3 independently* ***stand for*** *H, C1-6 alkyl, C2-6 alkenyl.*. Ezt az elöljárós igei alakot a *stand* ige vonzatkerettárába vettük fel (*V for N*). A kevés ilyen jellegű példa arra utal, hogy az (orvostudományi) szabadalmi szövegekre valószínűleg nem jellemzőek a nem kompozicionális igei szerkezetek (melyek – az angol nyelvben – lehetnek idiómák, illetve a elöljárós igék (‘phrasal verbs’)).

5.2 Módbeli segédigék

Módbeli segédigékkel kapcsolatosan azt figyeltük meg, hogy pl. a segédigeként és főigeként egyaránt funkcionálni képes *do* és *have* igéket tekintve eltérőek a tapasztalatok: a *do* kizárólag segédigeként szerepelt, míg a *have* kizárólag főigeként fordult elő a szabadalmi szövegekben. A *do* mint főige előfordulási hiánya – legalábbis részben – szintén a kompozicionalitás kérdésével lehet összefüggésben. Ugyanis főigeként általános szövegekben igen gyakran nem kompozicionális szerkezetekben (pl. *do away with*), vagy félig kompozicionális szerkezetekben fordul elő (pl. *do a favour*), amely szerkezetek viszont – mint fentebb említettük – határozottan nem jellemzőek e szabadalmi szövegekre. A *have* segédigeként történő előfordulásának hiányát pedig az magyarázhatja, hogy ilyen funkciójában olyan igeidőket, illetve -módokat (pl. a különféle befejezett igeidők, műveltető) fejez ki, melyek szintén nem jellemzik a szabadalmi szövegeket.

5.3 Egyéb jellemzők

Az általános nyelvvel szemben a tudományos szövegekre erőteljesebben jellemző további jelenség a vonzatok sorrendiségével kapcsolatos. Például a *prescribe* ige két vonzata általában a következő sorrendben szokott az ige után állni: *V N to N* (felír vmit vkinek), vagy a *V N N* (felír vkinek vmit). Azonban a szabadalmi szövegekben megfigyelhető, hogy a hosszabb és komplikáltabb tárgy a könnyebb érthetőség kedvéért a (*to* prepozícióval kifejezett) részeshatározó mögé kerül: *V to N N* (pl. *prescribing to the patient a therapeutically effective amount of quazepam*). (Az angol nyelvészeti terminológiában *heavy NP shift*-nek nevezik ezt a jelenséget.)

A fentebbieken kívül a jövőben még érdemes lenne megvizsgálni például azt, hogy a vonzatkerettárba jelenleg fel nem vett, előmódosítói szerepű, particípiumos szerkezetek és vonzataik hogyan építhetők be a vonzatkerettárba – a kötőjelezeléssel összefüggésben (pl. *diabetes-associated disorders*); illetve a többszavas igei kifejezések (vagy félig kompozicionális szerkezetek, l. [10], pl. *come into contact with N*) kezelési módját is érdemes tovább fejleszteni.

6 Összevetés más igei vonzatkerettárakkal

Az angol nyelvre már készültek korábban is igei vonzatkerettárak, illetve olyan korpuszok, amelyek tartalmaznak a vonzatkeretre vonatkozó információt. Ilyen például a VerbNet [3, 4, 5], a Proposition Bank [8] és a FrameNet [1]. A Proposition Bank a Penn Treebank szintaktikai szerkezeteihez rendel szemantikai szerepeket, a VerbNet a kibővített Levin-féle [7] igeosztályok szintaktikai kereteit, az argumentumok szemantikai szerepeit és a rájuk vonatkozó szelekciós megkötéseket tartalmazza, a FrameNet pedig a szemantikai keretek felől közelítve adja meg az adott keretbe illeszkedő igéket és azok argumentumainak szintaktikai és szemantikai tulajdonságait.

Noha a fenti adatbázisok is részletes információkat tartalmaznak az igei vonzatkeretekre nézve, mégsem ezeket választottuk vizsgálatunk alapjául, mivel ezek elsődle-

gesen a szemantikai szerepekre koncentrálnak, minket pedig elsősorban a szintaxis érdekelt. Azonban az egyes igékhez tartozó bejegyzések összevetése mindenképpen hasznos tanulságokkal szolgálhat. Példaként tekintsük a *substitute* igét!

Az általunk kialakított vonzatkerettárban a *substitute (helyettesít)* igének két vonzata szerepel: a) *valamit*: a régi entitás, melyet lecserélünk, és b) *valamivel*: az új entitás, amellyel helyettesítjük a régit (*V N for N*).

Nézzük meg, hogy a tematikus szerepekre koncentráló adatbázisok milyen kategóriákkal dolgoznak, és ott milyen jellemzőkkel jelenik meg a *substitute* ige.

A FrameNet a tematikus szerepeket alapvető és opcionális alcsoportokra osztja. A *substitute* igével kifejezett esemény jellemzésére a következő alapvető szerepeket határozza meg: ágens (aki a cselekvést végrehajtja), új entitás (amellyel az ágens betöltet egy szerepet), régi entitás (amely korábban betöltötte az adott szerepet). Az esemény opcionális szereplőiként pedig olyan szereplőket, illetve szerepeket nevez meg, amelyek szabad határozóként funkcionálnak (vagyis nem kötelező vonzatai az igének), pl: eszköz, mód, szerep, hely, cél, ok, idő stb. A Proposition Bank négy szerepet (argumentumot) határoz meg: ágens, egyes számú téma (Theme1), kettes számú téma (Theme2), és kedvezményezett / beneficiens; s nem jelöli meg ezek közül a kötelezőeket. A VerbNet a *substitute* igének szintén két kötelező vonzatát jelöli meg: téma 1 (THEME 1) és téma 2 (THEME 2).

A fentebbi vonzatkerettárak két fontosabb szempontból térnek el a szabadalmakra készített vonzatkerettárunktól. Egyrészt tárgykörükben térnek el egymástól: a fentebbi adatbázisok általános doménben alkalmazhatók, míg az általunk készített kerettár specifikus doménre készült. Másrészt pedig míg ez utóbbi a szintaxisra helyezi a hangsúlyt, a fentebbi vonzatkerettárak a szemantikai információkra fókuszálnak. Ez utóbbiakat a későbbiekben érdemes lehet beépíteni a szabadalmakra készített vonzatkerettárban szereplő argumentumok reprezentációjába. Mivel a kötelező bővítmények és a tematikus szerepek között egy az egyhez megfeleltetés figyelhető meg, vagyis minden kötelező vonzatnak egy és csakis egy tematikus szerepe lehet, viszonylag gyors és egyszerű a kettő közötti megfeleltetés.

Amennyiben csak a főbb szemantikai szerepekre szeretnénk koncentrálni, célszerű a Proposition Banket, illetve a VerbNetet használni, amelyeknek az az előnye is megvan, hogy – mivel kevesebb adattal operálnak e rendszerek – gyorsabb megoldásokat kaphatunk. Amennyiben azonban részletesebb szemantikai reprezentációra törekszünk, az összetettebb rendszerű FrameNetet érdemes használnunk. Ez azért is lenne előnyösebb a számunkra, mert olyan elemekhez is szeretnénk tematikus szerepet hozzárendelni, amelyeket a fentebb említett két másik rendszer nem tartalmaz. Ezek az elemek a szabad határozók, melyek tematikus szerepének leelemzése hosszadalmasabb folyamat, hiszen – a kötelező bővítményekkel ellentétben – egy-egy szabad határozónak többféle tematikus szerepe is lehet.

7 Főnévi vonzatkerettár

Az igei vonzatkerettáron kívül a főnévi vonzatkerettár is elkészült a szóban forgó szabadalmi korpusz alapján, a fenti elveket alkalmazva. A könnyebb kezelhetőség végett a főneveken belül elkülönítettük a perdurantokat (időbeli történést, esemény-

szerűséget jelölő főnevek, l. [9], amelyek több szempontból hasonlítanak az igékhez. Egyrészt hasonló a jelentésük, mivel eseményt fejeznek ki. Másrészt fontos azon jellemzőjük is, hogy szinte bármennyi és bármilyen szabad határozóval rendelkezhetnek. A perdurant jelentésű főneveket szemantikailag is egy kategóriába soroltuk az igékkel a reprezentáció során, hiszen a *method for treating Alzheimer's disease* és a *method for the treatment of Alzheimer's disease* jelentésében nincs különbség.

A vonzatkerettár szempontjából azért volt fontos megkülönböztetni a perdurant főneveket a nem perdurant főnevektől, mert az utóbbiaknál csak a Google Dictionaryben szereplő vonzatokat illesztettük, míg az előbbieknél szabad prepozíciós szerkezeteket is megengedtünk. Ez sokat javított a program hatékonyságán, mert volt olyan főnév is, amelynek 4 bővítménye is volt, ez pedig a *storage*:

storage (1) of the composition (2) for ten days (3) in an open Petri dish (4) at 40°C.±2°C.

Ezen esetekben, ha csak a vonzatkerettárat vennék alapul, akkor a (2-4) bővítményeket az előtte álló igéhez tettük volna. Általában véve igaz, hogy bármilyen főnévnek lehet *of* prepozícióval kezdődő vonzata, ezért azokat alapértelmezés szerint kivettük a vonzatkerettárból. Kevés olyan nem perdurant jelentésű főnévvel találkoztunk, ami szabadalomspecifikus lett volna. Ezek egyike volt a nagyon gyakran előforduló *means*, amelynek a Google Dictionary szerint csak *to*+inf vonzata lehet, de a szabadalmakban gyakran előfordult a *for* is.

A főnévi vonzatkerettárban 117 db főnév található összesen 162 vonzatkerettel.

8 Összegzés

Ebben a munkában beszámoltunk egy orvostudományi szabadalmak szövegein alapuló igei és főnévi vonzatkerettár létrehozásáról. Kiindulási alapnak egy általános célú szótárt, a Google Dictionary vonzatkereteit tekintettük. A vonzatkerettár létrehozása során kiderült, hogy léteznek szabadalomspecifikus igék, illetve szabadalomspecifikus vonzatkeretek, melyeket az általános célú szótár nem tartalmazott, így ezeket külön fel kellett vennünk, azaz az általános célú szótár csak korlátozottan használható a szabadalmak elemzésére. A vonzatkerettárakat a későbbiekben szeretnénk szemantikai jellegű információval is bővíteni, és ezáltal a vonzatokhoz tematikus szerepeket társítani. Az elkészült adatbázis eredményesen hasznosítható a szabadalmi szövegekre fejlesztett szintaktikai és szemantikai elemző fejlesztésében.

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Baker, C. F., Fillmore, C. J., Lowe, J. B.: The Berkeley FrameNet project. In: Proceedings of the COLING-ACL. Montreal, Canada (1998)
2. http://www.google.com/dictionary
3. Kipper, K., Dang, H.T., Palmer, M.: Class-Based Construction of a Verb Lexicon. In: AAAI-2000 Seventeenth National Conference on Artificial Intelligence (2000)
4. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending VerbNet with Novel Verb Classes. In: Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy (2006)
5. Kipper, K., Palmer, M., Rambow, O.: Extending PropBank with VerbNet Semantic Predicates. In: Workshop on Applied Interlinguas, held in conjunction with AMTA-2002 (2002)
6. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
7. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago, IL (1993)
8. Palmer Martha, M., Gildea, D., Daniel, Kingsbury Paul, P.: The Proposition Bank: an annotated corpus of semantic roles. Computational Linguistics Vol. (2005) 31 No. 1(1): (2005) 71–105
9. Ungváry R.: Az ontológiák legfelső generikus szintje, a csúcsfogalmak természetes rendszere és a DOLCE kritikája. In: Alexin Z., Csendes D. (szerk.): MSzNy 2006 – IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 85–96
10. Vincze, V., Csirik, J.: Hungarian Corpus of Light Verb Constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, China (2010) 1110–1118

Egy vertikális keresőrendszer készítése

Orosz György

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
e-mail: oroszgy@itk.ppke.hu

Kivonat A tanulmányban egy nyílt forrású alapokra épülő, nyelvi eszközökkel felvértezett vertikális keresőrendszer építésének lépéseit mutatom be. Az eljárás során sor kerül létező rendszerek összehasonlítására, egy kiválasztott kereső vizsgálatára, illetve a nyelvi modulok építésének bemutatására. A készített rendszert összevetem az eredetivel teljesítmény és relevancia szempontjából, majd megmutatom, hogy milyen területeken sikerült érdemben javítanom a működésen.

Kulcsszavak: információ-visszakeresés, keresőrendszerek, nyelvi kereső

1. Bevezetés

Keresőrendszerek használata mára már mindennapi rutinunk részét képezi. Internetes keresőkben egyre-másra tűnnek fel nyelvi eszközök, melyek célja, hogy minél pontosabb találatokkal szolgáljanak a felhasználóknak. Ezek az eszközök – nagy mennyiségű adatról lévén szó – többségében statisztikai alapúak.

Vertikális keresésről beszélünk akkor, ha a keresés az internet egy szeletén vagy valamilyen kritérium mentén szűkített korpuszon történik. Ez a kritérium leggyakrabban egy témakör, de lehet fájlformátum, dokumentumtípus stb. is. A továbbiakban bemutatom, hogy vizsgálataim alapján milyen lehetőségek kínálkoznak egy nyelvi vertikális kereső építésére.

Először áttekintem a nyílt forráskódú keresőrendszereket, megvizsgálom, milyen feltételek szükségesek egyáltalán egy ilyen projekt sikeréhez. Ezen elvárások mentén összehasonlítom a keresőket, majd egy választott képességeit megismerve a Magyar Országgyűlés plenáris üléseinek annotált korpuszához [5] készítek vertikális keresőt. Bemutatom a fejlesztés lépéseit, az integrált, illetve létrehozott új modulok képességeit, továbbá a készítés során megfogalmazott tapasztalatokat. Végül megmutatom, hogy az új rendszer hogyan teljesít: az egyes eszközökkel hol és milyen mértékben sikerült javítani az eredeti kereső működésén, különös tekintettel a relevancia szerinti eredményességen.

2. Keresőrendszerek összevetése

Munkám célja, hogy egy nyelvi eszközöket is használó általános célú vertikális keresőrendszert hozzak létre, majd vizsgáljam ennek működését. Ehhez alapul

egy létező nyílt forráskódú keresőt kerestem. A következőkben áttekintem, milyen alternatívák kínálkoztak.

2.1. Szempontok

Mielőtt elkezdtem volna a keresők közötti böngészést és azok tanulmányozását, megfogalmaztam azon szempontokat, melyek mentén a létező rendszereket vizsgálni kívánom:

- továbbfejlesztéshez használható API létezése, minősége
- projekt állapota: stabil verzió, aktív fejlesztők, dokumentáltság,
- integrált NLP eszözök száma, minősége,
- támogatott karakterkódolások (magyar nyelvű dokumentumok feldolgozásához szükséges legalább egy az alábbiakból: UTF-8, ISO 8859-2, WINDOWS CP-1250) volta.

2.2. Vizsgált keresőrendszerek

Munkám során igyekeztem a lehető legtöbb rendszert górcső alá venni. Ezek felkutatásához elsősorban az alábbi dokumentumokat vettem alapul:

- Emmanuel Eckard és Jean-Cédric Chappelierés - Free Software for research in Information Retrieval and Textual Clustering [4]
- Christian Middleton és Ricardo Baeza-Yates - A Comparison of Open Source Search Engines [2]

Így a következő rendszereket vizsgáltam: OpenFTS[1], Terrier[2], Lucene[3], Datapark Search Engine[4], Egothor[5], Xapian[6], ht://Dig[7] és Lemur/Indri[8]. Az Egothor, OpenFTS és a ht://Dig alkalmazásokat már munkám elején elvetettem, hiszen ezek fejlesztése évekkel ezelőtt abbamaradt, vagy fejlesztési dokumentáltságuk egyáltalán nem kielégítő.

A **Terrier** JAVA nyelven íródott, így UTF-8 támogatása biztosított, továbbá jól dokumentált API-val rendelkezik. Indexelés folyamán a rendszer egy pipelineon keresztül dolgozza fel az indexelendő kifejezéseket, mely jól használható (nyelvtechnológiai) modulok sorba kötéséhez. Fejlesztése folyamatos, a támogatás biztosított. Szótövező használatára – az említett csővezeték-kialakítás miatt – ad lehetőséget, de más nyelvi modulokat nem használ.

A **Lucene** egy teljes értékű keresőrendszer, mely szintén JAVA nyelven íródott, így a Terrierhez hasonlóan a magyar nyelv írásjeleit is tökéletesen kezeli. Jól

[1] http://openfts.sourceforge.net/
[2] http://terrier.org/
[3] http://lucene.apache.org/
[4] http://www.dataparksearch.org/
[5] http://www.egothor.org/
[6] http://xapian.org/
[7] http://www.htdig.org/
[8] http://www.lemurproject.org/

dokumentált alkalmazásprogramozási felülettel rendelkezik. Legfőbb erőssége, hogy jól skálázható és nagy teljesítményű indexelő program biztosítja az adatok gyors feldolgozását. Képes dátumok szerinti keresésre, mezők kezelésére, egyszerre több indexben való keresésre, továbbá többféle lekérdezéstípust is támogat. A projekt rendkívül jól támogatott, dokumentált. Nyelvi eszközök kapcsán az mondható el róla, hogy többféle szótövezőt és azok integrálását is támogatja.

A **Datapark Search Engine** egy C nyelven íródott kereső, melynek API-ja csak kis mértékben enged beavatkozni a rendszer működésébe. Ezt is elsősorban olyan esetekben használják, amikor a szoftvert egy komplex programba vagy weboldalba építik be. Némi dokumentáció is fellelhető a világhálón, de ez korántsem teljes. Több karakterkészletet is támogat, viszont nyelvi eszközökkel csak nehézkesen gyarapítható. Szótövezés csak Aspell, illetve Ispell alkalmazásokon keresztül valósítható meg, illetve szinonimák és mozaikszavak felismeréséhez szótárfájlok használatával biztosít lehetőséget.

A **Xapian** egy olyan eszközkészlet, melyet C++ nyelven írtak és támogatja Unicode karakterek használatát. Egy ráépülő népszerű keresőrendszer az **Omega**, mely integrált nyelvi eszközzel rendelkezik, úgymint szótövező (magyar nyelvű is!) és keresés közbeni szinonimahasználat. Képes még több adatbázis egyidejű indexként való használatára. A projekt jól dokumentált és folyamatosan karbantartott. A kereső alkalmazásprogramozási felületét is úgy tervezték, készítették, hogy elsősorban a program beágyazását tegye lehetővé, nem pedig a bővítését.

A **Lemur** eszközkészlet a Carnegie Mellon University és a University of Massachusetts támogatásával készült. Céljuk egy olyan keretrendszer létrehozása volt, mely nyelvmodellezési és adatbányászati kutatásokhoz jól használható. A program C++ nyelven íródott, de rendelkezik API-val más nyelvekhez is. Az angol nyelvi szövegeken kívül kínait és arabot is támogat (nyelvi eszközökkel is), továbbá beépített angol szótövezőkkel és mozaikszó-felismerővel is rendelkezik. A rendszer a dokumentumok feldolgozását pipeline-szerűen végzi, melyhez jól használható API-t biztosít, továbbá képes több indexfájl egyidejű használatára is. Beépített nyelvi eszközei: angol nyelvű szótövező (2 db), arab nyelvű tövező, betűszó-felismerés és stopwordtámogatás. Az **Indri** egy olyan keresőrendszer, mely a Lemur eszközkészletére épül. A fentieken kívül fontos tulajdonsága, hogy támogatja az UTF-8 kódolású dokumentumokat is.

A fentebb részletezett szempontok alapján az Indri keresőrendszer tűnt a legjobb választásnak, így a továbbiakban az ezzel való munkámat és eredményeimet mutatom be.

3. Integráció

A kereső készítése során a célom az volt, hogy megvizsgáljam, hogy egy létező rendszer nyelvi eszközökkel való kiegészítésére milyen lehetőségek mutatkoznak, illetve milyen eredményességgel lehetséges ez a munka. A fejlesztéshez a Magyar Országgyűlés plenáris üléseinek annotált korpuszát vettem alapul. Ezt tanulmányozva készítettem el a meglévő nyelvi eszközök integrációját, illetve hoztam létre újakat. Abból a hipotézisből indultam ki, hogy egy jól illeszkedő

lemmatizáló[9] modul és egy kontrollált szinonimaszótár[1] csak javíthat a kereső eredményességén. Ezeken kívül a következő eszközöket készíttetem el és integráltam az Indribe dátum szerinti, mértékegységek szerinti kereshetőség és személyek/felszólalók keresése. Az alábbiakban áttekintem a fejlesztés során tapasztalt nehézségeket, az ezekre adott megoldásokat illetve a használt módszereket.

3.1. Létező modulok integrációja

Az alábbiakban bemutatom, milyen létező eszközöket és hogyan használtam a keresőrendszer fejlesztéséhez.

A kapott **lemmatizáló** modul beépítése, a korábban említett pipeline-szerű feldolgozásnak köszönhetően, különösebb nehézségek nélkül működött. Az egyetlen probléma a modul és a rendszer különböző karakterkódolásai közti konverzió volt, amit az `iconv` alkalmazás segítségével oldottam meg. A kereső az indexelés folyamatában a dokumentumot mint szavak halmazát kezeli, és egy szóhoz egyetlen tövet enged tárolni. Így – mivel lehetőségem sem volt egy-egy szó kontextusának a vizsgálatára – minden szóhoz annak leggyakoribb és egyben legvalószínűbb szótövét tároltam. (Ezzel nyilván rontva a rendszer teljesítményét a fedést illetően.) Az Indri képes arra, hogy az index létrehozásakor megjegyezze a készítéskor használt lemmatizálót, így lekérdezéskor a keresőkifejezésen képes ugyanazt futtatni. Ebből kifolyólag az indexelés/lekérdezés folyamataiba, a tövező integrálása céljából összesen egy helyen volt szükséges beavatkozni.

A **szinonimaszótár** integrálása során azt találtam, hogy ennek a funkciónak a rendszer általi automatikus használata – a generált nagy mennyiségű zajos adat miatt – jelentősen ronthat [6] a kereső eredményességén. Így úgy döntöttem, hogy oly módon teszem elérhetővé ezt az eszközt, hogy a felhasználó kontrollálhassa annak kimenetét. Ennek módja a következő volt: az Indri lekérdezőnyelvét bővítettem egy új operátorral (`~`), melynek használatakor az utána álló szó vagy kifejezés szinonimáival kiegészítheti a felhasználó a keresőkifejezést. A felhasználói interfész létrehozásakor a kereső lekérdezőnyelvére tudtam támaszkodni, hiszen az több operátort is biztosít rokon értelmű szavak összevonására a kifejezés kiértékelése során. Hasonlóan a tövezőhöz, a modul integrációja során az egyetlen technikai probléma a karakterkódolások közötti konverzió volt.

A **tiltólistás szavak használata** egy elterjedt módszer keresőrendszerek használata során a nem kívánt – jelentős információtartalommal nem bíró – szavak keresésbőli szűrésére. Az általam használt alkalmazás két úton engedi szűrni az ilyen szavakat: egyszerű felsorolással, illetve egy, a pipeline-ba illeszkedő modul írásával. Az előbbit választva a szűrni kívánt szavak listájának alapjául egy szabadon elérhető kis méretű adatbázist[10] használtam, melyet a korpuszból kinyert szóhalmazzal bővítettem. Itt a kiválasztás emberi erővel történt az előforduló szavak leggyakoribb 1%-án.

[9] A rendszer elkészítéséhez a MorphoLogic Kft. moduljait használtam.
[10] http://snowball.tartarus.org/algorithms/hungarian/stop.txt

3.2. Saját fejlesztésű modulok

A rendelkezésre álló korpusz vizsgálata során azt láttam, hogy a dokumentumokban számos olyan információtartalom áll rendelkezésre, melyek kereshetővé tétele – feltételezésem szerint – javít a keresések pontosságán. Ezen információk egy része az annotált korpuszban már jelölve volt, míg másokat a készített alkalmazás tett jelöltté. Így a keresőben most lehetőség nyílik a dokumentumokban fellelhető dátumok és mértékegységek *egységes* keresésére. Ezen kívül a korpusz struktúrájában rendelkezésre álltak a felszólalások adatai, így az egyes felszólalók neve is. Ezt az információt felhasználva lehetőség nyílik személyekre való keresésre is.

A korpuszban található **dátumok** használata során azt tapasztaltam, hogy számos helyen sokféle típusú és formájú dátumokat használ az író. Így célom az volt, hogy a készítendő modul működése testreszabható legyen, és a lehető legtöbbféle formában megjelenő egységeket legyen képes felismerni, kereshetővé tenni. A korpuszban található dátumokat az alábbi csoportokra osztottam: pontosság szerint létezik évszázad pontos, évtized szerint pontos, évre pontos, hónap vagy évszak pontos, napra pontos, órára, percre, napszakra stb.-re pontos. Környezetfüggőség szerint létezik környezetfüggő[11], illetve környezetfüggetlen[12].

Az Indri rendelkezik dátumok kezelésének képességével. Ez a funkció nem csak az index létrehozásakor elérhető, hanem lekérdezésekkor is kitűnően használható (pl.: időintervallum lekérdezések futtatása). Indexeléskor a rendszer csak olyan dátumokat képes tárolni az adatbázisban, amiket a felhasználó a feldolgozandó dokumentumban dátummezőként előre jelzett. (Ezek közül is csak a környezetfüggetlen, napra pontos, és angolszász helyesírásúakat.) Az általam adott megoldás két részből áll. Az első egy olyan előfeldolgozó alkalmazás, mely az annotált korpuszban újabb annotációkat vezet be, míg a második ezeket a jelöléseket feldolgozva rögzíti azokat az indexbe az eredeti dátumkezelő funkció felületét használva. Ahhoz, hogy minél több dátumformát kezeljen a program, ezek megadásait a felhasználó kezébe adtam. Így egy konfigurációs fájl használatával a rendszer az ott megadott formákból egy-egy felismerő reguláris kifejezést készít, melynek segítségével jelöli és értelmezi a dátumokat. A megoldás sajnos magában hordozza a reguláris kifejezések korlátait, így az alábbi jelenlévő korlátok továbbra is élnek:

- csak környezetfüggetlen és napra pontos dátumokat vagyunk képes feldolgozni,
- gondolnunk kell a dátumok ragozott alakjára, és a konfigurációs fájlban megadni az ezeknek megfelelő dátumformákat,
- nem használhatunk ütköző[13] dátumformákat.

[11] Környezetfüggőnek nevezek olyan dátumokat, melyek csak a szövegkörnyezetükben értelmezhetőek pl.: tegnap, előző héten.

[12] Környezetfüggetlennek nevezek minden olyan dátumot, mely önmagában is értelmezhető pl.: 1986.04.30.

[13] Ütközőnek nevezek két dátumformát, ha létezik olyan dátum, ami mindkét formának megfelel.

Mennyiségek használatának kialakításakor támaszkodhattam az Indri azon funkciójára, hogy rendelkezik ún. numerikus mezők kezelésének lehetőségével. A megvalósított kiegészítő modul a dátumfelismeréshez hasonlóan két részből áll. Az első rész azt hivatott szolgálni, hogy a forrásfájlokban újabb annotációkat hozzon létre, melyek mennyiségeket jelölnek. Az előzőekhez képest itt annyival nehezebb a feladat, hogy tudnunk kell különbséget tennünk különböző mértékek között. Ezt úgy sikerült elérnem, hogy az annotálás folyamán különböző mértékek más-más címkével jelölődnek.

A magyar nyelvben a jelzős szerkezeteknek kötött a formája: a jelző megelőzi a jelzett szót. Mivel így tekinthetünk egy mennyiségi kifejezésre is, ezért feltételeztem, hogy egy mértékegység feldolgozásakor az azt megelőző szóval vagy számmal összetartoznak. A felismerő modul ebből az alapötletből kiindulva működik, azonban akadnak olyan esetek, mikor ez a hipotézis nem alkalmazható. Egy ilyen gyakori eset, amikor a mértékegységről mint fogalomról ír a szerző. Pl.: ” . . . az euró bevezetése. . . ” Egy másik nehézség, amivel a program fejlesztése során találkoztam, az, hogy az országgyűlési naplókban gyakoriak az olyan pénzösszegeket kifejező mennyiségek, melyekben vegyesen használnak betűvel és számmal írt mennyiséget, pl.: „70 milliárd forint”. A rendszer végül úgy lett kialakítva, hogy intelligensen felismerje és kezelje az ilyen formákat is.

A második modul, mely beépül a szövegfeldolgozási láncba, a mezők feldolgozásáért, reprezentálásáért felelős. Ez túllépve az eredeti kereső korlátain képes valós értékeket is tárolni, továbbá fontos tulajdonsága, hogy egységes (SI) formában reprezentálja az azonos típusú mértékeket. A készített modul egy gyakorlatban is sokat alkalmazható funkcionalitása, hogy képes betűvel írt – magyar nyelven megfogalmazott – számok felismerésére, átváltására. A modul nem kezel olyan törtszámokat, melyek hagyományos – tehát nem tizedes – alakban fordulnak elő. Az intervallumként megadott értékeknél a program a várttól eltérően működik, mivel csak az intervallum egy végét teszi kereshetővé.

A felsorolt eszközök elkészítésén kívül a kereső által nyújtott funkciókkal élve lehetővé tettem még a dokumentumokban fellelhető (annotált) **nevek, személyek, felszólalások, felszólalók visszakeresését**. Ennek megvalósításához felhasználtam az Indri indexelőjének azon funkcióját, hogy képes mezőket tárolni az indexben. Rendelkezésemre állt a korpuszban nagy mennyiségű annotált névelem[14] (főleg személynevek), így ezek indexelése kézenfekvő volt. Ezek az elemek jellemző módon a felszólalásoknál vannak jelölve (a felszólaló személye), de ezeken kívül az egyes beszédekben hivatkozott személynevek is sok esetben elérhetőek.

4. Eredmények

Ebben a fejezetben célom, hogy megvizsgáljam a kiegészített keresőrendszert, és több szempont mentén összehasonlítsam az Indrivel. Nyilvánvalóan nem tudom az összes új modul hatását teljes körűen mérni, mert néhányuk olyan új funkci-

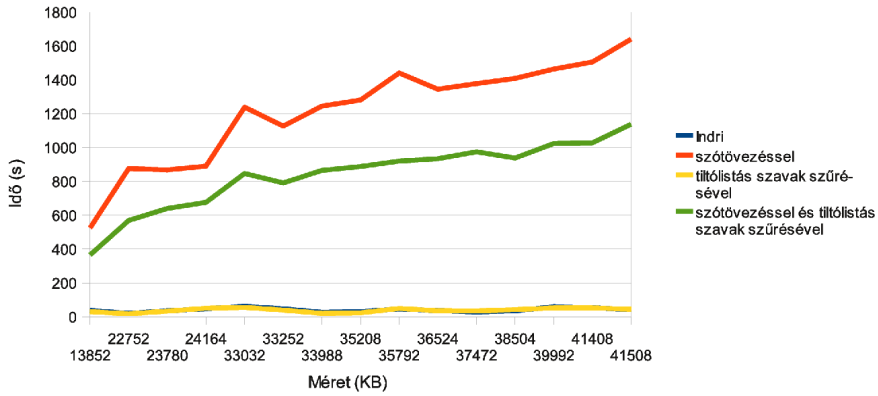
[14] Named Entity

onalitást ad az eszköznek, melyek nem vagy alig összehasonlíthatóak az eredeti program működésével.

4.1. Teljesítmény

Első lépésben azt vizsgáltam meg, hogy a keresőrendszer teljesítménye hogyan változik az általam készített eszközök hatására. Ez három helyen érhető tetten:

- indexelési idő változása,
- index méretének változása,
- keresési idő változása.

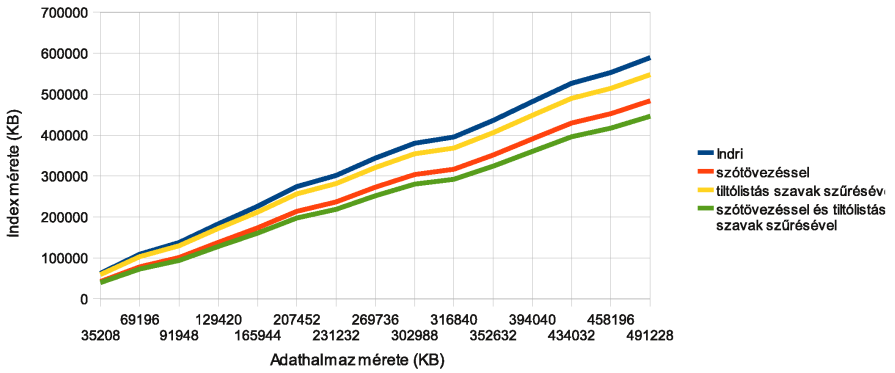


1. ábra. Indexelési idő az adathalmaz méretének függvényében

Indexelési idő. Az indexelési idő mérésénél a módszer az volt, hogy a dokumentumhalmazt valamilyen szempontok mentén csoportosítom, és mérem a rendszer indexeléssel töltött idejét az egyes halmazokon. Az ábrán jól látszik, hogy a szótövezés jelentősen lassította az indexelés folyamatát, viszont az is megállapítható, hogy a tiltólistás szavak szűrésével ez a teljesítménybeli visszaesés mérsékelhető.

Index mérete. Inkrementális indexelést végezve azt vizsgáltam, hogy hogyan változnak a készített indexek méretei, az eszközök használatának, illetve az adatmennyiség függvényében. Ennek a mérésnek az eredményét szemlélteti a 2. ábra.

Mint arra számítottam, az index mérete csökkent szótövezés és stopword-szűrés esetén. Tiltólistás szavak szűrésének alkalmazásakor, ez annak az egyszerű ténynek a következménye, hogy kevesebb kifejezés kerül az indexbe. Szótövezéskor pedig egy szó különböző ragozott alakjai ugyanahhoz a kifejezéshez tárolódnak. Az így nyert tárhelynövekedés a program eredeti működéséhez viszonyítva (átlagosan):



2. ábra. Index mérete az adathalmaz méretének függvényében

– szótövezéssel 22%,
– tiltólistás szavak szűrésével 6%,
– szótövezéssel és tiltólistás szavak szűrésével 28%.

Így tehát megállapítható, hogy e két eszköz használatával érzékelhetően lehet javítani egy keresőrendszer tárhelygazdálkodásán.

4.2. Relevancia

Módszer. A használt keresőrendszer eredménye egy rendezett lista, így az eredményesség mérésére az F-mérték csak korlátozott mértékben alkalmas. A kereső működéséhez jobban illeszkedő metrikák közül [1,3] az egyik legjobb tulajdonságokkal bíró az ún. *Mean Average Precision*[15]. Ezt a következő módon számolhatjuk: $\frac{\sum_{r=1}^{N}(P(r)\cdot \mathrm{rel}(r))}{R}$, ahol N a találati listában szereplő elemek száma, $rel(x)$ értéke 1, ha az x-edik elem a listában releváns, egyébként 0, $P(x)$ a pontosság a találati lista első x elemére szűkítve, R pedig a tárban található releváns dokumentumok számossága. A metrika szerint egy lekérdezés akkor 100%-os, ha az összes releváns elemet visszaadja, és ezek a találati lista legelején vannak (tehát nincs olyan dokumentum, ami nem releváns és megelőz egy relevánsnak mondottat). Látható, hogy ez a módszer nem érzékeny a zaj növekedésére, ha az a találati lista végén történik.

A mérésekhez minden esetben megfogalmaztam egy természetes nyelvű információigényt, majd azt az adott eszközökkel a lehető legpontosabban kódoltam. Mindehhez elkülönítettem a korpusz egy részhalmazát, mely elemeiről tudtam, hogy melyik milyen információigényhez hogyan kapcsolódik. A teszteket két indexen futtattam, összesen háromfélét[16]:

[15] A továbbiakban MAP.
[16] A továbbiakban 1., 2., illetve 3. mérés.

1. egyszerű[17] indexen, nyelvi eszközöket nélkülöző lekérdezéseket,
2. tövezett indexen a szótövezés és szinonimahasználat által nyújtott lehetőségeket használó lekérdezéseket,
3. tövezett indexen az előbbieken kívül az intelligens dátum- és mértékkeresés, illetve a személyek keresése használatával megfogalmazott lekérdezéseket.

A mérések készítése során felmerülő probléma, hogy hogyan is lehet objektíven leképezni az információigényt a kereső lekérdezőnyelvére. Érdekes kérdés, hogy mit feltételezek a felhasználóról:

- Mit tud a rendszerről? Mennyire ismeri a kereső lekérdezőnyelvét?
- Milyen forrásból fogalmazza meg a keresőkifejezést? (Az információigényt reprezentáló kérdés szavaiból vagy annak tágabb kontextusából?
- Milyen, a kérdéshez kapcsolódó többletinformációval rendelkezik a felhasználó?)
- Hány lekérdezést engedek futtatni? (Egy lekérdezés eredményének felhasználásával esetleg jobb keresőkifejezést lehet megfogalmazni.)

A tesztek futtatása során azzal a feltételezéssel éltem, hogy a felhasználó a lehető legteljesebben ismeri a lekérdezőnyelvet, illetve adott esetben rendelkezik a kérdéshez kapcsolódó plusz információval, és ezeket akár tágabb kontextusban is képes megfogalmazni. Viszont megszorításként csak egy lekérdezést futtattam egy-egy információigény megválaszolására.

Mérési eredmények. Korábbi méréseim [6] során már megmutattam, hogy szótövező használata érzékelhetően tudja növelni a kereső relevancia szerinti eredményességét, persze egyes esetekben (pontatlan keresőkifejezések használatakor) a zaj is növekszik. Méréseim megerősítettek abban a feltevésben, miszerint a lemmatizálással kombinált megfelelően kontrollált szinonimahasználat jelentős mértékű javulást tud eredményezni. A tesztek során az 1. mérésben az Indri eredményessége a MAP metrikával 71%-os volt. Ez azt jelenti, hogy az esetek nagy részében a releváns elemek a találati lista elején foglaltak helyet. Ezen sikerült javítani a 2. mérés folyamán, ami ezzel a metrikával 81%-os eredményt jelent. Szinonimákat használva azt tapasztaltam, hogy a keresőkifejezés nagyobb információtartalommal bíró szavaira alkalmazva jelentősen tud javítani az eredményeken (akár 20-50%-ot is). Ellenben egy-egy olyan szó esetén, mely a korpuszt tekintve gyakorinak mondható (pl.: beszéd, felszólalás), szinonimákat használni nem érdemes, mert ezekkel csak több zajt generálunk. A 3. esetről általánosan elmondható, hogy összességében itt is sikerült még további javulást eredményezni, így az említett módszerrel 85%-os eredményt kaptam. Mint várható volt, itt azon keresések eredményességén sikerült érdemben javítanom, amik valamilyen személlyel, mennyiséggel, időponttal kapcsolatos információt tartalmaztak. Elmondható még, hogy ha a felhasználó az információigényéhez kapcsolódóan többlettudással rendelkezik, úgymint 'kinek a felszólalásában található a kívánt információ?', 'kb. mikorra tehető a felszólalás időpontja?', 'milyen időpontok kap-

[17] szótövezést nem használó indexelővel készített

csolódnak a kérdéshez?', 'milyen mennyiségek (és milyen nagyságrendben) hozhatóak kapcsolatba a kérdéssel?', akkor további javulással számolhatunk. Összességében megállapíthatom, hogy a használt eszközökkel egy felhasználóbarátabb, relevánsabb találatokat produkáló rendszert sikerült építenem.

5. Összefoglalás

A cikkben bemutattam, hogy miként sikerült egy nyílt forráskódú keresőrendszert nyelvi eszközökkel bővíteni: várakozásaimnak megfelelően csak kisebb technikai akadályokba ütközött a rendszer bővítése. Az eredeti tervvel ellentétben nem volt lehetséges NLP-eszközök teljes integrációja kizárólag az Indri API-ját használva, így egyes esetekben módosítanom kellett a rendszer belső működését. Mindenképpen sikernek könyvelem el, hogy a rendszer új, eddig nem elérhető eszközökkel bővült, továbbá azt, hogy ezek javítanak, javíthatnak a kereső eredményességén. Az indexelési időben felmerült negatív eredményekről elmondható, hogy azok a végfelhasználót nem érintik, tehát érdemben nem rontottam a rendszer teljesítményén.

Hivatkozások

1. C. J Van Rijsbergen: Information Retrieval. London, Butterworths (1979).
2. Christian Middleton, Ricardo Baeza-Yates: A comparison of open source search engines. Jelentés, Universitat Pompeu, Fabra Department of Technologies (2007).
3. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information retrieval. New York, Cambridge University Press (2008).
4. Emmanuel Eckard, Jean Cédric Chappelier: Free Software for research in Information Retrival and Textual Clustering. Jelentés, Ecole Polytechnique Fédérale de Lausanne (2007).
5. Nagy István Zoltán: A Magyar Országgyűlés plenáris üléseinek annotált korpusza. Diplomamunka, Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar, Budapest (2008).
6. Orosz György: Nyelvtechnológiai alkalmazások integrálása keresőmotor(ok)ba. Diplomamunka, Eötvös Loránd Tudományegyetem, Informatikai Kar, Budapest (2010).
7. Prószéky Gábor, Kis Balázs: Számítógéppel emberi nyelven. Bicske, SZAK Kiadó (1999).
8. Thorsten Brants: Natural language processing in information retrieval. Konferenciakiadvány, 14th Meeting of Computational Linguistics in the Netherlands (2003).
9. Trevor Strohman, Donald Metzler, Howard Turtle, W. Bruce Croft: Indri: A language-model based search engine for complex queries. Konferenciakiadvány, International Conference on Intelligence Analysis (2004).

V. Beszédtechnológia

Környezetfüggetlen és sztochasztikus nyelvtanok összehasonlítása többnyelvű gépi beszédfelismerési feladatban

Mozsolics Tamás, Tarján Balázs, Mihajlik Péter, Fegyó Tibor

Távközlési és Médiainformatikai Tanszék,
Budapesti Műszaki és Gazdaságtudományi Egyetem
{mozsolics, tarjanb, mihajlik, fegyo}@tmit.bme.hu

Kivonat: A szituációs beszédfelismerés egyik legfontosabb eleme a szituációhoz jól alkalmazkodó beszédfelismerő hálózat tervezése. Ezért megvizsgáltunk néhány hálózatépítési módszert, hogy összehasonlítsuk teljesítményüket. Az építés és tesztelés folyamatát összesen hat nyelven végeztük el: angol, francia, magyar, német, olasz és spanyol. Tesztelés céljából a telefonos hálózaton keresztül az utcáról vagy járműből rögzített, tájékozódási célú kérdésekből és kijelentésekből álló adatbázist használtunk. Magyar, német, olasz és spanyol nyelvekre összehasonlítottuk a fonéma- és grafémaalapú tervezési technikákat, s a magyar modellt különböző paraméterek változtatása mentén is vizsgáltuk. A hálózatokat saját fejlesztésű, WFST-s modellező rendszeren építettük, saját felismerőn futtattuk és HTK-val értékeltük ki.

1 Bevezetés

A TELEAUTO projekt célja egy olyan tájékozódási szolgáltatás biztosítása az autósok számára, ahol a gépkocsivezető szóban kérhet segítséget egy céllal kapcsolatban, ahová el szeretne jutni. A kéréseket egy külső helyszínen, egy kétszintes kiszolgáló rendszer várja, s rájuk első körben egy számítógép próbál válaszolni, s amennyiben ez sikertelen, akkor a gépi rendszer továbbküldi a kérést a diszpécsernek. Válaszként mindkét esetben a kívánt cél GPS- (Global Positioning System) koordinátáit kapja az autóban található navigációs eszköz vissza.

Ebben a cikkben mi a projekt gépi kiszolgáló moduljának tervezésével s megvalósításával foglalkozunk. A gépi modul lelke az automatikus beszédfelismerő szoftver (ASR: Automatic Speech Recogniser), mely úgy működik, hogy egy előre betöltött ún. beszédfelismerő hálózat mondataihoz illeszti a bejövő kérést, s ezek közül kiválasztja a legjobban illeszkedőt. A beszédfelismerő hálózat az adott szituációban általunk várt mondatok gyűjteménye, melyből a beszédfelismerő szoftver mindig egyet választ. Ahhoz, hogy a gépi alapú kiszolgálás hatékony legyen, alapvetően két dologhoz, az autó akusztikai környezetéhez és a beszédszituációhoz kell jól alkalmazkodni. Ezek közül az első jelfeldolgozási, konkrétan szűrési, a második pedig a beszédfelismerő hálózat építéséhez kapcsolódó feladat. A cikk ez utóbbit tárgyalja.

2 Nyelvi modell építése

A kitűzött feladatból látható, hogy a projektben található gépi beszédfelismerőtől nem várjuk el, hogy adott nyelven bármit megértsen, sőt azt sem, hogy pontosan megértse a kérés minden részletét, elég, ha a célpontot jól megérti. Nincs szükség általános szöveget leíró beszédfelismerő hálózatra, sőt egy adott szituációra optimalizált modell sokkal hatékonyabb lehet. Mivel a téma elég speciális, általában nem áll rendelkezésre adatbázis a szituációban előforduló mondatokról, így nekünk kell azt összegyűjteni minden nyelvre.

2.1 Szituációhoz tartozó mondatok gyűjtése

Nézzünk meg néhány célpontkeresésre irányuló, várható példamondatot magyar nyelvre:

„Hol van a közelben McDonald's?”
„Hol van egy könyvesbolt?”
„Hol lehet egy OTP-automata?”
„Hol van a közelben kórház?”
„Hol található a közelben Tesco?”
„Hol van a Holokauszt Múzeum?”

Elméletileg persze végtelenféle mondat lehet, s biztosan nem tudjuk összegyűjteni mindet, de minél többet sikerül, annál jobb lesz a modell. Szerencsére az a megfigyelés, hogy az előforduló kérések szerkezete független a konkrét célponttól jelentősen csökkenti a variáció számát, hisz nem kell az összes *Hol van?* típusú kérdést felsorolni, s így időt, munkát, memóriát spórolhatunk. Ketté kell tehát választani a gyűjtési feladatot tipikus mondatszerkezetek és célpontok gyűjtésére. Ennek tükrében az előző példa mondatszerkezetei, illetve célállomásai (a [cél] változó) így festenek:

„Hol van a közelben [cél]?”	*„McDonald's/POI”*
„Hol van egy [cél]?”	*„könyvesbolt/POI”*
„Hol lehet egy [cél]?”	*„OTP/POI automata/POI”*
„Hol van a közelben [cél]?”	*„kórház/POI”*
„Hol található a közelben [cél]?”	*„Tesco/POI”*
„Hol van a [cél]?”	*„Holokauszt Múzeum/POI”*

Ez a fajta szétválasztás több szempontból előnyös. Egyrészt megkönnyíti a tervezést és bővítést, másrészt a külön célállomás (POI: Point Of Interest) lista kompatibilis a diszpécserközpont adatbázisával, mivel gépi felismeréskor lényegében csak a célpontok pontos felismerésére törekszünk, ezen célok szavai könnyen felcímkézhetők (l. /POI címkék a célpontok szavai után), megkönnyítve a gépi felismerést, lényegkiemelést. Ekkor egy *Hol van a közelben McDonald's/POI?* típusú találatból könnyedén kiemelhető a célállomásra vonatkozó rész, illetve ezen részek jelenlétében/hiányában könnyedén eldönthető, hogy lehet-e az adott gépi felismerésnek használható eredménye-e vagy sem. A gyakorlatban a nyelvi modellek ehhez a szituációhoz lényegében az itt leírt módon készültek, annyi különbséggel, hogy a mondatszerkezetek egy hatékonyabb leíró formátumban, a Phoenix cég Parser nevű rendsze-

rében (l. [6]) definiált GRA formátumban lettek a mondatszerkezetek definiálva, illetve a célpontlista, több különálló kategóriára lett bontva, pl.: bevásárlás, szállók, étkezdék. Készítsünk az előző példánkból GRA-modellt. Ehhez először rendezzük egymás mellé a hasonló mondat szerkezeteket, s ebből a GRA-modell:

„Hol van a [cél]?”
„Hol van a közelben [cél]?”
„Hol van egy [cél]?”
„Hol lehet egy [cél]?”
„Hol található a közelben [cél]?”

```
[varhato_keres]
    (Hol VAN NEVELO [cel]?)
VAN
    (van)(lehet)(található)
NEVELO
    (a)(a közelben)(egy) ;
```

A GRA-modellben a [ ] zárójelbe tett kifejezés a makró definíciót jelent, a makró *„ ; ”* jellel zárul. Az egyes makrókban definiálhatunk változókat a változatok leírására, ezeket csupa nagybetűvel írjuk, s még a makrón belül kifejtjük. A makrók, illetve változók által leírt egyes változatok ( ) zárójelbe kerülnek. A *„ *”* jel, azt jelenti, hogy *„vele vagy nélküle”*, tehát mindkét eset előfordulhat. A GRA formátum láthatóan nem támogatja az ékezetes karakterek használatát változó- és makródefiníciók esetén.

2.2 A CFG és az N-gram modellek

A CFG (Context Free Grammar) lényegében a 2.1. bekezdésben leírt módon összegyűjtött mondatszerkezetekből épített aciklikus modell, melybe behelyettesítjük a POI-listát.

Az N-gram modell a CFG-től eltérően már ciklikus felépítésű, melynek átmeneti valószínűségei ebben az esetben a CFG-ben összegyűjtött mondatokból mint tanítószövegből lettek tanítva egy simítási eljárást követően. Mindezt a CMU Logios nevű szoftverével végeztük el.

Ez a modell szerkezetéből adódóan elvileg toleránsabb az adott szituáció olyan nem várt mondataival szemben, ahol a várt szavak szerepelnek ugyan, de a várttól kissé eltérő sorrendben és/vagy kissé eltérő mondathossz mellett.

2.3 Emberi nyelv – gépi nyelv

Mi, emberek a beszédről szóegységekben gondolkodunk, s így a mondandónkat szósorozatok formájában fogalmazzuk meg. Ez számunkra természetes, így modelljeinket is szó-, vagy morfémaalapon készítjük. Viszont a számítógép szóalapú hálózathoz nem tud pontosan illeszteni. Ennek oka, hogy a szó mint alapegység számát és hosszát tekintve túl variábilis, illetve egy-egy mondat relatíve kevés szóból, illetve morfémából épül fel.

A beszédfelismerésre használható hatékony gépi modell – a beszéd sztochasztikus jellegéből adódóan – valószínűség-számítás alapú, s rejtett Markov-modellek (HMM: Hidden Markov Model) az elemei. Ilyen szövegkörnyezetben az illeszkedés jelentése, hasonló valószínűségi paramétervektorok birtoklása. Mivel az emberi nyelv, avagy a hálózattervezés hatékony szintje (szavak szintje), s a gépi nyelv, avagy a gépi beszédfelismerés hatékony szintje (HMM-ek szintje) nem esik egybe, ezért a megtervezett hálózatainkat át kell vinni a HMM-szintre a felismerő szoftverbe töltést megelőzően.

Ez a transzformáció három lépésben végezhető el, melyeket fonetikus átírásnak, környezetfüggősítésnek, illetve nyelvimodell-beillesztésnek nevezik. Trigráf alatt – a trifón analógiára – a szomszédjaitól mint környezettől függő grafémát értjük.

$$\text{szó sorozat} \rightarrow \left\{ \begin{array}{lcl} \text{fonéma sorozat} & \rightarrow & \text{trifón sorozat} \\ \text{graféma sorozat} & \rightarrow & \text{trigráf sorozat} \end{array} \right\} \rightarrow \text{HMM sorozat}$$

1. ábra. A nyelvi modellezés 4 szintje.

3 WFST Framework

A WFST Framework egy olyan egységes matematikai keretrendszer, melynek elemei speciális, címkézett és súlyozott irányított gráfok ún. WFST-k (Weighted Finite State Transducer) s a rajtuk végezhető műveletek. Ebben a matematikai modellben a hálózatoptimalizálás és a 2.3. bekezdésben látott szintlépések is elvégezhetők, l. [1, 2].

Egy beszédfelismerő hálózat mint WFST szerkezetileg optimális, ha determinisztikus, vagyis egy adott bemeneti sorozat egyértelműen meghatározza, hogy merre menjünk benne, minimális, vagyis a lehető legtömörebben épített és súlyaiban sztochasztikus, vagyis egynemű, kiugró értékektől mentes. A beszédfelismerő hálózat építését teljes egészében a 2. ábrán látható műveletsor írja le.

$$\underbrace{\underbrace{\underbrace{\underbrace{\text{H o wpush(min(det(C o det(L o } \underbrace{\text{G}}_{\text{szó szintű modell}} \text{))))}_{\text{fonetikus szintű modell}}}_{\text{trifón szintű modell}}}_{\text{HMM szintű modell}}$$

2. ábra. A nyelvi modelljeinkhez használt nyelv független WFST-műveletsor.

Ahol, a G (Grammar) transzducer a szószintű nyelvi modellünk, az L (Library) a nyelvi modellben szereplő szavak fonetikus vagy grafémás átiratait tartalmazó szótár, a C az ún. környezetfüggősítő (Context-Dependency) transzducer és a H (HMM-Library) az adott nyelv trifónjainak/trigráfjainak HMM-es átiratait tartalmazó szótár. Az *o* jelöli az ún. kompozíció műveletet, mellyel az egyes szintlépések elvégezhetők. A *det* gráfok determinizálásához hasonlóan a WFST egy olyan átépítését jelenti olyan ekvivalens WFST-vé, melyben a bemeneti szimbólumsorozatnak megfelelő haladás mindig egyértelmű. Sajnos ez a determinizált tulajdonságú ekvivalens WFST-k esetében nem mindig létezik, l. [5]. A *min* alatt itt olyan műveletek csoportját értem, melylyel az eredeti WFST-vel ekvivalens tömörebb WFST állítható elő. Ez lehet gráfminimalizáció és címkéket okosabban rendező/összevonó algoritmus is. A *wpush* a súlyok egyenletesebb eloszlását biztosítja a WFST-nkben.

A 2. ábrán látható műveletsor fontos tulajdonsága, hogy nyelvfüggetlen, ezért csak a H, C, L és G transzducereket kell előállítanunk minden nyelvre. Aciklikus nyelvi

modell esetén (pl.: CFG) a fonémaszintű modell determinizációja minimalizációval helyettesíthető a még tömörebb hálózat érdekében. Mivel az [1, 2] irodalmak egyértelműen leírják, hogy kell a H, C és L transzducereket felépíteni, ezért koncentrálunk mi is elsősorban ebben a cikkben a G nyelvi modell építésének ismertetésére.

4 Nyelvfüggő kihívások

Annak ellenére, hogy a 3. fejezetben megmutattuk, hogy a modellezési módszer nyelvfüggetlen, a tudásforrások összeállításánál akadnak nyelvfüggő részproblémák, mint pl.: helyhatározó és tárgyragok a magyarban, vagy a különböző nemű szavak névelői a németben.

4.1 Hely-és tárgyragok a magyar nyelvben

A TELEAUTO-s szituációra összegyűjtött mondatszerkezetek java részében a célpontok, tárgyként vagy helyhatározóként szerepelnek, hisz valamelyik *áruház**hoz***, *reptér**re***, *mozi**ba*** mennénk, s *éjjel-nappali**t*** keresünk.

Az természetesen ésszerűtlen elvárás lenne, hogy több tízezer POI-nak összes ragozott alakját kézzel állítsuk elő. Szerencsére ezen ragok (tipikusan a *-t*, *-ba/be*, *-ra/re* és *-hoz/hez/höz*) megfelelő alakjának kiválasztása jól automatizálható. Tekintsük át pl. a *-ba/be* ragok közötti választásra megadott alábbi szabályokat:

```
;a magánhangzók osztályozása
mghL == a á o ó u ú; mély
mghH == e é i í ö ő ü ű
            ; magas
;a ba/be ragozás szabályai
;1. hasonulások
e[ba] = é b e;
mghL-e[ba] = é b a;
mghL,msh-e[ba] = é b a;
a[ba] = á b a;
az[ba] = a b a;
ez[ba] = e b e;
```

```
;2. az utolsó magánhangzó dönt
  mghL-[ba] = b a;
  mghL,msh-[ba] = b a;
  mghL,msh,msh-[ba] = b a;
  mghH-[ba] = b e;
  mghH,msh-[ba] = b e;
  mghH,msh,msh-[ba] = b e;
  ;3.a korábbi mélymagánhangzó dominanciája
  mghL,mghH-[ba] = b a;
  mghL,mghH,msh-[ba] = b a;
  mghL,msh,mghH-[ba] = b a;
 mghL,msh,mghH,msh-[ba] = b a;
```

melyek közül a szabályok hosszúságuk szerinti sorrendben kerülnek sorra, tehát a szoftverek először mindig a hosszabbakat próbálják illeszteni. A módszer elve, hogy a magánhangzókat két osztályra, mély és magas magánhangzókra (mghL és mghH változók), s a szavak *-ba/be* ragot közvetlen megelőző része alapján (– előtti rész) hoz döntést. Nézzünk meg néhány példát:

```
mozi[ba] = mghL,msh,mghH-[ba] = b a
pizzéria[ba] = a[ba] = á b a
IKEA[ba] = a[ba] = á b a
parkoló[ba] = mghL-[ba] = b a
Debrecen[ba] = mghH,msh-[ba] = b e
edzőterem[ba] = mghH,msh-[ba] = b e
Allee[ba] = mghL,msh-e[ba] = é b a
```

Ezt a megoldást a magyar nyelvi modellekhez készített szótárakból egy 266 szavas részszótáron teszteltük, s csak 15-ször hibázott, ami kb. 95% pontosságnak felel meg.

5 Tesztelés

A beszédfelismerő hálózat építéséhez és teszteléséhez használt források fontosabb adatai kiolvashatók az 1. táblázatból.

A hat nyelven készített modellek építésekor nyelvenként átlagosan körülbelül 25000 mondat szerkezetet sikerült összegyűjteni/generálni. Az összehasonlítós tesztek nagyjából 500 POI-val készültek, átlagosan 900 szavas teljes szótárméret mellett. A teszteléshez összesen 56 beszélő által felmondott, nagyjából 500 felvétel állt rendelkezésünkre. Ezek mindegyike valós körülmények között, gépkocsiban felvett, telefonos beszélgetés minőségű felvétel volt. A felvételeket nyelvenként számos, különböző nemű, korú, natív beszélőtől rögzítettünk.

5.1 Felismerési pontossági jellemzők

A tesztelés táblázataiban tipikusan az alább ismertetett 5 paraméter jelenik meg.

Az S_{ACC}, illetve $S_{ACC,POI}$ azt mutatja meg, hogy a mondatok, illetve az POI-kkal kapcsolatos mondatrészek hány százalékát sikerült hibátlanul felismerni. Hasonlóan értelmezhetők a W_{ACC}, illetve $W_{ACC,POI}$ paraméterek az elhangzott szavakra.

Azt, hogy a felismerési idő hogyan arányul a tesztszöveg időbeli hosszához, mutatja meg az RTF, az ún. Real-Time Factor, tehát pl. ha RTF=0.2, akkor az elhangzási idő ötöde kell a feldolgozáshoz.

1. táblázat: A teszteléshez használt modellek és felvételek főbb paraméterei.

nyelvi modell	angol	francia	magyar	német	olasz	spanyol	összes
mondatszerkezetek	36746	4066	18782	68296	5425	17604	25153
POI-k száma	501	501	518	519	506	529	512
szótárméret	624	727	1886	677	689	748	892
átírás adatai	angol	francia	magyar	német	olasz	spanyol	átlag
fonémák száma	44	35	38	39	57	25	39,66
grafémák száma	--	--	33	31	29	33	31,50
akuszt. modell	angol	francia	magyar	német	olasz	spanyol	átlag
tanítószöveg[óra]	17,8	57,9	28,9	62,1	93,7	56,5	52,81
tesztfelvételek	angol	francia	magyar	német	olasz	spanyol	összes
száma	57	40	266	26	70	28	487

5.2 Hardver- és szoftverkörnyezet

2. táblázat: A fonetikus átírásokhoz használt szoftverek és szótárak 6 nyelvre.

angol	francia	magyar	német	olasz	spanyol
M-Phon	Liaphon	M-Phon	Txt2pho	M-Phon	Txt2pho

A teszteléshez használt hardver egy T7300 nevű Core2Duo architektúrájú, 2 GHz-es processzorú laptop volt, 2 GB RAM-mal, 32 bites Windows operációs rendszer alatt. A beszédfelismerő hálózatokat a saját készítésű M-System nevű szoftverrendszerrel építettük, s a szintén saját VOXerver nevű felismerőn futtattuk le, majd az eredményeket HTK-val (l. [4]) értékeltük ki. A fonetikus átíráshoz a 2. táblázatban látható szoftvereket és szótárakat használtuk.

5.3 Teszttípusok

5.3.1 CFG vs. N-gram

Ebben a tesztben fonémaalapú CFG- és N-gram-modelleket hasonlítottunk össze 6 különböző nyelven. Mivel a modellek nagyjából hasonló kondíciók mellett 500 POI készültek, ez a teszt alkalmas a nyelvek osztályozására is e feladat kapcsán. A teszt eredményeit a 3. táblázatban foglaltuk össze.

Ha a két modellt akarjuk egymáshoz hasonlítani, akkor elsősorban a táblázat utolsó oszlopát érdemes megvizsgálni, ugyanis itt szerepelnek az eredményeknek a felvételek számával súlyozott átlagai. Ezek alapján elmondható, hogy a két megoldás között nincs különösebben nagy különbség. Az N-gram-megoldás a fő paraméterekben $W_{ACC,POI}$ és $S_{ACC,POI}$ 1-1.5%-al túlteljesíti a CFG-t pontosságban, cserébe a feldolgozási idő kb. 28%-kal nő.

Ha a különböző nyelvek egymáshoz képesti viszonyát tekintjük, akkor az angol az egyetlen, melynek paraméterei kb. 10%-nál jobban eltérnek a többitől. Ennek okai egyrészt a rendelkezésre álló relatíve kisebb adatbázis (l. 1. táblázat), másrészt az angol nyelv beszélt és írott alakja közötti hatalmas eltérés, mely a fonetikus átírás során nehezen leküzdhető feladat elé állítja a mérnököket.

3. táblázat: A fonémaalapú CFG- és N-gram-modellek eredményei 6 nyelvre.

CFG	angol	francia	magyar	német	olasz	spanyol	átlag
W_{ACC} [%]	55,16	73,86	78,16	68,71	73,11	79,78	73,98
$W_{ACC,POI}$ [%]	46,53	79,34	74,68	84,78	76,81	80,60	72,95
S_{ACC} [%]	16,07	40,00	57,42	34,62	37,14	46,43	46,39
$S_{ACC,POI}$ [%]	36,36	72,50	72,33	76,92	71,43	82,14	68,81
RTF [$*t_{valós}$]	0,344	0,108	0,164	0,191	0,072	0,140	0,167
N-gram	angol	francia	magyar	német	olasz	spanyol	átlag
W_{ACC} [%]	51,92	71,59	76,74	71,78	73,37	84,70	73,12
$W_{ACC,POI}$ [%]	48,51	75,21	75,78	84,78	80,43	86,57	74,31
S_{ACC} [%]	14,29	37,50	52,36	26,92	31,43	53,57	42,39
$S_{ACC,POI}$ [%]	36,36	72,50	72,98	73,08	77,14	78,57	69,58
RTF [$*t_{valós}$]	0,508	0,170	0,171	0,290	0,117	0,266	0,214

5.3.2 Szituációs modellek vs. POI-lista

Egy jogosan felmerülő kérdés lehet, hogy mennyivel kapnánk rosszabb eredményt, ha egyszerűen a POI-listából, a szituációhoz alkalmazkodó mondatszerkezetek nélkül építenénk fonémaalapú beszédfelismerő hálózatot.

Összevetettük hát ezt a szólistás hálózatot a korábbi két modellel magyar nyelv esetében, s ez a teszt a 4. táblázatban látható eredményeket hozta. Ha kiolvassuk a fő paraméterek ($W_{ACC,POI}$ és $S_{ACC,POI}$) értékeit, láthatjuk, hogy a szituációkhoz alkalmazkodó mondatszerkezetek alkalmazásával kb. 2.5-3-szor pontosabb felismerés lehetséges.

4. táblázat: A POI-listából épített fonémaalapú modell összevetése a CFG- és N-gram-modellekkel, magyar nyelvre.

CFG	szólista	CFG	N-gram
W_{ACC} [%]	12,18	78,16	76,74
$W_{ACC,POI}$ [%]	24,27	74,68	75,78
S_{ACC} [%]	00,00	57,42	52,36
$S_{ACC,POI}$ [%]	30,08	72,33	72,98
RTF [$*t_{valós}$]	0,201	0,164	0,171

5.3.3 Fonémás vs. grafémás

Ezek a tesztek csak négy nyelven (magyar, olasz, német és spanyol) készültek, mivel a grafémás modellek csak olyan nyelvek esetében működnek jól, ahol a kiejtés és az írott alak között elég szoros kapcsolat van.

Az 1. táblázatból látható, hogy a nyelveknek átlagosan kisebb a grafémakészlet nagysága, mint a fonémáké, ezért a grafémás modellektől tömörebb felépítést s valamivel pontatlanabb nyelvleírást, s ezáltal némileg pontatlanabb felismerési eredményeket vártunk.

A POI-k vonatkozásában sok külföldi eredetű szó lehetséges egy adott nyelvterületen, pl. a magyar modell teszteléséhez használt tesztfelvételeink esetében a 412 POI-hoz kapcsolódó szavunkból 55 (kb. 15%) volt külföldi (pl.: Erste, McDonald's, Renault), ezért a fonéma- és grafémaalapú modellek összehasonlítása csak úgy lehetett fair, ha mindkét esetben kivételszótárat használunk a külföldi eredetű szavakra.

Ezen szótárak közötti eltérés csak annyi, hogy míg a fonémaalapú megoldásnál az idegen szavak magyar kiejtése, addig a grafémaalapú megoldásnál azoknak megfelelő grafémás alak szerepel a kivételek között pl.: a *„Rossmann”* szó átiratai

```
Rossmann = r o sz m a n ;    //fonema alapu kivetel szotar
Rossmann = r o s s z m a n ; //grafema alapu kivetel szotar
```

magyar modellek esetében.

5. táblázat: A grafémaalapú CFG-modellek eredményei 4 nyelvre.

CFG	angol	francia	magyar	német	olasz	spanyol	átlag
W_{ACC} [%]	--	--	77,17	63,80	77,28	79,78	76,49
$W_{ACC,POI}$ [%]	--	--	72,02	82,61	84,78	83,58	75,85
S_{ACC} [%]	--	--	59,40	23,08	38,57	39,29	51,80
$S_{ACC,POI}$ [%]	--	--	72,08	73,08	77,14	78,57	73,52
RTF [$*t_{valós}$]	--	--	0,137	0,235	0,092	0,163	0,137

A teszt során kapott eredmények az 5. táblázatban láthatók. A nyelvek közül kettő (magyar és német) eredményei valamivel rosszabbak, a spanyolé nagyjából egyforma, az olaszé pedig meglepő módon jóval jobb lett, mint a fonémás változaté. Ez utóbbi oka valószínűleg a saját kezűleg összegyűjtött olasz átírási szabályok hiányosságaiban keresendő.

A két módszer átlagos pontosságáról a 6. táblázat alapján elmondható, hogy az nagyjából egyforma, s a fonetikus megoldás előnye feltehetően egy jobb olasz fonetikus átírási megoldás mellett sem több néhány %-nál.

Kivételszótárat mindenképpen ajánlott használni grafémás esetben is, hiányukban az adott 4 nyelvre nagyjából 10%-os idegen szó arány mellett a tesztfelvételeken átlagosan 5%-kal rosszabb $S_{ACC,POI}$, $W_{ACC,POI}$ eredményt kaptunk a fonémás megoldáshoz képest.

6. táblázat: A graféma- és fonémaalapú modellek súlyozott átlaga 4 nyelvre.

Felvételek számával súlyozott átlagok	grafémás CFG	fonémás CFG
W_{ACC} [%]	76,49	76,74
$W_{ACC,POI}$ [%]	75,85	76,16
S_{ACC} [%]	51,80	51,47
$S_{ACC,POI}$ [%]	73,52	73.18
RTF [$*t_{valós}$]	0,137	0,148

5.3.4 A felismerési pontosság függése a POI-k számától

Fontos kérdés, hogy a POI-k számának növelésével, hogyan alakulnak a felismerési pontossági értékek. Nyilvánvalóan csökkeni fognak, hisz a POI-k növekedésével egyre több hasonló nevű célpontot kapunk, melyek között egyre nehezebb választani. Nem mindegy azonban, hogy ez a csökkenés milyen függvény szerint és milyen ütemben történik. Ennek vizsgálatához az 5.3.1. bekezdés fonetikus magyar CFG-modelljének POI-készletét bővítettük 5 lépésben egészen 5000-ig.

7. táblázat: A felismerési pontosság jellemzőinek alakulása az $500<N_{POI}<5000$ tartományban, magyar CFG-modell esetén.

POI-k száma	500	1000	1500	3000	4000	5000
W_{ACC} [%]	77,28	76,71	76,63	75,18	73,41	72,35
$W_{ACC,POI}$ [%]	72,41	71,03	69,77	67,09	65,47	61,95
S_{ACC} [%]	56,25	54,90	54,90	53,54	49,21	46,85
$S_{ACC,POI}$ [%]	70,75	69,29	68,90	67,33	65,60	63,05

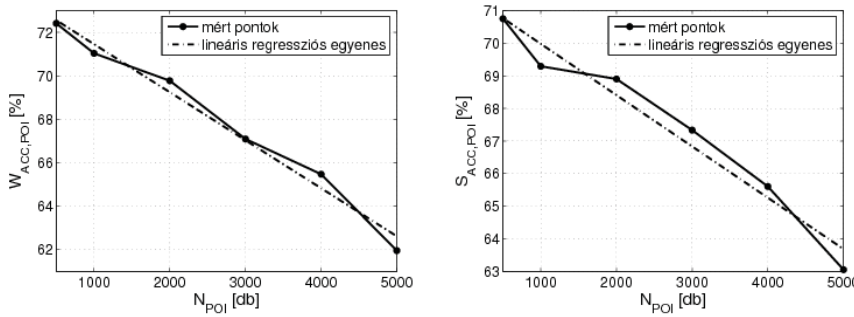
A tesztek eredményei kiolvashatók a 7. táblázatból. Az itt látható 4 pontossági paraméter közül minket leginkább a $W_{ACC,POI}$ és $S_{ACC,POI}$ alakulása érdekel, hisz az előbbi azt adja meg, hogy az esetek hány százalékában találunk a POI-hoz kapcsolódó szavakat, illetve hány százalékban találjuk meg az egyes POI-khoz tartozó összes szót. Ezen jellemzők alakulásának könnyítése végett, az $500<N_{POI}<5000$ tartományban a mért értékeket analitikus tesztfüggvényekkel közelítettük regressziót alkalmazva. A lineáris regressziós tesztfüggvények (lásd a 4. ábrát) mindkét esetben, pontosságban felülteljesítették az exponenciális, illetve hatványkitevős társaikat, ezért kijelenthető, hogy a vizsgálati tartományban az N_{POI} értékének növelésével a $W_{ACC,POI}$ és

$S_{ACC,POI}$ felismerési pontosságok egyenletes ütemben csökkennek. A közelítő egyenesek pontos egyenletei megtalálhatók a 8. táblázatban. Ezen egyenletek alapján a POI-k számának 1000-rel növelése 2,21% $W_{ACC,POI}$, illetve 1,57% $S_{ACC,POI}$ csökkenéssel jár.

Mondhatjuk persze, hogy még 5000 POI sem túl sok, s hogy az autók navigációs rendszere ennél nyilván többet tartalmaz. Ennek ellenére a POI-k száma jól kordában tartható, ha minden esetben az autótól egy bizonyos hatósugarú körben levő célokra szűkítjük a keresést.

8. táblázat: A $W_{ACC,POI}$ és $S_{ACC,POI}$ felismerési pontosságot közelítő lineáris regressziós egyenesek egyenletei az $500<N_{POI}<5000$ tartományban, magyar CFG-modell esetén.

	regressziós egyenes egyenlete
$W_{ACC,POI}$ [%]	$73.6610 - 0.002209 \cdot N_{POI}$
$S_{ACC,POI}$ [%]	$71.5529 - 0.001574 \cdot N_{POI}$



3. ábra. A $\mathbf{W_{ACC,POI}}$ és $\mathbf{S_{ACC,POI}}$ felismerési pontosságot közelítő lineáris regressziós egyenesek az $500<N_{POI}<5000$ tartományban, magyar CFG modell esetén.

5.3.5 A felismerési hiba szórása az egyes szituációkban

Az 5.3.1-5.3.4. bekezdésekben kaptunk átlagos felismerési pontossági adatokat egy-egy adott nyelvre készített adott típusú modell esetén. Eddig arról viszont nem volt szó, hogy mekkora eltérések lehetnek egy-egy eltérő szituációban (eltérő jel-zaj viszony, zavaró környezeti zajok, pl. mentőautó hangja) ezen átlagértékektől.

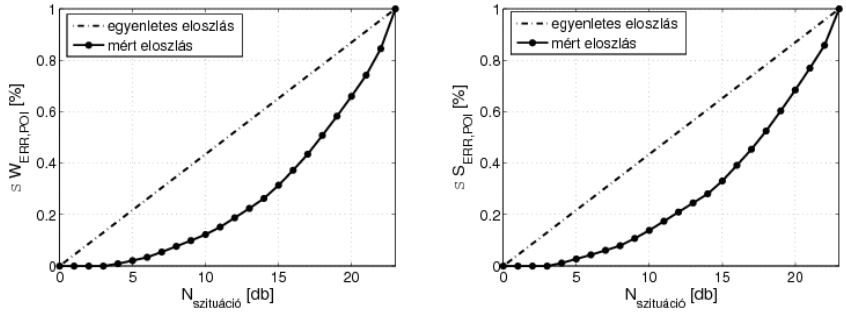
Ennek szemléltetéséhez a fonémaalapú magyar CFG-modellt szituációnként is leteszteltük. Az eredetileg 27 felvételsorból összevontuk az 5 legkisebbet, melyekhez 5-nél kevesebb felvétel tartozott, s előállítottuk 23 különböző szituáció átlagos $W_{ERR,POI}$ ($=1-W_{ACC,POI}$) és $S_{ERR,POI}$ ($=1-S_{ACC,POI}$) paramétereit.

9. táblázat: A fonémaalapú magyar CFG-modell 23 különböző szituáció alapján számított statisztikus paraméterei.

	átlag	terjedelem	szórás
$W_{ERR,POI}$ [%]	26,51	100	20,56
$S_{ERR,POI}$ [%]	28,63	100	21,44

Ezen tesztelés a 9. táblázatban összefoglalt eredményei alapján elmondható, hogy a $W_{ERR,POI}$ és $S_{ERR,POI}$ paraméterek a [0%,100%] tartományban mozogtak, átlagos értékük (ahogy azt már korábbról is ismertük) 26.5-28.5% volt, s az egyes szituációkban ettől 20.5-21.5%-kal tértek el átlagosan.

A koncentráltságról a 4. ábrán látható eloszlási függvények segítségével tájékozódhatunk. Az ábrán az átlót követő függvény képviseli az egyenletes hibamegoszlást az egyes szituációk között, az alattuk futó konvex alakú függvények pedig a mért eloszlás függvények. Ezek alapján pl. a 23-ból a 6 legrosszabb eredmény felelős az összes $W_{ERR,POI}$ és $S_{ERR,POI}$ hibák 50%-áért, a 7 legrosszabb pedig a hibák 60%-áért.



4. ábra. Az összes hiba eloszlása a fonémaalapú magyar CFG-modell 23 különböző szituációjában.

5.3.6 A felismerési pontosság függése a mondatszerkezetek számától

Miután az 5.3.2. bekezdésben láthattuk, hogy hasznos, hogy POI-listánkat az adott szituációhoz illeszkedő mondatszerkezetekbe illesztjük, érdemes lenne megvizsgálni, hogy ezen szerkezetek variációinak száma, hogy hat a felismerési pontosságra.

A 10. táblázatból látható, hogy a mondatszerkezetek számának, s ezáltal közvetve a találati arány csökkenésével az N-gram felismerési pontosság előnye, rugalmas szerkezeti felépítésének köszönhetően, a korábbi 1-2%-ról 5-6%-ra emelkedett.

10. táblázat: A felismerési pontosság alakulása a mondatszerkezetek számának csökkenésével fonémaalapú magyar CFG és N-gram esetén.

fonéma CFG				
rel. mondatszerkezetszám	1,000	0,841	0,682	0,540
$W_{ACC,POI}$ [%]	72,41	71,90	55,22	45,38
$S_{ACC,POI}$ [%]	70,75	70,36	56,18	42,74
fonéma N-gram				
rel. mondatszerkezetszám	1,000	0,841	0,682	0,540
$W_{ACC,POI}$ [%]	73,18	70,08	61,44	49,41
$S_{ACC,POI}$ [%]	70,97	68,13	60,24	47,22

6 Összefoglalás

A TELEAUTO projekt beszédalapú tájékozódást segítő szolgáltatás készítését tűzte ki céljául autóvezetők számára, növelve ezzel biztonságukat és komfortérzetüket. Segítségkéréskor egy külső központot hívhatunk, s egyszerű hétköznapi kérésekért cserébe az autónk navigációs rendszere a kívánt cél koordinátáit kapja vissza.

Ennek a rendszernek fontos eleme egy automatikus beszédfelismerő rendszer, mely jó esetben jelentősen csökkenti a diszpécserek munkamennyiségét, azáltal, hogy a beérkező kérdések jelentős hányadát megválaszolja. Mivel a felismerő szoftver a beérkezett kéréseket egy előre betöltött beszédfelismerő hálózathoz hasonlítja, a megoldás kulcsa ennek a hálózatnak a tervezésében rejlik.

Bemutattuk a WFST Framework nevű matematikai modellt, s annak keretein belül történő beszédfelismerő hálózatépítés előnyeit, mely szerint egyszerű, nyelvfüggetlen megoldást kapunk az emberi és gépnyelv közötti alapegység transzformációra szavakról egészen a HMM-ekig, illetve a beszédfelismerő hálózat optimalizálására is.

Szemléltettük a magyar nyelvben a tárgy- és helyragok illesztésének problémáját, mint egy nyelvfüggő kihívást, s hatékony automatikus módszert adtunk a megoldásához.

Ismertettünk két nyelvi modell típust, a CFG-t és az N-gram-ot. Megmutattuk, hogy mindkét megoldás hasonlóan jó, az N-gram kicsit több erőforrásért cserébe kicsit pontosabb eredményt ad, főleg abban az esetben, ha az eredeti várakozásoktól eltérő hosszúságú vagy szerkezetű kérdésekre kell válaszolni. Tervezett hálózatoknak mindkét megoldásnál két fő eleme van. Egyrészt az elérni kívánt pontok (POI-k) listája, melynek bővítése a felismerési pontosságot lineáris ütemben csökkenti, illetve a szituációhoz illeszkedő mondatszerkezetek, melyek hiányában a hibák száma drasztikusan nőne.

Bemutattuk és teszteltük a grafémaalapú tervezést mint a fonéma alapú rendszer alternatíváját olyan nyelvekre, ahol az írott és beszélt nyelv kapcsolata szoros. Ez kis hibanövekedés mellett rengeteg energiát spórol meg, melyet nyelvenként a külső fonetikus átíró szoftverek tesztelésére és rendszerbe illesztésére, vagy átírási szabályok gyűjtésére és tesztelésére fordítanánk. Fontos, hogy a grafémaalapú módszerrel modellezett nyelvekhez illeszkedő kivételszótárakat alkalmazzunk a modellezett nyelvhez képest idegen POI-vonatkozású szavakra.

A különböző nyelvekre készített modellek teszteléskor láthattuk, hogy az adott 6 nyelvre, az angolt kivéve, megoldásunk egyformán jól működik. Remélhetőleg a jövőben a rendelkezésünkre álló, az akusztikus modelltanításra használható adatbázis bővülésével az angol modellünk eredményei is felzárkóznak a többiéhez.

Köszönetnyilvánítás

Ez a kutatás az OM-00102/2007 számú "TELEAUTO" projekt keretén belül készült.

Bibliográfia

1. Szarvas, M.: Efficient Large Vocabulary Continuous Speech Recognition Using Weighted Finite-state Transducers – The Development of a Hungarian Dictation System. PhD Thesis, Department of Computer Science, Tokyo Institute of Technology, Tokyo (2003)
2. Mohri, M., Pereira, F. C. N., Riley, M.: Speech recognition with weighted finite-state transducers. In: Rabiner, L., Juang, F. (szerk.): Handbook on Speech Processing and Speech Communication, Part E: Speech recognition. Springer-Verlag, Heidelberg, Germany (2008)
3. Magyar Telefonos Beszéd Adatbázis: http://alpha.tmit.bme.hu/speech/hdbMTBA.php
4. Young, S., Ollason, D., Valtchev, V., Woodland, P.: The HTK book. (for HTK version 3.4) (2009) http://htk.eng.cam.ac.uk
5. Allauzen, C., Mohri, M.: Efficient algorithms for testing the twins property. Journal of Automata, Languages and Combinatorics Vol. 8 No.2 (2003) 117–144
6. Center for Spoken Language Research of Colorado: Phoenix parser for spontaneous speech. http://cslr.colorado.edu/~whw/phoenix/

Magyar nyelvű nagyszótáras beszédfelismerési feladatok adatelégtelenségi problémáinak csökkentése nyelvimodell-interpoláció alkalmazásával

Tarján Balázs[1], Mihajlik Péter[1,2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{tarjanb, mihajlik}@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.

Kivonat: A lineáris interpolációt elterjedten alkalmazzák in-domain és out-of-domain nyelvi modellek egyesítésére folyamatos, nagyszótáras gépi beszédfelismerési feladatokon. Nyelvünk gazdag morfológiája azonban szükségessé teszi, hogy morfémaalapon is megvizsgáljuk a módszer hatékonyságát, és összevessük az interpolációs és a tanítókorpuszok sima egyesítésével kapható eredményeket. Cikkünkben bemutatunk egy új megközelítést morfémaalapú nyelvi modellek interpolációjára, mellyel 3gram modellek esetén sikerült megjavítani a korpuszegyesítéses módszer eredményét. A nyelvimodell-komplexitást 4gramra növelve azonban az interpolációval nyerhető előny eltűnik, így megítélésünk szerint a morfémaalapú interpolációra vonatkozóan további vizsgálatok szükségesek. Kísérleteink során sikerült 12% alá csökkenteni a szóhibaarányt a tesztelési célokra használt hangoskönyvrészleten, mely legjobb tudomásunk szerint az eddigi legalacsonyabb eredmény magyar nyelvű, nagyszótáras feladaton.

1 Bevezetés

A nagyszótáras beszédfelismerő rendszerek pontosságát döntően befolyásolja a nyelvi modell mérete és minősége. Minél nagyobb és a felismerési feladathoz jól illeszkedő szövegkorpusz áll rendelkezésünkre a rendszer tanításához, annál precízebben írható le a szótári elemek kapcsolata az n-gram modellben. Azonban a gyakorlati tapasztalat szerint jó minőségű tanítóanyagok csak korlátozott mennyiségben hozzáférhetők, így a nyelvi modell robosztusságát gyakran a feladathoz nem vagy csak lazán kapcsolódó tanítóadat bevonásával kell növelni.

Több megoldás is létezik arra, hogy különböző szöveges tudásforrások egy közös nyelvi modellben hasznosuljanak. Szokás a rendelkezésre álló szövegeket egyszerűen összemásolni, és az így létrejött korpusszal tanítani egy n-gram modellt. Az eljárás hátránya, hogy egy nagyméretű kiegészítő korpusz könnyedén elnyomhatja a kisebb, de a feladat szempontjából releváns tanítószöveg szókapcsolati statisztikáit. Erre kínál megoldást a **nyelvimodell-interpoláció**, mellyel különböző nyelvi modellek n-gram becslései egyesíthetőek tetszőlegesen megválasztott súlyozó tényezővel. A

nyelvimodell-interpolációs technikák közül az egyik legegyszerűbb, ám igen hatékony eljárás a nyelvi modellek ún. **lineáris interpolációj**a [6]. Megvalósítása az alábbi képlet alapján történik. (**4**)

$$P(w|h) = \sum_{s \in S} \alpha_s P_s(w|h) \quad (1)$$

Ahol *w* jelöli az interpolált modell megbecsülendő szótári elemét, *h* az előtörténetet, *S* a forrásmodellek összességét, míg α_s és $P_s(w|h)$ a s-edik modellhez tartozó **interpolációs súly**t, valamint nyelvimodell-becslést. Új modell generálásakor α_s értékek változtatásával tudjuk az egyes forrásmodellek részvételi súlyát változtatni. Az interpolációban részt vevő modellek optimális arányának megállapítása általában in-domain szöveg perplexitásvizsgálatán alapul.

A lineáris interpoláció kiforrott és elterjedten használt technikának számít szóalapú nyelvi modellek esetén. Azonban a morfológiailag gazdag nyelveknél – mint amilyen a magyar – a jelentős szóalaki változatosság miatt fellépő adatelégtelenség megkérdőjelezi a szóalapú megközelítés létjogosultságát. Összehasonlító kísérletek bizonyítják, hogy magyar nyelven szóalapú helyett morfémaalapú nyelvi modelleket használva szignifikáns felismerésipontosság-növekedés érhető el [9, 11]. Felvetődik tehát a kérdés, hogy morfémákra cserélve az egyesítendő nyelvi modellek alapját, vajon a szóalapú megközelítéshez hasonló mértékben növekszik-e a felismerési pontosság, illetve ha nem, milyen módon növelhető mégis a morféma alapon interpolált nyelvi modellek teljesítőképessége.

Kísérleteink során megvizsgáljuk, milyen módszerekkel interpolálhatók hatékonyan a morfémaalapú nyelvi modellek, és összevetjük a szóalapú nyelvimodell-interpolációs eredményekkel. Emellett, hogy az interpoláció hatékonyságát általában is értékelni tudjuk, összehasonlítjuk az interpolált és az egyszerű korpuszegyesítéses modellek eredményeit is. Cikkünk további részében először a kísérletekhez használt tanító-, illetve tesztadatbázist ismertetjük, majd kitérünk az akusztikus és nyelvi modellek tanításánál alkalmazott módszerek bemutatására. A felismerési feladat részletes áttekintése után kiértékeljük a különböző interpolációs technikákat egy e célból létrehozott tesztanyagon, míg végül összefoglalását adjuk kísérleteink legfontosabb következményeinek.

2 Felismerési feladat és módszertan

A bevezetésben felvetett kérdések megválaszolásához először egy olyan felismerési feladatot kellett találnunk, mely alkalmas a különféle interpolációs módszerek vizsgálatára. Választásunk egy beszédfelismerési kísérletekhez már korábban is felhasznált [12] hangoskönyvre esett, mely Krúdy Gyula Szindbád történeteinek felvételét tartalmazza Gáspár Sándor előadásában. Fontos szempont volt, hogy olyan feladatot válasszunk, melyhez könnyen elérhető jól illeszkedő tanítószöveg, illetve hogy egy a feladattól távolabb álló, de műfajában kötődő, nagyobb méretű tanítókorpusz is gyűjthető legyen hozzá. Emellett további előnye a hangoskönyvnek, hogy a felvételeken a háttérzaj és a beszéd spontán jegyeiből adódó artikulációs pontatlanságok hatá-

sa elhanyagolható, így biztosított, hogy a felismerési pontosságok változása valóban a nyelvi modellek eltérő teljesítményéhez köthető. A rendelkezésünkre álló felvételt a [12]-ben leírtakkal megegyező módon két részre osztottuk. A nagyobbik, 186 perces részt az akusztikus modell tanításához használtuk fel, míg a kisebbik, 26 perceset a felismerő hálózatok tesztelésére.

2.1 Akusztikus modell tanítása

Akusztikusmodell-tanításhoz a hangoskönyv teszteléshez nem használt része, összesen 186 perc állt rendelkezésre. Figyelembe véve, hogy ez a több mint 3 óra egyetlen beszélőtől származik, úgy döntöttünk, hogy egy új, beszélőfüggő akusztikus modellt tanítunk. Először egy, az MRBA [13] beszédadatbázison tanított beszélőfüggetlen akusztikus modell segítségével kényszerített felismerést hajtottunk végre a tanítóanyagon, melyhez felhasználtuk az érintett Szindbád-novellák szövegét is. Ezután a kényszerített felismerés kimenete alapján háromállapotú, balról-jobbra struktúrájú, környezetfüggő rejtett Markov-modelleket tanítottunk. A létrejött akusztikus modell 1400 egyenként 7 Gauss-függvényből álló állapotot tartalmaz. A felismerési kísérletek során mindvégig ezt az akusztikus modellt használtuk.

2.2 Tanítószövegek gyűjtése és előkészítése

Mint a bevezetőben kitértünk rá, a nyelvimodell-interpolációs technikát gyakorta használják arra, hogy egy, a feladathoz jól illeszkedő kisebb és egy feladathoz csak lazán kötő nagyobb nyelvi modell előnyeit egyesítsék. Esetünkben a feladathoz jól illeszkedő modell tanításához tanítószövegként Krúdy Gyula műveinek gyűjteménye szolgált. A létrehozott korpusz 1,4 millió szót tartalmaz, forrása a Magyar Elektronikus Könyvtár [8]. Ez az általunk **jól illeszkedő (JI)** korpusznak keresztelt szöveg nem tartalmazza sem a tesztanyag, sem az akusztikusmodell-tanításhoz használt felvételek leiratát. A JI korpusz kiegészítéséhez három forrásból gyűjtöttünk, további összesen 16,6 millió szót tartalmazó tanítószöveget: Magyar Elektronikus Könyvtár, Digitális Irodalmi Akadémia [3], Elektronikus Periodika Archívum és Adatbázis [4]. Ez a tanítószöveg – melyre a továbbiakban **gyengén illeszkedő (GYI)** korpuszként fogunk hivatkozni – Krúdy Gyula kortársainak és hozzá stílusban közel álló szerzők szépirodalmi műveire épül.

Szóalapú tanítószöveg-előállítás

Egy beszédfelismerési alkalmazás a szöveges tanítóadatok előfeldolgozását követeli meg. A rendszer tanításához felhasznált szépirodalmi szövegek olyan elemeket is tartalmaznak, melyeket nem lehet, vagy eredeti alakjukban nem lehet beszédhangokkal leírni. Ennek megfelelően az írásjeleket eltávolítottunk a tanítószövegből, míg a számokat szöveges átiratukkal helyettesítettük. Végül minden karaktert kisbetűsre alakítottunk. Az így előállt előfeldolgozott tanítószöveget használtuk a szóalapú nyelvi modellek tanításához.

Morfémaalapú tanítószöveg-előállítás

A morfémaalapú tanítószövegek előállításához további lépések szükségesek. Először speciális szóhatárjelölő karaktereket (<w>) helyeztünk a szövegbe, melyeket külön morfémaként kezeltünk a nyelvi modellben. Szerepük abba rejlik, hogy segítségükkel vissza tudjuk állítani a morfémaalapú kimenetben a szóhatárokat. Ezután létre kellett hozni egy, a szavakat morfémák sorozatára átíró szótárat. Cikkünkben felhasznált morfémaalapú tanítószövegek az ún. **Morfessor Baseline (MB)** statisztikai szegmentáló eljárással [2] készültek. A MB egy felügyelet nélküli, nyelvfüggetlen morfémaszegmentáló eljárás, melyet kifejezetten beszédfelismerési célokra fejlesztettek ki finn kutatók. Segítségével csupán a szótár megadásával összerendelhetők a szavak morfémabontásukkal. A szóhatárjelölő szimbólummal ellátott, előfeldolgozott tanítószövegben ezután már csak a szavakat kellett morfémaszegmentálásukkal helyettesíteni.

Kétféle elv szerint hoztuk létre a tanítószövegekhez tartozó morfémakészleteket. Először a két tanítószöveghez tartozó szótáron egymástól függetlenül alkalmaztuk a MB szegmentálást. Ezt a megoldást **független szótáras (FSZ)** megközelítésnek neveztük el. Bár morfémaalapú hálózatok interpolációjával kapcsolatban nemzetközileg is kevés a tapasztalat, a független szótáras megoldás alkalmazása felvet egy problémát. Ha a statisztikai feldolgozó egymástól függetlenül szegmentálja az interpolálandó nyelvi modellek szótárát, akkor nagy valószínűséggel merőben eltérő morfémakészlet keletkezik. Ennek következtében a nyelvi modellek összefűzése során kevés közös n-gram lesz a két szótárban, ami ronthatja az interpoláció hatásfokát.

A probléma kezelésére több módszert kidolgoztunk, melyek közül egy ún. **közös szótáras (KSZ)** megközelítés vált be a legjobban. Ennek lényege, hogy a két tanítószöveg szótárát egyesítettük, majd ezen a közös szótáron futtattuk a statisztikai szegmentálást. A két tanítószövegben így minden közös szó ugyanarra a morfémasorozatra íródott át, ezzel biztosítva a lehető legtöbb közös n-gramot nyelvi modellekben. A kétféle módszert csak interpolációban részt vevő nyelvi modellek esetén alkalmaztuk. Korpuszegyesítés esetén a szótár a két részkorpusz közös szótárának adódik, így az itt alkalmazott szó-morféma átírás megegyezik a közös szótáras módszernél kapottal. A tanítószövegekkel kapcsolatos részletes statisztikákért lásd az **1. táblázatot**.

1. táblázat: A nyelvi modell tanító adatbázisokhoz kapcsolódó statisztikák

Tanító-korpusz	Méret [millió szó]	Szótár [ezer szó]	FSZ morféma-készlet [ezer morf.]	KSZ morféma-készlet [ezer morf.]	Szó-perplexitás [–]	OOV arány [%]
JI	1,4	152	18	36	1559	4,9
GYI	16,6	800	64	65	2905	2,6
Egyesített	18,0	840	–	66	2121	1,9

2.3 Nyelvi modellek tanítása

Mind a JI, GYI, mind az egyesített korpuszból készült nyelvi modellek módosított Kneser-Ney simítás [1] használatával készültek az SRI-LM [10] nyelvi modellező

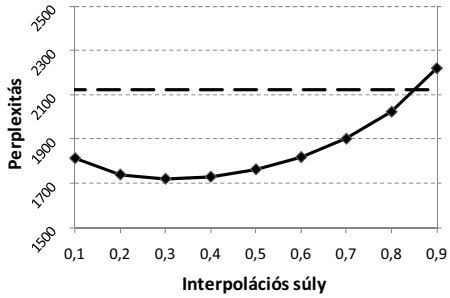
toolkit segítségével. Modellmetszést egyetlen esetben sem alkalmaztuk. Az interpolált nyelvi modellek előállításához azt az elterjedten használt technikát [7] alkalmaztuk, mely szerint egy kisebb méretű, in-domain (JI) és egy nagyobb méretű, out-of-domain (GYI) nyelvi modellt tanítottunk egymástól függetlenül, majd ezeket az SRI-LM toolkit-be épített lineáris interpolációs eljárás segítségével különböző arányban egyesítettük. A tanítókorpuszokra vonatkozó perplexitásértékek és szótáron kívüli szóarányok jól illusztrálják (**1. táblázat**), hogy bár a GYI korpusz kevésbé illeszkedik jól a tesztanyaghoz, több, a tesztanyagban előforduló szót képes modellezni, mint a JI.

3 Felismerési eredmények

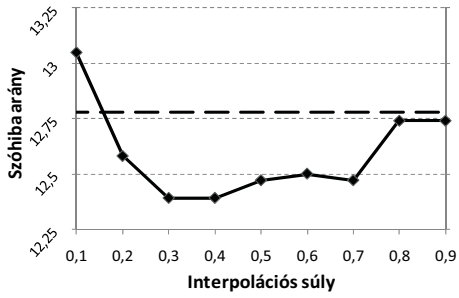
A tesztfelvétel lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre és ún. vak csatornakiegyenlítő eljárást is alkalmaztunk. A súlyozott véges állapotú átalakítókra (WFST) épülő felismerőhálózatok generálását és optimalizálását az Mtool keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztéshez a VOXerver [5] nevű WFST-dekódert használtuk. A felismerő rendszerek teljesítményének értékeléséhez szóhibaarányt (WER) számoltunk. Az egyes rendszerekkel elérhető WER értékek összehasonlításához a (**2**) képletben definiált mérőszámot használtuk.

$$Relatív\ \ WER\ csökkenés = \frac{WER_{referencia} - WER_{új}}{WER_{referencia}} * 100\% \qquad (2)$$

3.1 Szóalapú 3gram eredmények



1.1 ábra. Szóperplexitás az interpolációs súly függvényében.



1.2 ábra. Szóhibaarány az interpolációs súly függvényében.

Az **1.1 ábrán** látható, hogyan alakul a tesztanyagon vizsgálva a különböző interpolációs súllyal készült szóalapú 3gram nyelvi modellek perplexitása. A súly értéke a GYI korpuszból készült modell részarányát jelöli. Megfigyelhető, hogy a kiegészítő korpusz részarányának növelése egy pontig csökkenti a perplexitást, majd a 0,3-as

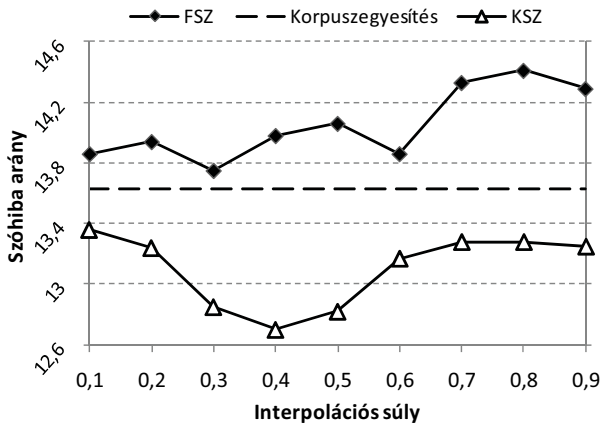
értéktől kezdve az újra növekedni kezd. Hasonló tendencia figyelhető meg a **1.2 ábrán**, mely a szóhibaarányokat ábrázolja a súly függvényében. Mindkét grafikonon szaggatott vonal jelöli a korpuszegyesítéses módszerrel elérhető perplexitást, illetve szóhibaarányt. Az a tény, hogy a folytonos vonal nagy része a szaggatott vonal alatt halad, szemléletesen mutatja, hogy szóalapú modellek esetén az interpoláció hatékonyabb, mint a korpuszok egyszerű egyesítése. Az elérhető legnagyobb pontosság esetén az interpolációval kapható relatív WER-csökkenés 3%-ot tesz ki.

3.2 Morfémaalapú 3gram eredmények

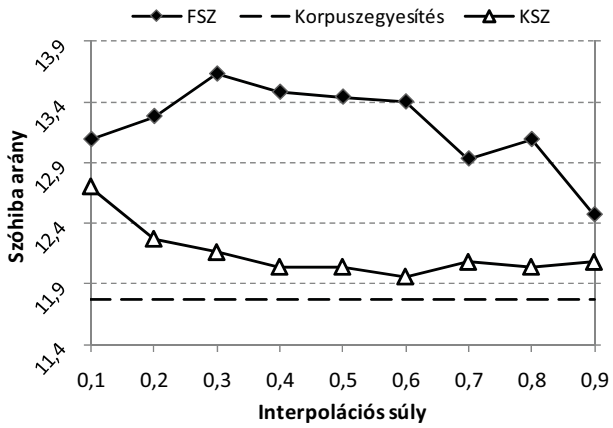
A morfémaalapú nyelvi modellek előállításához két különböző szegmentálási módszert is alkalmaztunk. Az első ún. független szótáras (**FSZ**) esetén nem készítjük fel a nyelvi modelleket az interpolációra, így azok morfémakészlete egymástól független optimalizálás eredménye (**2. ábra**). Ezt a megközelítést alkalmazva láthatóan egyetlen interpolációs súly esetén sem tudjuk javítani a korpuszegyesítéssel kapható szóhibaarányt. Ezzel szemben, ha a morfémaszegmentálás az általunk bevezetett közös szótáras (**KSZ**) módszerrel történik, akkor szóalapú eredményekhez hasonlóan csökkenteni lehet interpolációval a szóhibaarányt. A korpuszegyesítéses módszerhez képest mérhető maximális relatív WER-csökkenés (7%) felülmúlja a szó alapon kaphatót.

3.3 Morfémaalapú 4gram eredmények

Korábbi kutatásaink során többször tapasztaltuk, hogy morfémaalapú nyelvi modellezéskor 3-ról 4gram-ra növelve a nyelvi modell komplexitását szignifikánsan növekedett a felismerési pontosság [9, 11]. Ezért fontosnak láttuk morfémaalapon a 4gram modellek vizsgálatát is. Némi meglepetésre 4gram nyelvi modellek interpolációjakor nem sikerült javítani a korpuszegyesítéssel kapható felismerési eredményen. Azonban a közös szótáras (**KSZ**) megoldás itt is felülmúlja a független szótárasat (**FSZ**) felismerési pontosság tekintetében. (**3. ábra**)



2. ábra. Morfémaalapú 3gram szóhibaarányok az interpolációs súly függvényében.



3. ábra. Morfémaalapú 4gram szóhibaarányok az interpolációs súly függvényében.

4 Összefoglalás

Cikkünkben a nyelvi modellek lineáris interpolációjának alkalmazási lehetőségeit vizsgáltuk elsősorban morfémaalapú beszédfelismerő rendszerek esetén. Felismerési feladatként egy képzett beszélőtől származó hangoskönyvrészletet használtunk, melyhez egy kisebb in-domain és egy nagyobb out-of-domain tanítószöveget gyűjtöttünk. Az idealisztikus körülményeknek hála, sikerült 12% alá szorítani rendszerünk szóhibaarányát, mely legjobb tudomásunk szerint az eddig publikált legalacsonyabb érték nagyszótáras, folyamatos magyar nyelvű beszédfelismerési feladaton.

Az interpolációval és a tanítókorpuszok sima egyesítésével kapható eredményeket folyamatosan összevetettük, hogy képet kapjunk az interpolációval járó előnyökről. Hagyományos szóalapú interpolációval 3%-os WER-javulást tudtunk regisztrálni. Ez a javulás 7%-osra nőtt 3gram morfémaalapú felismerővel, ám csak abban az esetben, ha az általunk bevezetett új, a morfémaszegmentálást a tanítókorpuszok közös szótárán végző módszerrel hajtottuk végre. Ha a szótárakon függetlenül végeztük a morfémabontást, az interpoláció hatástalan eljárásnak bizonyult. Növelve a morfémaalapú nyelvi modell komplexitását 3-ról 4gramra, eltűnt az interpolációval kapható előny, és a korpuszegyesítéses módszerrel nagyobb felismerési pontosságot értünk el.

Jelen kísérletünkben nagyobb komplexitású morfémaalapú nyelvi modell esetén a lineáris interpoláció nem növelte a pontosságot a standard eljáráshoz képest. Annak eldöntéséhez, hogy ez a megfigyelés általános érvényű, vagy csupán a felismerési feladat sajátosságaiból következik, további vizsgálatok szükségesek. Éppen ezért a későbbiekben vizsgálni szeretnénk a lineáris interpolációt olyan feladatokon, melyekhez a mostaninál nagyobb tesztanyag érhető el, így kiküszöbölve a mérési hibát. Illetve ki szeretnénk próbálni a közös szótáras morfémainterpolációt olyan esetekre is, amikor a jelenleginél sokkal kevesebb adat áll rendelkezésre in-domain nyelvi modell tanításához.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani az AITIA International Zrt.-nek és a THINKTech Kutatási Központ Nonprofit Kft.-nek a rendelkezésünkre bocsátott eszközökért. Kutatásunkat részben a KMOP-1.1.3-08/A-2009-0006-os és TAMOP-4.2.2-08/1/KMR-2008-0007-es projekt támogatta.

Bibliográfia

1. Chen, S. F., Goodman, J.: An Empirical Study of Smooting Techniques for Language Modeling. Technical Report TR-10-98. Computer Science Group, Harvard University (1998)
2. Creutz, M., Lagus, K.: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. In: Comp. and Inf. Sci., report A81. HUT (2005)
3. Digitális Irodalmi Akadémia. http://www.irodalmiakademia.hu
4. Elektronikus Periodika Archívum és Adatbázis. http://epa.oszk.hu
5. Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., Tatai, G.: VOXenter - Intelligent voice enabled call center for Hungarian. In: EUROSPEECH (2003) 1905–1908
6. Jelinek, F., Mercer, R.L.: Interpolated estimation of Markov source parameters from sparse data. In: Proc.Workshop on Pattern Recognition in Practice (1980)
7. Liu, F. et al.: IBM Switchboard progress and evaluation site report. In: LVCSR Workshop, Gaithersburg, MD. National Institute of Standards and Technology (1995)
8. Magyar Elektronikus Könyvtár. http://www.mek.oszk.hu
9. 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. In: IEEE Transactions on Speech and Audio Processing, Vol. 18 No. 6, (2010) 1588–1600
10. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing. Denver (2002) 901–904
11. Tarján, B., Mihajlik, P.: On Morph Based LVCSR Improvements. In: Proc. of the 2nd Int. Workshop on Spoken Language Technologies for Under-resourced Languages (2010) 10–15
12. Tóth L.: Beszédfelismerési kísérletek hangoskönyvekkel. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2009) 206–216
13. Vicsi K., Kocsor A., Teleki Cs., Tóth L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2004) 348–359

Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal

Gosztolya Gábor[1], Tóth László[1]

[1]MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport,
Szeged
{ggabor, tothl}@inf.u-szeged.hu

Kivonat: A beszédadatbázisok kereshetővé tételéhez szöveges címkékkel kell azokat ellátni. A kézenfekvő megoldás szószintű átirat készíttetése lenne nagyszótáras beszédfelismerővel. A felismerők azonban zárt szótárral dolgoznak, így előfordulhat, hogy számunkra fontos keresési kifejezéseket (tulajdonneveket, névelemeket) esélyünk sem lesz megtalálni, pusztán mert azok nem szerepelnek a felismerő szótárában. Jelen cikkben olyan megoldásokat hasonlítunk össze, amelyek csupán beszédhang szinten végzik el az előzetes indexálást, így tetszőleges keresési kifejezésre (hangsorozatra) képesek rákeresni. A vizsgált módszerek találati pontossága gyakorlati szempontból is használhatónak ígérkezik, köszönhetően az eleve magas beszédhang-felismerési pontosságnak. A futási időt tekintve azonban még a leggyorsabb módszer is sokkal lassabbnak bizonyul, mint ami egy ilyen alkalmazástól elvárt lenne. Ezért a későbbiekben kifinomult indexálási technikák bevetésére lesz szükség.

1 Bevezetés

A beszédadatbázisok kereshetővé tételének legkézenfekvőbbnek tűnő módja egy beszédfelismerő lefuttatása a hanganyagon: ez elvileg szöveges átiratot készít a felvételekről, amelyeken ezután már a hagyományos szöveges keresési és indexálási módszereket alkalmazhatjuk. A gyakorlatban azonban az általános célú, nagyszótáras felismerők még elég nagy hibaaránnyal dolgoznak (magyar nyelvre 80% körüli szópontosság a legjobb ismert eredmény [13]). Nyilvánvaló módon a hibásan felismert szavakat a szöveges keresés során biztosan elveszítjük. Ezen a problémán lehet valamelyest segíteni oly módon, hogy nemcsak a legvalószínűbb átiratot generáltatjuk le a felismerővel, hanem ún. „N-best" kimenetet készítünk, amelyben a bizonytalan pontokon több lehetséges illeszkedő szó is szerepel (de ezzel a keresési időt és a „vakriasztás" esélyét is növeljük). A hibásan felismert szavak mellett van azonban egy másik, kevésbé nyilvánvaló probléma is a fent leírt technológiával: az, hogy a beszédfelismerők mindig egy zárt szótárral dolgoznak, így a szótárukban nem szereplő szavakat soha nem fogják megtalálni. A zárt szótár problematikája legfőképpen a főnevek, azon belül is a tulajdonnevek, illetve tágabban véve a névelemek kategóriáját érinti: ezek azok a szófajok, amelyeken belül folyamatosan keletkeznek új szavak,

A kutatást részben az NKTH TÁMOP-4.2.2/08/1/2008-0008 programja támogatta.

vagy legalábbis előtérbe kerülnek olyan szavak, amelyek a szótár összeállításakor nem forogtak közszájon, s így a szótárba sem kerültek be. Viszonylag újonnan keletkezett köznévre lehet példa a „teljesítményvolumen-korlátozás", névelemre pedig egy újonnan bejegyzett cégnév, pl. „Sokasara Kft.". A háttérből előbukkanó, majd újra eltűnő tulajdonnév esetét példázza Biszku Béla neve, amely egy hétig naponta szerepelt a híradókban, előtte viszont évekig nem, és azóta ismét nem. A beszédfelismerők szótárát, illetve nyelvi modelljét statisztikai úton, nagyméretű szöveges korpuszok alapján automatikusan szokás összeállítani. Amennyiben tehát egy szó vagy névelem nem fordult elő a tanítókorpuszban – akár mert még nem létezett, akár mert „lappangott" –, akkor az a szó nem kerül be a felismerő szótárába, és így felismerni sem fogja tudni azt. Az ilyen szavakat OOV – „out of vocabulary" – névvel illeti a szakirodalom. Egy alaposan összeállított nyelvi modell mellett ezek az OOV szavak viszonylag ritkák, így például egy diktálási feladatnál csak kevés hibát okoznak. Teljesen más azonban a helyzet, ha nem diktálásról, hanem hanganyagokban való keresésről van szó. Kereséskor ugyanis jóval gyakrabban írunk be főneveket, illetve névelemeket, mint amilyen azoknak a természetes szövegekben való előfordulási gyakorisága. A Yahoo vizsgálatai szerint a webes keresőjükbe beírt keresési kifejezések 70%-a főnév, amelynek több mint fele (40%) tulajdonnév [2]. A Microsoft hasonló elemzése szerint a keresések 71%-a tartalmaz névelemet [4], egy harmadik tanulmány szerint a webes keresések 11-17 százaléka irányul személynévre [1]. Mindez azt mutatja, hogy kereséskor pont azok a szavak fontosak, amelyeknél a legnagyobb a kockázata annak, hogy a beszédfelismerő nem ismeri őket. Természetesen elvileg megoldható a felismerő nyelvi modelljének folyamatos bővítése, ekkor azonban a teljes felismerést is újra és újra le kell futtatni, ami nehézkes és nagyon időigényes.

Az OOV szavak elkerülésére a felismerő előzetes lefuttatásával szemben elvileg lehetséges az a megoldás is, hogy a felismerőt csak a keresési kulcsszó megadása után futtatjuk le, természetesen csak az adott szóra. Nagyon nagy adatbázis esetén azonban ez nem járható út, mert még az egyetlen szóval történő teljes felismertetés is túl sokáig tart. Olyan megoldást kell tehát találnunk, amely bizonyos szintig elvégzi a felismerést, de a szószintű valószínűségek kiszámításánál hamarabb megáll. Ennek egyik módja lehet, ha nem szavakkal indexáltatjuk a beszédkorpuszt, hanem annál kisebb egységekkel, például beszédhangokkal. A felhasználó által beadott keresési kifejezést tehát a felismerő beszédhang szintű kimenetében fogjuk keresni. Sajnos ennek a módszernek is megvan a maga hátránya: a beszédhang-felismerési pontosságok a szószintű pontosságnál jóval rosszabbak, általában 50-70% közé esnek. Emiatt tehát egy hibákkal erősebben terhelt kimenetben kell keresnünk, és a környezet (szószint) sem segít, emiatt magas lesz például a vakriasztások száma rövid szavak esetén. A keresés maga is bonyolultabb, és emiatt lassabb lesz, mint szószintű címkézés esetén. Az irodalomban emiatt a szóalapú és a beszédhang alapú technikák kombinált használatát tartják a legjobbnak [6].

Jelen cikkben többféle beszédhang alapú keresési technikát hasonlítunk össze híradófelvételekben való keresési feladaton. A híradós felvételeknek csak a hírközlő által bemondott blokkjaiban keresünk, azaz alapvetően jó hangminőségű és szépen artikulált beszédről van szó, aminek köszönhetően viszonylag magas, 80% fölötti beszédhang-felismerési pontosságot sikerült elérnünk. A cikkben bemutatjuk magát a beszédkorpuszt, valamint a felismerésben használt neuronhálós technológiát. Ezután

kiindulási alapként két olyan kulcsszókereső módszert is kipróbálunk, amelyek a neuronháló adatkeret szintű kimenetén dolgoznak, tehát a lokális valószínűségek letárolásától eltekintve gyakorlatilag a teljes felismerést lefuttatják az adott kulcsszóval. Mint említettük, ez a megoldás viszonylag lassú, ezért kipróbálunk egy olyan algoritmust, amely a felismerő által kiadott N-best beszédhanghálóban dinamikus programozással, a tévesztési mátrixot figyelembe vevő szerkesztési távolság alapján keresi meg a kulcsszó feltételezett előfordulásait. A negyedik algoritmus pedig csak a felismerő által kiadott legvalószínűbb fonémasorozatot dolgozza fel, így várhatóan pontatlanabb, de gyorsabb, mint a teljes hálón dolgozó megoldás.

2 A felhasznált beszédadatbázis és feldolgozása

A kulcsszódetektálási kísérletekhez 70 híradót rögzítettünk nyolc tévécsatornáról (ATV, Hálózat TV, Hír TV, M1, M2, RTL, Tv2). A felvételeket néhány mondatos blokkokra vágtuk és három kategóriába soroltuk minőség szerint: a „tiszta" kategóriába kerültek azok a felvételek, amelyeken szépen artikulált beszéd hallható, és a háttérzaj minimális. Tipikusan ide tartoznak a stúdióban, a műsorvezetőktől elhangzó részletek. „Zajos" besorolást kaptak a tervezett beszédet tartalmazó, de magasabb zajszintű felvételek – jellemzően a külső helyszínen tartózkodó riporterek bejelentkezése. Végezetül, a „spontán" címkét kapták a spontán beszédet tartalmazó blokkok – ezek tipikusan a riportalanyok szájából elhangzó mondatok. Jelen cikkben csak a tiszta besorolású felvételeket használtuk fel, abból kiindulva, hogy többnyire minden hír előtt elhangzik egy stúdiós felvezető, így ezek kereshetővé tételével a teljes hírblokkot is meg lehet találni. A 70 híradót 44-9-17 arányban osztottuk fel betanítási (train), fejlesztési (development) és tesztelő (test) blokkokra, ügyelve arra, hogy a tévécsatornák mindegyikéből kerüljön mindegyik részhalmazba. Időtartamot tekintve kb. 5 és fél óra - 1 óra - 2 óra arányban oszlanak el a felvételek a blokkok között. A felvételek mindegyikét legépeltük, az ortografikus átiratot utólagosan is ellenőriztük. A leiratok fonetikus átiratát egy egyszerű fonetikus átíróval készítettük el, amely csak egyetlen átiratot rendel minden szóhoz, és csak egyszerű hasonulási szabályokat használ.

3 Nagy pontosságú beszédhang-felismerés neuronhálókkal

Kísérleteinkben a beszédfelismerőt szószintű nyelvi modell nélkül fogjuk futtatni, azaz pusztán a fonetikai szintű kimenete alapján szeretnénk a keresendő kifejezések előfordulásait megtalálni. Ezért érthető módon rendkívül sok múlik azon, hogy a fonetikai kimenet mennyire pontos. Angol nyelvre a TIMIT beszédadatbázison végezték a legtöbb beszédhang szintű felismerési kísérletet, és az eredmények azt mutatják, hogy neuronhálós technikákkal jobb eredményeket lehet elérni ezen a téren, mint a hagyományos rejtett Markov-modelles (HMM) megoldásokkal [11, 7]. Ezért kísérleteinkben mi is egy neuronhálót használtunk a beszédjel adatkereteinek fonetikai címke-valószínűségekkel való ellátására. A későbbiekben bemutatandó

kulcsszókereső algoritmusok egy része közvetlenül ezeket a keretszintű valószínűségi értékeket használja fel. Más részük viszont beszédhang szintű felismerési kimenetet, azaz beszédhang-sztringet vagy hálót igényel bemenetként. A felismerés lefuttatásához a neuronháló kimenete integrálható a hagyományos rejtett Markov-modellbe, így kapjuk az úgynevezett hibrid HMM/ANN rendszereket [3]. Mi a közismert HTK rendszert [18] módosítottuk úgy, hogy képes legyen a neuronháló által nyújtott lokális valószínűségekből felismerést végezni. Ily módon a hagyományos (Gauss-görbékkel dolgozó) és a hibrid modell közvetlen összehasonlítására is lehetőségünk nyílt.

A beszédjelek fonetikai címkézésére, illetve a keresendő kulcsszavak fonetikai átírására 52 címkét használtunk, ezek lényegében megfelelnek a magyar nyelv hangkészletének. A beszédtechnológiában az akusztikus modellezésben azonban igen elterjedt megoldás az úgynevezett környezetfüggő vagy trifón modellek használata. Ennek lényege, hogy a címkézést tovább finomítjuk oly módon, hogy az egyes hangok különböző hangkörnyezetben előforduló változatai különböző címkéket kapnak. Ezzel megkönnyítjük az algoritmusok számára a címkék elkülönítését, így a felismerési pontosság javul. A módszer hátránya, hogy a modellek száma megnő, így a tanítás és a felismerés is lassabb lesz, és a trifón címkék kezelése speciális problémákat is okoz. Szerencsére a HTK csomag fel van készítve a trifón modellek készítésére, így alapmodellként egy monofón és egy trifón HMM modellt is tanítottunk a korábban ismertetett adatbázison. Akusztikus jellemzőként a szokványos 13 mel-kepsztrális (MFCC) együtthatót használtuk, azok első és második deriváltjaival. Az egyes beszédhangokhoz háromállapotú modelleket rendeltünk, a HTK trifón-készítő algoritmusa ezeket 1073 környezetfüggő állapotra („szenonra”) képezte le. Nyelvi modellként csak egy szimpla beszédhangbigramot alkalmaztunk. A monofón, illetve trifón modellekkel kapott beszédhang-felismerési pontosságokat az 1. táblázatban láthatjuk.

A neuronháló betanításához adatkeret szintű fonetikai címkézésre van szükség, ezt az előzőekben betanított HMM-ek segítségével, kényszerített illesztéssel állítottuk elő. Háromféle címkézéssel is kísérleteztünk: az egyik esetben a monofón címkét rendeltük minden kerethez, azaz az 52 címke egyikét. A második esetben a monofón modell állapotának azonosítójára tanítottunk, ez esetben 3*52=153 elemből állt a címkekészlet. Végezetül, a harmadik kísérletben a trifón modell állapotazonosítóit taníttattuk a hálóval, ez esetben az osztályok száma 1073 volt. A rejtett neuronok száma minden esetben 5000-re volt állítva, és egyszerű packpropagation tanítást végeztünk. Inputként a hálózat 9 egymás melletti MFCC-vektort használt.

Viszonylag új felfedezés, hogy a neuronháló pontossága tovább javítható, ha a kimeneteire egy újabb hálót tanítunk [10, 5]. Ez a második háló a kontextus segítségével képes az első háló hibáin javítani, és részben a hangkapcsolatok modellezését is átveszi a bigram nyelvi modelltől. Ezt a technikát „kétfázisú” megoldásnak fogjuk nevezni a továbbiakban. Az 1. táblázatban soroltuk fel a különböző neuronháló-címkézési és -tanítási stratégiákkal kapott modellek beszédhang-felismerési pontosságát. Az eredmények megerősítik azt a korábbi megfigyelésünket [16], hogy a monofón tanítású hibrid körülbelül olyan teljesítményre képes, mint a hagyományos trifón HMM. Ha pedig a neuronhálót a trifón címkékhez igazítjuk, akkor további jelentős pontosságnövekedést tudunk elérni. A legjobb rendszer 83%-os pontossága alapján bíztunk abban, hogy a kulcsszavak megtalálása is lehetséges lesz pusztán a

fonetikai kimenet alapján. A következő fejezetben az e célból bevetett algoritmusokat ismertetjük.

1. táblázat: Beszédhang-felismerési pontosság a különböző akusztikus modellekkel.

Akusztikus modell	Pontosság
HMM, monofón	67,18%
HMM, trifón	75,38%
Hibrid, monofón, 1 állapot	75,56%
Hibrid, monofón, 3 állapot	76,93%
Kétfázisú hibrid, monofón, 1 állapot	77,46%
Kétfázisú hibrid, monofón, 3 állapot	79,18%
Kétfázisú hibrid, trifón	83,33%

4 Kulcsszókeresési megközelítések

A kulcsszókeresési probléma egy információ-visszakeresési (*information retrieval*, IR) feladat: adott hanganyagban keressük egy adott kulcsszóhalmaz előfordulásait. Egy kulcsszókeresési eljárás tehát találatok egy listáját adja vissza, melyek jellemzően rendelkeznek valószínűséggel is, mely alapján rangsorolhatók. Ezt a rangsort azonban, néhány más információ-visszakeresési problémával ellentétben, nem a felhasználónak adott felsorolás sorba rendezésére használjuk, hanem a találatok további szűrésére (melyet pl. a FOM kiértékelési metrika is kihasznál, l. 5.1 alfejezet).

Érdemes megjegyeznünk, hogy az angol nyelvű szakirodalomban a probléma megnevezésére két kifejezés is elterjedt: *keyword spotting*, illetve *spoken term detection* (STD). A legtöbb szerző egyetért abban, hogy vannak különbségek a két terület között, de abban már nem, hogy pontosan mik is ezek: a legkorábbi különbségtétel szerint a keyword spotting magában a hanganyagban keres, míg az STD valamilyen köztes reprezentációban. Más források szerint a fő eltérés a kulcsszavak szótárának zártságában (keyword spotting) vagy nyíltságában (STD) van; végül különbséget szokás tenni a pontosság mérésére szolgáló metrikák használata alapján is (keyword spotting: FOM, spoken term detection: ATWV; részletesebben l. 5.1 alfejezet) [17]. Az angol terminológia bizonytalansága miatt mi egységesen a kulcsszókeresés kifejezést használjuk.

A kulcsszókeresési problémában jelenleg több megközelítésnek is van létjogosultsága. Mivel a keresett kulcsszóval ellentétben a felvételek, amelyekben keresünk, előre ismertek, azok feldolgozását bizonyos mértékig előre elvégezhetjük. Az egyes megközelítések közötti alapvető különbség az, hogy a teljes feldolgozás mekkora része történik ebben az előkészítő szakaszban. A feldolgozás bizonyos lépései a keresett kulcsszó ismerete nélkül csak közelítőleg végezhetőek el; amennyiben ezeket is az előkészítő szakaszhoz soroljuk, azzal a keresési rész időigényét csökkentjük, azonban egyúttal információt is veszítünk, mely könnyen vezethet a pontosság kisebb-nagyobb mértékű csökkenéséhez. A következőkben áttekintjük a legelterjedtebb megközelítéseket.

4.1 Viterbi-keresés

A jelenlegi beszédfelismerési technikákhoz legközelebb álló megközelítésben a keresett kifejezést a beszédfelismerésben megszokott módon, a keretszintű valószínűségeket aggregálva próbáljuk ráilleszteni a bemondásokra. Találatot akkor jelzünk, ha az illeszkedés valószínűsége egy bizonyos küszöb fölé esik. Ekkor tehát az előkészítő részbe soroljuk a bemondásokon végzett szokásos műveleteket egészen a jellemzőkinyerési és fonémavalószínűség-számítási fázisokig; ezek után a keretszintű fonémavalószínűségeket tároljuk el.

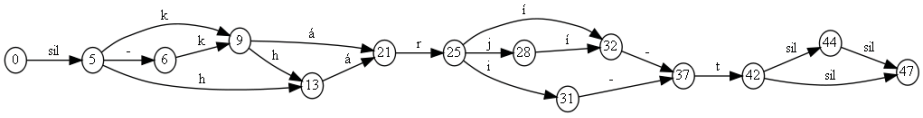
A megközelítés előnye, hogy szinte kizárólag a beszédfelismerésben szokásos technikákat alkalmazza, melyek azon a területen már hatékonynak bizonyultak. Ugyanígy, a későbbiekben a beszédfelismerés területén bevezetett pontosságnövelési technikák is könnyen átvehetők. Hátránya viszont egyrészt a bemondásokhoz eltárolandó adatmennyiség nagysága, másrészt a keresés nagy műveletigénye.

Ebben a cikkben ezt a megközelítést két konkrét implementációval valósítottuk meg; az első egy dinamikus táblatöltéses eljárás, mely az egyes keretszintű fonémavalószínűségeket kombinálja össze. A keresés végeredménye azon nem átfedő hipotézisek listája, melyek (fonémaszámmal normált) valószínűsége egy bizonyos küszöb fölé esik. Algoritmusunk futási ideje két okból is lényegesen magasabb az optimálisnál: egyrészt implementációja Matlabban történt, másrészt a konkrét megvalósításban minden kulcsszóra teljesen külön keresést végzünk azok párhuzamosítása helyett.

Alternatív megoldásként a HTK beszédfelismerő rendszerrel [18] is készítettünk egy kulcsszókereső modellt. Mivel egy beszédfelismerőnek minden jelszakaszhoz kell outputot adnia, ezért kulcsszókeresésre úgy használható fel, ha a szótárába a kulcsszavak mellé ún. „filler” vagy „garbage” elemeket veszünk fel; mi a legegyszerűbb megoldásként a beszédhangmodelleket használtuk ilyen célra. Ilyenkor a rendszer beszúrási büntetésének állításával hangolhatjuk be a találatok és a vakriasztások arányát.

4.2 Hálóalapú keresés

Az előző megközelítés két jelentős hátrányán (nagy eltárolt adatmennyiség, időigényes keresés) javíthatunk, ha további lépéseket sorolunk át az előkészítő szakaszba. Erre a legelterjedtebb megoldás a hálóalapú keresés: ennek során a keretszintű valószínűségeken fonéma N-gram nyelvtant használva hagyományos beszédfelismerést végzünk, majd a talált legjobb hipotézisek gráfját tároljuk el, a keresendő kifejezéseket pedig erre a hálóra (*lattice*) illesztjük (l. 1. ábra). E megoldás kétségtelen előnye, hogy a bemondásonként eltárolandó adatmennyiség lényegesen kevesebb, valamint – a háló méretétől függően – a keresés is gyorsabb lehet. Hátránya viszont, hogy nagymértékben támaszkodik a fonémafelismerés pontosságára: az ekkor elkövetett hibákat a későbbi lépésekben már nehezen vagy egyáltalán nem lehet korrigálni. Mivel az, hogy a keresett kifejezés összes fonémáját hibátlanul azonosítsuk, igen valószínűtlen, a keresés során bizonyos büntetésekkel beszúrást, törlést és csere műveletet is meg szokás engedni (bár ez lassítja a keresést). Emellett, a fonémafelismerés hibáit ellensúlyozandó, annak tévesztési mátrixa alapján fonémánként eltérő büntetősúly rendelhető a beszúrás és törlés műveletekhez, fonémapáronként eltérő pedig a cseréhez.



1. ábra. Példa fonémahálóra; a csúcsok időpontoknak felelnek meg, az élek címkéi az adott szakaszhoz rendelt fonémák. „sil" a csendet, „-" a zárhangok zárfázisát jelöli.

Cikkünkben ezt a megközelítést egy külső rendszer használatával képviseltettük: a Brnói Műszaki Egyetem LatticeSTD rendszerét használtuk [12]. Az előfeldolgozást és az előkészítő szakaszt a HTK [18] rendszerrel végeztük, a hálót fonéma 2-gram nyelvtant használva generáltattuk le, a háló méretét – a HTK rendszer megfelelő programjának, a HVite-nek N-best paraméterét használva – 3-ra állítottuk.

4.3 Legvalószínűbb fonémasorozaton alapuló keresés

A legvalószínűbb találatok hálóban tárolása még mindig elég bonyolult adatreprezentációt igényel, melyben a keresés is időigényes. A bemondás eltárolt modelljének további egyszerűsítésével mindkét hátulütőn javíthatunk. Kézenfekvő egyszerűsítés, ha csak a beszédfelismerési keresés során legvalószínűbbnek talált fonémasorozatot tároljuk el (természetesen a fonémák közti határok helyével és az egyes fonémaelőfordulások valószínűségével együtt) a legjobb hipotéziseket leíró háló helyett. Ennek egyértelmű előnye a reprezentáció egyszerűsége, amely a keresési algoritmus időigényét is csökkenti. Az egyszerű adatformátum lehetővé teheti olyan reprezentáció használatát is, mellyel a keresés nagymértékben felgyorsítható (indexálás). További előny, hogy az eltárolt adatok mennyisége a hálóban tároltnál egy nagyságrenddel kisebb. A megközelítés hátránya viszont, hogy a beszédfelismerési keresés során szuboptimálisnak bizonyult utak elhagyásával információt veszítünk, és még a hálóalapú keresésnél is nagyobb mértékben hagyatkozunk a fonémaosztályozóra.

Ezt a megközelítést is egy saját implementációjú kereső módszer (egy dinamikus táblatöltési eljárás) képviseli. A hálóban kereséshez hasonlóan itt is megengedünk beszúrást, cserét és törlést is, melyeket szintén a fonémafelismerés tévesztési mátrixából számolunk. A keresés végeredménye azon nem átfedő hipotézisek listája, melyek (fonémaszámmal normált) valószínűsége egy bizonyos küszöb fölé esik. Ennek a módszernek az implementálása is Matlabban történt, így futási ideje mindenképpen magasabb, mint egy gépközeli (C++, Java) megvalósításé lenne, emellett itt is az egyes kulcsszavak keresése a többitől függetlenül történik.

5 Tesztelés és eredmények

A kulcsszókeresési probléma és az alkalmazott algoritmusok leírása után most rátérünk a tesztkörnyezet ismertetésére: bevezetjük a pontosság- és sebességmérésre

használt metrikákat, vázoljuk a tesztelés menetét, végül prezentáljuk és elemezzük az elért eredményeket.

5.1 Kiértékelési metodikák

A kulcsszókeresési probléma egy információ-visszakeresési probléma, emiatt hagyományos IR-metrikákkal: pontossággal (*precision*) és fedéssel (*recall*) is mérhető egy adott algoritmus-konfiguráció teljesítménye [15, 17]. A legtöbb információ-visszakeresési területen a két metrikát azok (parametrikus) harmonikus közepével, az F-mértékkel (*F-measure*) szokás egyetlen értékké aggregálni, azonban a kulcsszókeresés területén más metrikák terjedtek el. Leggyakrabban a Figure-of-Merit (FOM) mérőszámot használják, mely az óránként és kulcsszavanként 1, 2, … 10 hibás találat megengedése esetén elért fedési értékek számtani közepe. A másik elterjedt mérőszámot az amerikai National Institute of Standards and Technology (NIST) vezette be 2006-os kulcsszókeresési versenyén: ez az aktuális kulcsszósúlyozott érték (*Actual Term-Weighted Value*, ATWV), mely a következőképpen definiált:

$$\mathrm{ATWV} = 1 - \frac{1}{T}\sum_{i=1}^{T}\left(P_{Miss}(t) + \beta P_{FA}(t)\right), \tag{1}$$

ahol $P_{Miss}(t)$ az adott kulcsszó eltévesztésének, $P_{FA}(t)$ pedig hibás találatának valószínűsége; azaz

$$P_{Miss}(t) = 1 - \frac{N_{corr}(t)}{N_{true}(t)} \quad \text{és} \quad P_{FA}(t) = 1 - \frac{N_{FA}(t)}{T_{speech} - N_{true}(t)}, \tag{2}$$

ahol $N_{corr}(t)$ az adott kulcsszó helyes találatainak, $N_{true}(t)$ a valós előfordulásainak, $N_{FA}(t)$ a hamis találatainak száma, T_{speech} pedig az átfésülendő felvételek összhossza másodpercben [8, 9]. β értéke általában 1000. Egy tökéletesen működő rendszer ATWV pontszáma 1,0, egy olyané, amely egyáltalán nem ad vissza találatokat, 0,0. Feltételezve, hogy T_{speech} lényegesen nagyobb, mint $N_{true}(t)$, egy olyan rendszer, amely az összes előfordulást megtalálja, de minden kifejezésre óránként 3,6 hamis találatot produkál, szintén 0,0 értéket fog kapni, így ez a metrika jóval szigorúbb, mint a FOM. További különbség, hogy az ATWV az összes visszaadott találatot figyelembe veszi, míg FOM esetén csak a valószínűbbeket (melyek megtartásával a hamis találatok száma még óránként és kulcsszavanként 1, 2, …, 10 alatt marad). A tesztek során elsősorban az ATWV metrikát alkalmaztuk, de az adott konfigurációhoz tartozó FOM pontszámot is feltüntettük.

Az egyes módszerek teljesítménye mellett fontos tényező azok futási ideje is. Ezt általában egyórányi hanganyagra és egy kulcsszóra vetített, másodpercben mért időigényben szokás megadni [9, 17], így mi is ezt az utat követtük.

5.2 A tesztelés menete

A tesztelést 25 darab, 2-6 szótagos kulcsszóval végeztük, melyek kellő számban fordultak elő az adatbázisban, és amelyeket valószínű keresési kifejezéseknek ítéltünk. Figyelembe véve a magyar nyelv agglutináló voltát, azt is helyes találatnak értékeltük, amennyiben a szövegben előforduló szó teljes egészében tartalmazza a keresett kulcsszót, vagy annak olyan alakját, melyben a szóvégi magánhangzó meghosszabbodott (pl. az „Amerika" kulcsszó esetén az „Amerikában" is helyes találat). A híradós felvételekből körülbelül egyórányi anyagot (a fejlesztési részt) a rendszerek fejlesztése, paramétereik finomhangolása során vettünk igénybe; az így optimalizált keresőeljárások teljesítményét pedig a mintegy kétórányi tesztfelvételhalmazon mértük le.

A módszerek hatékonyságának mérésére az ATWV metrikát alkalmaztuk. Mivel a két saját módszer esetében a találati listára akkor veszünk fel egy lehetséges találatot, ha valószínűsége egy bizonyos határ fölött van, ennek a küszöbnek a kiválasztása sem triviális; ezt úgy tettük meg, hogy minden lehetséges küszöbértékre kapott listára kiszámoltuk az ATWV metrikát a felvételek fejlesztési részén, és azt a küszöböt választottuk, amely a maximális értékhez vezetett. Ezek után a végső tesztfelvételhalmazon ezt a küszöböt alkalmazva számítottuk ki az ATWV értéket. Az érdekesség kedvéért feltüntettük az MTWV metrikát is: ezt úgy kapjuk, hogy az összes lehetséges küszöbérték használatával megszűrt találatlistára kiszámítjuk az ATWV-t, majd vesszük ezek maximumát. Mivel ezt a tesztfelvételekre tettük meg, ez lényegében azt mutatja meg, hogy optimálisra választott küszöb esetén mekkora ATWV értéket érhetnénk el.

Harmadikként kiszámítottuk a FOM százalékot is. Megjegyzendő, hogy a HTK és az általunk tesztelt, hálóban kereső módszer (LatticeSTD) elég szűk találati listát ad vissza: a hamis riasztások száma gyakran az óránként és kulcsszavanként 2-t sem éri el, míg a valós FOM érték meghatározásához szükséges az összes találat megadása felvételóránként és kulcsszavanként 10 hamis riasztásig; ebből következően ennél a két rendszernél a feltüntetett FOM százalékok csak tájékoztató jellegűek, a valós pontszám ezeknél feltehetően valamivel (2-3 százalékponttal) magasabb.

Az eljárások futási idejét az adatbázis tesztelésre fenntartott részén mértük (egy 3,0 GHz-es Intel Core2 Duo számítógépen 2GB RAM-mal), és egyórányi felvételre és egy kulcsszóra igénybe vett másodpercben fejeztük ki. Azon rendszerek esetében, melyek futási ideje függ a paraméterbeállítástól is (HTK, LatticeSTD), csak olyan paramétereket használtunk, melyekkel a lefutást még „kivárhatónak" ítéltük.

5.3 Eredmények

A 2. táblázatban láthatók az egyes módszerek által elért eredmények. Az előkészítő fázis során kipróbáltuk mind az egyszerűbb (egyállapotú) monofón, mint a bonyolultabb, de pontosabb trifón modellt; az utóbbi modelleket a Viterbi-keresés általunk tesztelt implementációja már nem tudta kezelni, így ezt az eljárást csak monofón modellel próbáltuk ki. (A fennmaradó módszerek közül a HTK képes trifónokkal is dolgozni, a hálóalapú és a legjobb fonémasorozatban kereső módszerek esetén pedig

ez a kérdés csak az előkészítő szakaszt érinti, maguk a keresőeljárások már csak az 52 önálló fonémát tartalmazó hálót, illetve sorozatot kapják meg.)

2. táblázat: Az egyes keresési megközelítések teszteredményei monofón és trifón fonémamodellt alkalmazva.

	Monofón fonémamodell			Trifón fonémamodell		
Keresési módszer	ATWV	MTWV	FOM	ATWV	MTWV	FOM
Viterbi	0,54	0,60	89,98%	–	–	–
HTK	0,52	0,58	90,42%	0,62	0,63	88,93%
Hálóalapú (LatticeSTD)	0,48	0,48	73,24%	0,65	0,65	82,64%
Legjobb fon.sorozatban	0,46	0,52	85,03%	0,43	0,58	88,37%

Az eredmények alapján a módszerek már a gyakorlatban is használható pontosságot adnak. Az is látható, hogy trifón modellt használva lényegesen javulnak az egyes módszerek teljesítményei: a növekedés jóval nagyobb, mint amit a fonémaszintű pontosság 77,46%-ról 83,33%-ra emelkedésétől várnánk, ami valószínűleg annak tudható be, hogy a tárgyalt kulcsszókeresési módszerek alapvetően a fonémaosztályozóra hagyatkoznak. Az egyes megközelítések teljesítményét összevetve jól látható, hogy ahogy egyre több részfeladatot helyezünk át az előkészítő szakaszba, és ennélfogva egyre kevesebb információ alapján végezzük a keresést, úgy csökken a keresés hatékonysága. A 3. táblázatból (mely az egyes megközelítések időigényét tartalmazza másodpercben, egy kulcsszóra és egy órányi hanganyagra vetítve) azonban az is kiviláglik, hogy mindez együtt jár a keresési idő jelentős csökkenésével is. Megjegyzendő, hogy a HTK-val és a hálóalapú keresőrendszerrel ellentétben (melyek C++-ban íródtak) a két saját eljárás Matlabban íródott, így nem igazán futási időre optimalizált. További hátrányuk, hogy a kulcsszavak keresése egymástól függetlenül történik, míg például a HTK rendszer a 25 kulcsszót párhuzamosan illesztette a bemondásokra. Valószínűleg ez a magyarázata a HTK kiugró gyorsaságának monofón fonémamodell esetén, azonban a gyakorlatban ritka az a szituáció, mikor egynél több kifejezést keresnénk párhuzamosan.

3. táblázat: Az egyes keresési megközelítések (előkészítő szakaszt nem tartalmazó) keresési időigénye másodpercben, egy kulcsszóra és egyórányi hanganyagra.

Keresési módszer	Monofón modell	Trifón modell
Viterbi	122,29	–
HTK	2,25	21,30
Hálóalapú (LatticeSTD)	34,91	34,87
Legjobb fonémasorozatban	10,75	10,75

Látványos az ATWV és MTWV értékek szignifikáns különbsége a Viterbi és a legjobb fonémasorozatban keresés módszerek esetén, míg ez a differencia a HTK rendszernél (trifón esetben legalábbis) igen kicsi, a LatticeSTD esetében pedig nulla. Ebből arra következtethetünk, hogy a két saját módszernél az egyik adatbázison (a

fejlesztési részen) megállapított küszöb nem stabil, más felvételhalmazon (esetünkben a tesztelési részen) jóval az optimális alatt teljesít, ami felveti valamilyen más küszöbszámítási módszer (pl. a keresett kulcsszó fonémaszáma helyett a találat időtartama alapján normalizálás) szükségességét.

Összességében azt mondhatjuk, hogy míg az egyes rendszerek pontossága már eléri a gyakorlati felhasználás szintjét, időigényük még meghaladja a tolerálható mértéket. Például egyhónapnyi híradófelvételben egyetlen kulcsszó megtalálása a leggyorsabb eljárásnak is majdnem három percébe kerülne, ami egy átlagos felhasználó szempontjából egyértelműen túl sok. Emiatt további programozási és indexálási trükköket szoktak bevetni, amelyek további részeredményeket leszámolva növelik ugyan a tárigényt, de gyorsítják a visszakeresést. Például a szerkesztési távolság számítása (a cserélési, törlési és beszúrási lehetőségek végigvizsgálata) gyorsítható oly módon, hogy az összes lehetséges (max k. hosszú) részstring távolságát előre kiszámítjuk és eltároljuk [14].

6 Konklúzió

Cikkünkben egy nagypontosságú fonémafelismerőre alapozva különféle kulcsszókeresési algoritmusokat hasonlítottunk össze. Várakozásainknak megfelelően azt találtuk, hogy ha a számítások egyre nagyobb részét toljuk át az előkészítő fázisba, annál gyorsabb lesz ugyan a keresés, de a pontosság is egyre romlik. Mindezzel együtt is úgy véljük, hogy az elért pontosságértékek gyakorlatilag is használhatóak lehetnek – a futási időkön viszont feltétlenül csökkenteni kell. Ezért további fejlesztésként különféle kifinomult indexálási technikák bevetése lenne a legfontosabb. Szerencsére jelenleg ez nagyon aktív kutatási terület, és az irodalom számos megoldást kínál erre a problémára. További érdekes kutatási lehetőség lenne a nagyobb egységekkel (pl. szótagok) történő indexálással való kiegészítés, valamint – mint bevezetőnkben említettük – a szószintű rendszerekkel való kombinálás, hiszen az általunk javasolt módszerek főleg az OOV szavak esetén ígérnek jelentős javulást.

Bibliográfia

1. Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 Evaluation Campaign: Overviews of the Web People Search Clustering Task. In: Proc. WWW 2009 (2009)
2. Barr, C., Jones, R., Regelson, M.: The Linguistic Structure of English Web-Search Queries. In: Proc. EMNLP (2008)
3. Bourlard, H., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer (1994)
4. Guo, J., Xu, G., Cheng, X., Li, H.: Named Entity Recognition in Query. In: Proc. SIGIR (2009)
5. Ketabdar, H., Bourlard, H.: Enhanced phone posteriors for improving speech recognition systems. IEEE Trans. ASLP, Vol. 18, No. 6 (2010) 1094–1106
6. Mamou, J., Mass, Y., Ramabhadran, B., Sznajder, B.: Combination of multiple speech transcription methods for vocabulary independent search. In: Proc. SIGIR (2008)

7. Mohamed, A.-R., Dahl, G., Hinton, G.: Deep Belief Networks for phone recognition. In: Proc. NIPS 22 workshop on deep learning for speech recognition (2009)
8. NIST: The Spoken Term Detection (STD) Evaluation Plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. http://www.nist.org/speech/tests/std (2006)
9. Pinto, J., Szöke, I., Prasanna, S.R.M., Hermansky, H.: Fast Approximate Spoken Term Detection from Sequence of Phonemes. In: Proc. SIGIR (2008) 28–33
10. Pinto, J. et al.: Analysis of MLP based hierarchical phoneme posterior probability estimator. IEEE Trans. ASLP, megjelenés alatt (2010)
11. Siniscalschi, S.M., Schwarz, P., Lee, C.-H.: High-accuracy phone recognition by combining high performance lattice generation and knowledge-based rescoring. In: Proc. ICASSP (2007) 869–872
12. Szöke, I., Schwarz, P., Matejka, P., Karafiát, M.: Comparison of Keyword Spotting Approaches for Informal Continuous Speech. In: Proc. Interspeech (2005)
13. Tarján, B., Mihajlik, P., Tüske, Z.: Nagyszótáras híranyagok felismerési pontosságának növelése morfémaalapú, folyamatos beszédfelismerővel. In: MSZNY (2009) 185–194
14. Thambiratnam, K., Sridharan, S.: Rapid Yet Accurate Speech Indexing Using Dynamic Match Lattice Spotting. IEEE Trans. ASLP, Vol. 15, No. 1 (2007) 346–357
15. Tikk, D. (szerk.): Szövegbányászat. Typotex, Budapest (2007)
16. Tóth. L., Tarján, B., Sárosi, G., Mihajlik, P.: Speech Recognition Experiments with Audiobooks. Acta Cybernetica, megjelenés alatt.
17. Wang, D.: Out-of-Vocabulary Spoken Term Detection. PhD thesis, University of Edinburgh (2010)
18. Young, S.J. et al: The HMM Toolkit (HTK) (software and manual). http://htk.eng.cam.ac.uk/ (1995)

Szótagok automatikus osztályozása spontán beszédben spektrális és prozódiai jellemzők alapján

Beke András[1], Szaszák György[2]

[1] MTA Nyelvtudományi Intézet Fonetikai Osztály
beke.andras@gmail.com
[2] BME Távközlési és Médiainformatikai Tanszék
szaszak@tmit.bme.hu

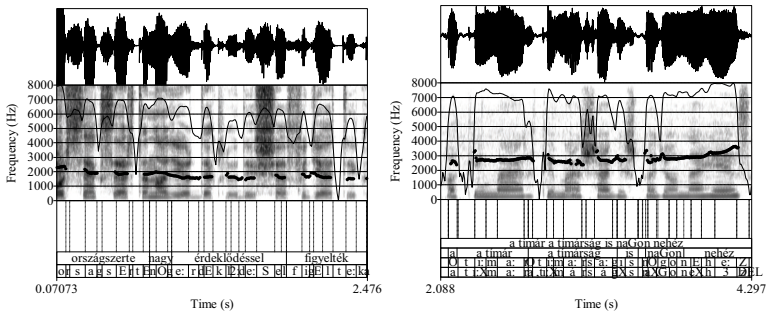
Kivonat: A beszédfolyam automatikus, szavaknak vagy néhány szóból álló szócsoportoknak megfelelő szintaktikai egységekre való tagolásában bizonyítottan fontos szerepe van a prozódiai jegyeknek, az alapfrekvenciának és az intenzitásnak. A prozódiai jegyek mellett a magánhangzó minősége is alkalmazható lehet, elsősorban a szótag eleji–nem szótag eleji szótagok osztályozására, másodsorban pedig a szóhatár meghatározására is. A jelen kutatásban azt vizsgáljuk, lehetséges-e a magánhangzó-minőség alapján a redukálódott magánhangzók automatikus elkülönítése spontán beszédben, illetve magánhangzó-minőség alapján elvégezhető-e a hangsúlyos szótagok automatikus detektálása.

1 Bevezetés

A beszédfolyam automatikus, szavaknak vagy néhány szóból álló szócsoportoknak megfelelő szintaktikai egységekre való tagolásában bizonyítottan fontos szerepe van a prozódiai jegyeknek, az alapfrekvenciának és az intenzitásnak [30]. A prozódiai jegyek mellett a magánhangzó minősége is alkalmazható lehet, elsősorban a szótag eleji–nem szótag eleji szótagok osztályozására, másodsorban pedig a szóhatár meghatározására is [7, 22, 25, 29, 33]. Ha a magánhangzó az eredeti minőségében realizálódik, akkor a hangsúlyos szótag megjelenésének esélye növekszik, míg ha a magánhangzó redukálódott formában realizálódik, akkor a hangsúlyos szótag megjelenésének esélye csökken [17, 33]. A magánhangzó redukciójáról akkor beszélünk, amikor annak képzésekor az artikulációs konfiguráció a centrális irányba tolódik el, megváltoztatva ezzel a magánhangzó minőségét. A jelenség a spontán beszéd esetében fokozottabban nyilvánul meg [1].

Az izolált szavas beszédfelismerésben a szóhatárokat egyértelműen jelzi a szünet jelenléte. A folyamatos felolvasásban a szünet mellett a szupraszegmentális akusztikai jellemzők is hozzájárulnak a szóhatárok pontos gépi meghatározásához. A spontán beszédben azonban a szavak között szinte alig jelennek meg szünetek, a beszéd folyamatos és megakadásokkal tarkított (1. ábra). A korábbi kutatások kimutatták, hogy a humán beszédpercepció szegmentálási eredménye csökken spontán beszédben (felolvasásban 90%-os [2], míg spontán beszédben 70%-os), és az sem egyértelmű, hogy a kísérletben részt vevők milyen akusztikai, szemantikai, pragmatikai jellemzők

alapján jelölték be a megnyilatkozási egységet a szövegben [12]. Az akusztikai kutatások eredményei azt mutatják, hogy a spontán beszédben az artikulációs megvalósítás túlnyomórészt ösztönös, a beszélő nincs feltétlenül tudatában annak, hogy mely szegmentális vagy szupraszegmentális tényezőt alkalmazza tagoló funkcióban, illetőleg meglehetősen nagyok az egyéni eltérések [21].



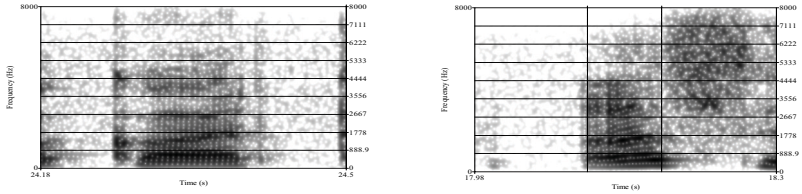
1. ábra. Egy felolvasott mondat részének és egy spontánbeszéd-megnyilatkozás részének akusztikai képe és annotációja.

A gépi felismerés számára a spontán beszédben azonban nem csak a megnyilatkozáshatárok felismerése jelenthet nehézséget, hanem már a szóhatárok bejelölése is. Az egyes idegennyelv-oktatásra irányuló kutatásokban például kimutatták, hogy a szavak szegmentálásának eredménye idegen nyelvben romlik az anyanyelvi szegmentálási eredményekhez képest. A romlás a nem ismert szavak miatt történt [31]. A spontán beszéd gépi felismerésében nagyon nehézkes a szemantikai, pragmatikai jellemzőket beépítése, illetve a szegmentális és szupraszegmentális jellemzők sem egyértelműek, emiatt a spontán beszédben megjelenő szavak határainak meghatározása ezért igen nehéz feladat.

Kutatásunkban arra is keressük a választ, hogy lehetséges-e a szótagok automatikus osztályozása spontán beszédben a magánhangzó minőségét meghatározó spektrális jellemzők alapján. Az osztályozást rejtett Markov-modellel (HMM), valamint szupport vektor gépekkel (SVM) végezzük.

2 Anyag, módszer, kísérleti személyek

A jelen kutatásban a BEA [15] korpuszból 19 magyar beszélő spontán beszédét dolgoztuk fel (8 férfi és 11 nő). Minden hangfájlnak elkészítettük a fonetikus átiratát. Az annotáció során a beszédhangokat kézzel szegmentáltuk a hangszínképük alapján a PRAAT beszédelemző szoftverben. A jelen kutatásban a következő magánhangzókat elemeztük: [ɔ], [aː], [ɛ], [eː], [i], [iː], [o], [oː], [u], [uː]. A beszédhangot akkor jelöltük az annotáció során svá magánhangzónak, (i) ha a beszédhang a centrálishoz közeli formánsstruktúrával rendelkezett (2. ábra), illetve (ii) ha a beszédhang a lehallgatás során svának hatott.



[ɔ] [ə]

2. ábra. Az [ɔ] magánhangzó és a svá [ə] magánhangzó spektrogramja (ugyanazon beszélőtől)

A szegmentálás során az annotációban azt is jelöltük, hogy a magánhangzó hangsúlyos vagy hangsúlytalan szótagon realizálódott. A magyar nyelvben mindig a szó első szótagjára esik a hangsúly [20, 32]. Azt, hogy egy szótag valóban hangsúlyos vagy sem, a szótag F0- és intenzitásértékével ellenőriztük [30].

A svát akkor jelöltük [ə]-val, ha az egységes, osztatlan beszédhangként szerepelt a hangfelismerésben. Nagy karakterrel jelöltük a svá variációkat: [A], [E], [O], ha a svát mint az eredeti magánhangzótól függő realizációt szerepeltettük a modellben. A kutatásban 4000 magánhangzót annotáltunk. A tanításhoz 2500, a teszteléséhez 1500 magánhangzót használtunk. A csoportosításra használt modellek működésének kiértékelésére és összehasonlítására meghatároztuk az osztályozás pontosságát. A pontosság (Acc) azt jellemzi, hogy az osztályozó algoritmus milyen mértékben azonosítja helyesen a beszédhangokat:

Acc=helyesen osztályozott előfordulások / összes előfordulás száma *100%

2.1 A rejtett Markov-modell

A rejtett Markov-modellezéskor az akusztikai előfeldolgozás tekintetében a beszédfelismerésben a beszédhangok akusztikai modellezéséhez használt előfeldolgozási láncot meghagyjuk, és a gyakran alkalmazott MFC (mel-frekvenciás kepsztrális) együtthatókat számítjuk 39 elemű jellemzővektorokat létrehozva. Ezek a jellemzővektorok delta és delta-delta együtthatókat is tartalmaznak. A modellkomplexitást legfeljebb 16 komponenst tartalmazó Gauss keverék (GMM) sűrűségfüggvényig növeljük. Az alkalmazott modellek topológiájuk tekintetében minden esetben 3 állapotú balról-jobbra modellek voltak.

A mintaillesztési megközelítést azonban kissé módosítjuk: felismeréskor nem szósorozatokat illesztünk, hanem beszédhangsorozatokat, az osztályozhatóság szempontjából a vizsgálatunkban érdektelen beszédhangokat azonban töltelékmodellbe vonjuk össze. Ha tehát például egyes magánhangzók és a svá irányába tolódott realizációik elkülönítése (osztályozása) a cél, akkor a magánhangzókra és a svá variánsaikra külön-külön modellek készülnek, minden egyéb beszédhangot töltelékmodellbe vonunk össze, hasonlóan a kulcsszavas beszédfelismeréshez.

A rejtett Markov-modell által időben dinamikus vetemítéssel végzett mintaillesztést kétféle módon valósítjuk meg: az első esetben hagyományos módon a teljes be-

szédmintára illesztünk. Ennek során problémaként jelentkezhet, hogy a töltelékmodellek viszonylagos univerzalitása (általános beszédhangmodell) miatt a mintaillesztés időben pontatlan, különösen hosszú beszédminták esetén. Tapasztalataink szerint ez jelentősen növelheti a törléses-beszúrásos hibák számát. Ha ilyet tapasztaltunk, akkor második mintaillesztési megközelítésként célzottan az osztályozandó magánhangzót tartalmazó részt vágtuk ki közvetlen környezetével együtt, mintegy 100 ms hosszú átmeneti részt meghagyva a magánhangzó előtt és után. Az osztályozás ezután következett, a nyelvi modell azonban kötelező jelleggel 3 modellből álló illesztési szekvenciát engedélyezett csak, amely töltelékmodellel indít és azzal is zárul. A középen elhelyezkedő magánhangzó pedig osztályozandó, az egyes variánsok egyenlő valószínűséggel szerepeltek. Ez a megközelítés kizárja a törléses hibát, és csak helyettesítési hiba fordulhat elő. Ha az osztályozást környezetből kivágottan végezzük, akkor a modellek betanítása is ugyanezen stratégiával, tehát környezetből kivágottan történik. A szerzők a magánhangzók osztályozását hangsúlyos/hangsúlytalan szempontból is vizsgálták ugyanezen megközelítésben. A rejtett Markov-modell alapú osztályozókat HTK környezetben valósítottuk meg [35].

2.2 SVM (Support Vector gépek)

Az SVM (Support Vector Machine) a felügyelt tanulási módszerek családjába tartozik, célja egy olyan szeparáló hipersík keresése, amely jól választja el egymástól a két osztály elemeit (lehet többosztályos is). Az SVM-ek működésének lényege, hogy az eredeti megfogalmazásában még komplex nemlineáris megoldást igénylő feladatot, azaz a feladatból származó mintákat, nemlineáris transzformációk segítségével egy, a bemeneti mintatér dimenziójánál több dimenziós térbe transzformálja, ahol az már lineárisan megoldható. A módszer egyik legnagyobb előnye, hogy egy garantált felső korlátot ad az approximáció általánosítási hibájára. Egy másik fontos jellemzője, hogy a tanulási algoritmus törekszik a modell méretének minimalizálására (ritka modellt alkot), ami a hiba rovására történik, de mértéke egy paraméterrel szabályozható [8, 34]. A hagyományos SVM alkalmazásának legnagyobb akadálya a módszer nagy algoritmikus komplexitása és a nagy memóriaigény, ami tipikusan a nagy adatmennyiség kezelését teszi lehetetlenné. A probléma megoldására számos megoldás született. Ezek az algoritmusok többnyire iteratív megoldások, melyek a nagy optimalizálási feladatot kisebb feladatok sorozatára bontják [3]. A nemlineáris osztályozáshoz a legelterjedtebbet, a radiális bázis (RBF – Radial Basis Function) kernelfüggvényt alkalmaztuk. A hangsúlyos/hangsúlytalan szótagok osztályozását elvégző algoritmus megvalósítása a MATLAB programban történt. Az osztályozáshoz az OSU SVM függvénykészletet használtuk [27]. Az SVM tanításához a magánhangzókból kinyert MFC-jellemzőket használtuk.

3 Eredmények

3.1 A magánhangzó és a redukálódott magánhangzó osztályozása

A semleges magánhangzók akusztikai realizációi jóval változatosabbak, mint a magánhangzókéi [5, 24]. A szegmentális és szupraszegmentális modellekben fontos szerephez juthat a svá automatikus felismerése, hiszen a folyamatos szófelismerésben a magánhangzó nem redukálódott, teljes realizációja jelezheti a szó kezdetét a beszédben [9, 25, 26]. Kopecký [22] a beszédfelismerő rendszerébe beépítette a svá fonémát, amely a rendszer felismerési pontosságának javulását eredményezte.

A magánhangzókat és a svá-realizációkat 3-állapotú HMM-mel modelleztük. A tanítás során "V" szimbólummal jelöltük a magánhangzókat, és "S" szimbólummal a svá-realizációkat. Mind a két modellt rendre 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mindkét hangmodellt ("V", "S") egyenlő súllyal rögzítettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 4 Gauss-os modell adta (1. táblázat).

1. táblázat: A magánhangzók és a svák felismerési eredményei (4 Gauss).

	Összesen	Acc [%]
"V"	706	79,46
"S"	157	71,97

A semleges magánhangzókra tanított HMM-modell nem veszi figyelembe az eredeti magánhangzót, illetve a szótag hangsúlyosságát. Az eredmények azt mutatják, hogy a spontán beszéden tanított HMM-modellel a svá-realizációk 71,97%-át osztályozta helyesen a rendszer. Az eredményből arra következtethetünk, hogy a semleges magánhangzó rendelkezik egy jól meghatározható spektrális karakterrel, amely megkülönbözteti a többi magánhangzótól. A svá akusztikai realizációi között azonban további kisebb csoportok vannak, amelyek lehetséges svá-alcsoportokra utalnak. Ilyen csoport lehet az, amelyik átfedésben lehet a magánhangzókkal is.

3.2 A magánhangzók és az egységes svá modell

Flemming [10] kimutatta, hogy a svá fonéma realizációinak lehetnek különböző alcsoportjai: közép-centrális svá és kontextusfüggő svá. A svá-realizációk variációinak egy része a kontextus hatására megváltozik, és egy sajátos kontextusfüggő hangminőséget hoz létre, amely a svának egy akusztikai alcsoportja lehet [29]. A nemzetközi és a hazai szakirodalom sem egységes abban, hogy milyen tényleges okai vannak a svá variáltságának, illetve melyek a lehetséges svá-alcsoportok. Flemming szerint nyilvánvaló, hogy a semleges magánhangzónak két típusa létezik, azonban a levonható következtetések nem egyértelműek. A magánhangzó redukciója jelezheti a hangsúlytalan és hangsúlyos szótag közötti szembenállást is, ami az angol nyelvben szabályszerűnek tekinthető. A közép-centrális svá a hangsúlytalan alacsony nyelvállású magánhangzóból jön létre kismértékű redukció során, éppen ezért nem minden magán-

hangzó-minőségből keletkezhet. A svá nem közép-centrális variánsai a magas nyelvállású magánhangzókból jönnek létre a redukció során.

Annak érdekében, hogy a svá-realizációk egységességét megvizsgáljuk, illetve hogy meghatározzuk, melyik magánhangzó minőséghez esik a legközelebb, négy HMM-modellt építettünk. Három modellt készítettünk a három leggyakrabban előforduló magánhangzóra [ɔ, ɛ, o] és egy egységes modellt a semleges magánhangzóra "S". A három magánhangzó-minőséget és a svá-realizációkat 3-állapotú HMM-mel modelleztük. A tanítás során [ɔ], [ɛ], [o] szimbólummal jelöltük a magánhangzókat, és "S" szimbólummal a svá-realizációkat. Mind a négy modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A legjobb felismerési eredményt a 4 Gauss-os modell adta (2.a. táblázat).

2.a. táblázat: Az [ɔ], [ɛ] és [o] magánhangzók, és az egységes svá "S".

	Összesen	Acc [%]
S	140	65
[ɔ]	167	70,65
[ɛ]	225	75,11
[o]	115	73,04

Az eredmények azt mutatják, hogy a svá magánhangzók helyes osztályozásának eredménye 7%-kal romlott ezzel az eljárással. A 2.b. táblázatban a négy modell tévesztési mátrixa mutatja, hogy a svá magánhangzó az [o] modellhez van a legközelebb, mivel az [o] hangok 18%-át téveszti össze a rendszer a svá magánhangzóval.

2.b. táblázat: Az [ɔ], [ɛ] és [o] magánhangzók, és az egységes svá "S" tévesztési mátrixa.

Magánhangzók	[ɔ]	[ə]	[ɛ]	[o]
[ɔ]	118	9	10	16
[ə]	15	91	8	18
[ɛ]	16	13	169	19
[o]	12	7	4	84

A nagyobb tévesztési arány oka az lehet, hogy az [o] vokális artikulációs konfigurációja közel esik a semleges magánhangzóéhoz, illetve az [o] időtartama alacsonyabb, mint az [ɔ] és [ɛ] időtartama [13, 14]. Ennek igazolására kimértük a spontán beszédben előforduló [ɔ], [ɛ], [o] és a redukálódott magánhangzók időtartamát. A svá időtartama szignifikánsan rövidebb, mint a magánhangzóké. A svá magánhangzó időtartama átlagosan 53 ms, míg a magánhangzóké 84 ms (ANOVA: $F(1, 2917) = 252{,}757$;. $p = 0{,}000$**). Ez a tendencia megegyezik a nemzetközi és hazai szakirodalomban leírtakkal [4, 10, 14, 33].

Az adatok szerint a három magánhangzó időtartama közül az [o] magánhangzóé áll a legközelebb a svá időtartamához. Az [o] magánhangzó időtartama (77 ms) szig-

nifikánsan rövidebb, mint az [ɔ] (83 ms) vagy az [ɛ] (90 ms) időtartama (ANOVA: $F(2, 2313) = 19{,}86$ $p = 0{,}000$**; csoportok közötti különbség (post hoc test) $p > 0{,}000$**).

A magánhangzók realizációi átfedésben vannak egymással és a redukálódott magánhangzókkal is az első két formánsértéket tekintve. Bondarko et al. [4] kimutatta, hogy a magánhangzó átmeneti része minden magánhangzó esetében meghatározható mind az olvasott, mind a spontán beszédben, azonban a magánhangzó tisztafázisa a spontán beszédben sokszor eltűnik a magánhangzók redukálódása miatt. A jelen kutatás adatai szerint az [o] magánhangzó időtartama jelentősen rövidebb, mint az [ɔ] és [ɛ] magánhangzóé, ami utal a magánhangzó ejtésekor bekövetkezett célalulmúlásra, ez pedig a magánhangzó tisztafázisának redukciójához vezethet: így az [o] magánhangzó redukálódása olykor erősebb lehet. Eredményeinket alátámasztja Padget [28] kutatása is, amelyben 9 beszélő beszédében a magánhangzók redukálódását vizsgálta. Azt találta, hogy az [ɔ] és az [o] magánhangzót nehezebben lehet elkülöníteni a többi magánhangzótól mind felolvasásban, mind spontán beszédben.

3.3 A magánhangzók és a magánhangzófüggő svá

A helyettesítő funkcióban realizálódott svá akusztikai képe feltételezésünk szerint függ az eredetileg kiejteni kívánt magánhangzó artikulációs konfigurációjától is, amely helyett megjelenik a beszéd során. Ha a svá realizációi függenek a helyettesített magánhangzó minőségétől, akkor a svá-realizációk modellezhetőek a helyettesített magánhangzó minősége mentén. A svá-realizációnak a következő alcsoportjai léteznek: az [ɔ] magánhangzót helyettesítő svá [A], az [ɛ] magánhangzót helyettesítő svá [E], az [o] magánhangzót helyettesítő svá [O]. A tanítás során [ɛ], [o], [ɔ]-val jelöltük az eredeti minőségben realizálódott magánhangzókat, míg [A], [E], [O]-val a helyettesített magánhangzó minőségétől függő svá-realizációkat. Mind a hat modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mind a hat hangmodellt egyenlő súllyal szerepeltettük (azaz osztályozáskor egyformán valószínűek voltak). A legjobb felismerési eredményt ismét a 4 Gauss-os modell adta (3. táblázat).

3. táblázat: Az [ɛ], [o], [ɔ] és az [A], [E], [O] osztályozásának eredményei.

	Összesen	Acc [%]
[ɔ]	169	65,08
[A]	47	68,08
[ɛ]	227	69,60
[E]	65	63,07
[o]	116	61,20
[O]	29	62,06

Az eredmények azt mutatják, hogy az osztályozó az [ɔ] magánhangzó helyett realizálódó [A] svát osztályozta a legjobb arányban. Az [o] magánhangzót és az [o] magánhangzó helyett realizálódott svát az algoritmus nem tudta elválasztani olyan pontosan a többi modelltől. Az osztályozás legnagyobb nehézsége ebben az esetben az [o] magánhangzó és a helyette realizálódott svá minőségének variabilitása és az időtartamának csökkenése. A véletlen találgatásnál sokszorosan jobb eredmények mindenesetre alátámasztják, hogy a svá realizációja helyettesítő funkcióban függ a helyettesített magánhangzó eredeti célkonfigurációjától.

3.4 Veláris − palatális magánhangzó és veláris − palatális svá

A magánhangzófüggő svá jobb osztályozhatóságának vizsgálata érdekében megpróbáltunk modelleket összevonni. Korábban megjegyeztük, hogy számos nemzetközi tanulmány foglalkozik a svá-realizációk csoportosítási lehetőségével. A tanulmányok többsége a magánhangzó F2 dimenziójának és időtartamának módosulását tartja a magánhangzó-redukálódás akusztikai paraméterének, ezért a modelleket az F2 dimenzióban vontuk össze. Ha a svá veláris magánhangzó helyett realizálódik, akkor a redukálódott magánhangzóra a veláris magánhangzó-minőség lesz jellemző a svá-realizációkon belül. Ha a svá palatális magánhangzó helyett realizálódik, akkor a redukálódott magánhangzóra a palatális magánhangzó-minőség lesz jellemző a svá-realizációkon belül. Elsősorban a svá-realizációk palatális és veláris alcsoportjainak elkülöníthetőségét teszteltük. Jason [18] a svá lehetséges palatális − veláris alcsoportját HMM modellel tanította és tesztelte fonetikailag variábilis, angol nyelvű, egy beszélőtől származó korpuszon. Eredményei alátámasztották, hogy a svá-realizációknak létezik egy veláris és egy palatális alcsoportja. Azt is kimutatta, hogy a svá magánhangzók kezdeti fázisukban különíthetőek el egymástól, míg a végső fázisukban nem.

A redukálódott magánhangzók realizációiban ugyanúgy létezik veláris − palatális különbség, ahogy a magánhangzók realizációiban. A palatális svá realizációk az F1/F2 térben közelebb vannak a palatális magánhangzókhoz: magasabb F2-értékkel realizálódnak.

A négy magánhangzó-minőséget 3-állapotú HMM-mel modelleztük. A tanítás során VV-vel jelöltük a veláris magánhangzókat, PV-vel a palatális magánhangzókat, VS-sel a veláris svákat és PS-sel a palatális svákat. Mind a négy modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A legjobb felismerési eredményt a 4 Gauss-os modell adta (4.a táblázat).

4.a. táblázat: Az osztályozás eredményei a VV, PV, VS és PS modellekre.

	Összesen [db]	Acc [%]
Veláris mgh.	318	56,91
Palatális mgh.	375	53,86
Veláris svá	76	63,15
Palatális svá	66	40,90

Az eredmények azt mutatják, hogy a svá-realizációk felbonthatóak veláris és palatális svá-realizációkra, amely alátámasztja a formánsok alapján leírt megállapításokat. Az osztályozásban a veláris svá modell adta a legjobb eredményt (63,15%). A palatális svák viszonylag alacsony osztályozási képessége azzal magyarázható, hogy a palatális magánhangzó realizációinak artikulációs tere jóval nagyobb, mint a veláris magánhangzóké, ezért jóval magasabb a realizációk variációinak a száma is (azaz a modell nagyobb szórást enged meg, és emiatt relatíve pontatlanabb modellezést tesz csak lehetővé). Az eredményeinket alátámasztják Bunnel [6] eredményei is: Bunnel megállapította, hogy a palatális svák felismerési eredménye jobb, mint a veláris sváké. Az osztályozási feladatot a veláris és a palatális svá elkülönítésére egyszerűsítve jól látható, hogy a veláris svá felismerése sokkal biztosabb (4.b táblázat).

4.b. táblázat: Az osztályozás eredménye a VS és PS modellekre környezetből kiragadott modellezési technikával.

	Összesen [db]	Acc [%]
Veláris svá	89	79,77
Palatális svá	69	66,66

Ez visszavezethető arra, hogy a veláris svá sokkal egységesebb kategóriát képez, ami megegyezik a nemzetközi szakirodalomban leírtakkal [11]. Harmegnies–Poch-Olivé [16] kimutatták, hogy a redukálódás markánsabban jelenik meg a palatális magánhangzók esetében, mint a veláris magánhangzók esetében.

A nemzetközi eredmények és a jelen kutatás eredményei azt mutatják, hogy a svának helyettesítő funkcióban két alcsoportja különíthető el: a palatális és a veláris svá.

3.5 A modellek kiértékelése

A jelen tanulmányban használt HMM modellek közül az egységes svát és az egységes magánhangzót modellező 3-állapotú HMM-ek pontossága volt a legjobb (78%), 4 Gauss kibocsátási valószínűséget leíró függvénnyel, ami azt jelenti, hogy ezekkel a modelleken osztályozta helyesen a legtöbb hangot az algoritmus (5. táblázat). A legkevesebb helyes találatot az eredeti minőségű magánhangzót és a helyettesített magánhangzó-minőségtől függő svát modellező 3-állapotú HMM-ek adták (69,46%). A veláris és palatális svákat modellező 3-állapotú HMM-ek pontossága 74%.

5. táblázat: A tanított modellek pontossága.

A tanított modellek	Acc [%]
Eredeti magánhangzó-minőség és a helyettesített magánhangzó-minőségtől függő svá	69,46
Veláris palatális magánhangzó és veláris palatális svá (monofon)	70,35
Veláris palatális magánhangzó és veláris palatális svá (környezetből kiragadva)	74,05

Egységes svá és magánhangzók [ɛ],[o], [ɔ]	75,86
Egységes magánhangzó és egységes svá	78,09

3.6 Hangsúlyos − hangsúlytalan szótagok osztályozása a magánhangzó minőségének segítségével

A hangsúlyos és hangsúlytalan szótagok osztályozásához gondosan felszegmentált anyagon, 3-állapotú HMM-eket tanítottunk. A hangsúlyos szótagokat „XA”-val, a hangsúlytalan szótagokat „XT”-vel jelöltük. Mind a két modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mind a két hangmodellt egyenlő súllyal szerepeltettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 8 Gauss-os modell adta (6.a. táblázat).

6.a. táblázat: A XA és a XT osztályozásának eredményei.

Szótagok	Összesen [db]	Acc [%]
XA	309	82,80
XT	855	70,72

Az eredmények szerint a hangsúlytalan szótagok osztályozása kevésbé pontos, ami arra utal, hogy ez az osztály nem egységes a modellezett szótagok hangminősége szempontjából.

Az SVM-mel tanított és tesztelt magánhangzó-minőségen alapuló osztályozó pontossága 54%. A szókezdő pozícióban lévő magánhangzókat mindössze 56%-ban, míg a nem szókezdő pozícióban lévő magánhangzókat 58%-ban osztályozta helyesen az algoritmus (6.b. táblázat).

6.b. táblázat: A szókezdő és nem szókezdő pozícióban lévő magánhangzók osztályozási eredménye (Acc) SVM-mel.

	Szókezdő	Nem szókezdő
Szókezdő	61%	39%
Nem szókezdő	42%	58%

Az SVM-mel végzett osztályozásban is a szókezdő pozícióban lévő magánhangzók eredménye jobb.

A hangsúlytalan szótag modellezésében ronthatja az osztályozási eredményeket, hogy a hangsúlytalan szótagban a magánhangzó minősége nem egységes. A hangsúlytalan szótagban a magánhangzó megjelenhet redukálódott magánhangzóként, illetve az eredeti magánhangzó artikulációs konfigurációnak megfelelő minőségben is. Ennek igazolására három HMM-et építettünk. A kísérlet során modelleztük a hangsúlyos szótagban realizálódott magánhangzót (XA), a hangsúlytalan szótagban realizálódott eredeti magánhangzó-minőséghez közeli magánhangzót (XT) és a hangsúlytalan szótagon megjelenő redukálódott magánhangzót (XS). Ezeket a magánhangzó-minőségeket 3-állapotú HMM-ekkel modelleztük. Mind a három modellt ismét rendre 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk.

Az osztályozás nyelvtanában mind a három hangmodellt egyenlő súllyal rögzítettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 8 Gauss-os modell adta (6.c. táblázat).

6.c. táblázat. A három magánhangzó-minőség (XA,XT,XS) osztályozási eredménye.

Magánhangzó-minőségek	Összesen [db]	Acc [%]
XA	299	80,70
XT	646	68,80
XS	218	73,20

A modellek átlagos osztályozási pontossága lényegesen nem változott az előbbi esethez képest, a hangsúlytalan szótagok helyes osztályozása (közösen az XT- és az XS-modell) ismét közel 71%-os. A hangsúlytalan modell kettéválasztása az eredeti hangminőséghez közeli magánhangzóra és redukálódott magánhangzóra viszont azt mutatja, hogy ha a magánhangzó redukálódik is, akkor valamelyest kisebb eséllyel osztályozza az osztályozó hangsúlyosnak. Meg kell jegyeznünk, hogy a hibák nagyobb része törlésből és nem tévesztésből származott, ilyenkor az osztályozó egyszerűen átugrik egy szótagot (vagy ha úgy tetszik, összevonja azt az előtte-mögötte állóval).

4 Következtetések

A jelen tanulmány célja az volt, hogy a helyettesítő funkcióban lévő svá-realizációkat spektrális jellemzőik alapján modellezze HMM-ekkel magyar nyelvű spontán beszédben.

Az elemzések során bemutattuk, hogy (i) a [ə] és a svá-realizációk MFC-együtthatók alapján előfeldolgozva HMM-ekkel modellezhetőek a magyar nyelvű spontán beszédben; (ii) a svá-variációk realizációi függnek az általuk helyettesített magánhangzó artikulációs konfigurációjától. Ezen megállapítások igazolására hat különböző modellt építettünk, amelyek reprezentálták a „svá" és a lehetséges svá-alcsoportok realizációit. Jóllehet az osztályozás során globális pontosság szempontjából a legjobb eredményt az osztatlan svá, az eredeti minőségben realizálódott magánhangzó modellhalmaz adta, a svá-realizációk közötti leghatékonyabb osztályozást a veláris és palatális svá-alcsoportra épített HMM-ek adták (79%).

A vizsgálat során összehasonlítottuk az eredeti minőségben realizálódott magánhangzók és a redukálódott magánhangzók akusztikai szerkezetét (időtartam és formánsszerkezet). Az eredmények azt mutatták, hogy az [o] magánhangzó artikulációs konfigurációjában közelebb áll a svá artikulációs konfigurációjához a spontán beszédben, mint a vizsgálatban szereplő többi magánhangzó. Ennek oka az, hogy az [o] artikulációs konfigurációja és időtartama jóval nagyobb variációt mutat, mint a vizsgálatban szereplő többi magánhangzóé.

A svá realizációk lehetséges alcsoportjait HMM-ekkel modelleztük. A hipotézisünk és a nemzetközi szakirodalom szerint a svá-realizációk alapvetően két csoportra bonthatók, méghozzá palatális és veláris svá-variációkra. Az eredmények azt mutatták, hogy a svá-realizációknak ez a két alcsoportja létezik: veláris és palatális svá. Ennek oka az, hogy a helyettesítő funkcióban lévő svá függ az általa helyettesített magánhangzó minőségétől: a veláris magánhangzó redukálódása közben megőrzi az alapvető veláris spektrális jegyeket, ahogy a palatális svá is megőrzi a palatális magánhangzó spektrális jellemzőit. Ez természetesen nem zárja ki azt, hogy a svá-realizációk esetleg más dimenziókban is elválaszthatók egymástól.

A hangsúlyos és hangsúlytalan szótagokban realizálódott különböző magánhangzó-minőségeket HMM-ekkel modelleztük MFC-előfeldolgozás alapján. A hangsúlyos és a hangsúlytalan szótagok osztályozásának eredménye 73,76% volt. A hangsúlyos szótagok felismerése ezen belül 82,8%-kal a leghatékonyabb volt. A svá-modell beépítése összességében javított a hangsúlyos és hangsúlytalan szótagok csoportosításában. A magánhangzó-minőséggel modellezett hangsúlyos és hangsúlytalan szótagok felismerésének eredménye jobbnak bizonyult a hasonló nemzetközi kutatások eredményeihez képest (vö.[19, 23]).

Bibliográfia

1. Beke A.: A veláris magánhangzók stabilitása a spontán beszédben. In: Gecső Tamás − Sárdi Csilla (szerk.) A kommunikáció nyelvészeti aspektusai. Kodolányi János Főiskola−Tinta Kiadó, Székesfehérvár−Budapest (2009) 27–31
2. Batliner, A, Kompe, R., Kießling, A., Mast, M., Niemann, H., Nöth, E., Oth, E.N.: M=Syntax+Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases. Speech Communication Vol. 25. (1998) 193–222
3. Bennett, K. P., Campbell, C.: Support Vector Machines: Hype or Hallelujah?. SIGKDD Explorations Vol. 2 No. 2 (2000) 1–13
4. Bondarko, Liya V., Volskaya, Nina B., Tananaiko, Svetlana O., Vasilieva, Ludmila A.: Phonetic properties of Russian spontaneous speech. In: Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona, 3-9 August (2003) 2973−6
5. Browman, C. P., Goldstein L.: Articulatory phonology: An overview. In: Phonetica Vol. 49 (1992) 155−80
6. Bunnel, H. T., Lilley J.: Schwa variants in American English. In: Proceeding of the InterSpeech 2008. Brisbane, Australia (2008) 1159−62
7. Cruttenden, A.: Intonation (2nd ed.). Cambridge University Press, New York (1997)
8. Girosi F.,: An equivalence between sparse approximation and support vector machines. Neural Computation Vol. 10 No. 6 (1998) 1455–1480
9. Dressler, W. U.: Explaining Natural Phonology. In: Phonological Yearbook 1 (1984) 29−50
10. Flemming, E.: The phonetics of schwa vowels. Manuscript, MIT (2007)
11. Flemming, E., Johnson S.: Rosa's roses: reduced vowels in American English. In: Journal of the International Phonetic Association Vol. 37 (2007) 83−96
12. Gósy M.: Virtuális mondatok a spontán beszédben. In: Gósy M. (szerk.): Beszédkutatás 2003. MTA Nyelvtudományi Intézet, Budapest (2003) 19−44
13. Gósy M.: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest (2004)
14. Gósy M.: The manifold function of schwa. Grazer Linguistische Studien 62 (2004) 15−26

15. Gósy M.: Magyar spontánbeszéd-adatbázis – BEA. In: Gósy M. (szerk.): Beszédkutatás 2008. MTA Nyelvtudományi Intézet, Budapest (2008) 194–207
16. Harmegnies, B., Poch-Olivé D.: A study of style-induced vowel variability: laboratory versus spontaneous speech in Spanish. In: Speech Communication Vol. 11 (1992) 429−37
17. Heuvel, H., Kuijk D., Boves L.: Modeling lexical stress in continuous speech recognition for Dutch. In: Speech Communication Vol. 40 (2003) 335–50
18. Jason, L.: Data-driven investigation of subphonemic variation: "Front" schwa vs. "back" schwa. Paper read at the Cognitive Science Graduate Student Conference 2008. Delawer, Friday April 18th (2008)
19. Jenkin, K. L., Scordilis M. S.: Development and comparison of three syllable stress classifiers. In: International Symposium on Chinese Spoken Language, ICSLP (1996) 733−6
20. Kálmán, L., Nádasdy Á.: A hangsúly. In: Ferenc Kiefer (szerk.): Strukturális magyar nyelvtan 2: Fonológia. Akadémiai Kiadó, Budapest (1994) 393−467.
21. Kopecký, J., Glembek O., Karafiat M.: Advances in acoustic modeling for the recognition of Czech. Paper read at the International Conference on Text, Speech and Dialogue; TSD (2008)
22. Kohle, K. J.: Prosodic boundary signals in. German. Phonetica Vol. 40 (1983) 89–134
23. Kuijk, D., Boves L.: Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication Vol.27 No.2 (1999) 95−111
24. Ladefoged P.: A course in phonetics. Third edition. Harcourt Brace Jovanovich, New York. (1993)
25. Ladefoged P., Maddieson I.: Vowels of the world's languages. In: Journal of Phonetics Vol. 18 (1990) 93−122
26. Madelska, L., Dressler W. U.: Postlexical stress processes and their segmental consequences illustrated in Polish and Czech. In: Hurch, B., Rhodes, R. A. (szerk.): Natural Phonology: The state of the art. Mouton de Gruyter, Berlin & New York. (1996) 189−200
27. OSU SVMs Toolbox for MATLAB (http://www.ece.osu.edu/~maj/osu_svm/)
28. Padgett, J., Tabain M.: Adaptive Dispersion Theory and phonological vowel reduction in Russian. In: Phonetica Vol. 62 (2005) 14−54
29. Pennington, M. C.: Phonology in English language teaching: An international approach. Longman, London (1996)
30. Szaszák Gy.: A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben. Ph.D. értekezés, Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék (2009)
31. Simon O.: Anyanyelvi és idegen nyelvi percepciós működések összefüggései az általános iskolában. In: Navracsics J., Tóth Sz. (szerk.): Nyelvészet és interdiszciplinaritás. Generalia, Veszprém (2004) 438−449
32. Siptár, P., Törkenczy M.: The phonology of Hungarian. Oxford University Press, Oxford (2000)
33. Swerts, M., Kloots H., Gillis S., Schutter, G.: Vowel reduction in spontaneous spoken Dutch. In: Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition. Tokyo, Japan (2007) 31–34
34. Valyon J., Horváth G.: Least squares szupport vektor gépek adatbányászati alkalmazása. Híradástechnika Vol. 60 No.10 (2005) 33–38
35. Young, S., Evermann, G., Gales, M.,Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.3). Cambridge University Engineering Department, Cambridge (2005)

Spontán beszédben rejlő nem verbális hangjelenségek – érzelmek, hanggesztusok – vizsgálata

Vicsi Klára, Sztahó Dávid, Kiss Gábor, Czira Anita

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium,
1111 Budapest, Sztoczek u. 2.
{vicsi, sztaho}@tmit.bme.hu

Kivonat: Ebben a cikkben azokat a vizsgálatokat tárgyaljuk, amelyek a spontán beszéd nem verbális hangjelenségeinek a kutatására vonatkoznak. Elsősorban a nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalom jellegzetességeit, feldolgozási nehézségeit tárgyaljuk, amelyek prozódiai jellemzőkkel jutnak kifejezésre a beszédben a nyelvi tartalommal összefonódva. Továbbá csoportosítva tárgyaljuk azokat a nyelvi tartalomtól elhatárolt, attól független hangi jelenségeket, amelyek a spontán beszédben előfordulnak, és bemutatjuk az általunk létrehozott hanggesztustárat.

1 Bevezetés

Az emberi beszédkommunikációban a beszédinformáció feldolgozása két egymástól elkülönült módon történik. Az egyik feldolgozási mód esetében az üzenet nyelvi tartalmát dolgozzuk fel (verbális csatorna); a másik információfeldolgozási mód (a nem verbális csatorna) ahol a beszélő általános érzelmi, egészségi állapotát, hangulatát érzékeljük [1]. Az utóbbi évtizedben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna jelentősége ez idáig kisebb volt, és működését kevésbé értjük.

1. ábra. Az emberi kommunikáció két egymástól elkülönült feldolgozási csatornája.

Az emberi beszéddel a beszédtartalmon túl sok mást is ki lehet fejezni. Ezeket a beszélő különböző beszédformákkal (változatok) tudja érzékeltetni. A hangszínezet, az intonáció, a ritmusváltozások mind széles körben használatosak arra, hogy a beszélő érzelmi, hangulati vagy egészségi állapotát is egyidejűleg kifejezzék.

Korábban a beszédtartalom vizsgálatakor rendszerint olvasott, vagy szépen kiejtett beszéd volt a vizsgálat alapja, viszont a beszédtechnológiai alkalmazásokban a valóságos spontán beszéd feldolgozása szükséges!

Spontán társalgásban számos nem nyelvi elem fordul elő, amelyek hozzájárulnak ahhoz, hogy a beszélgetőpartnerek jobban megértsék egymást. A beszédkommunikációban a lelki állapot, az érzelem, az egyetértés vagy egyet nem értés közvetítése azt a célt szolgálja, hogy a beszélgetőpartnert informáljuk, még ha ezeket az információkat szavakkal nem is fejezünk ki a társalgás során. A spontán társalgás jelfeldolgozás szempontjából történő megismeréséhez elengedhetetlenül szükséges ezeknek a nem verbális jelenségeknek a kutatása.

A BME TMIT Beszédakusztikai Laboratóriumban éppen ezért, ezeket a beszédben rejlő nem verbális információkat hordozó hangjelenségeket vizsgáljuk. Ezek a nem verbális hangjelenségek a következők:

1. Nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalom, amely prozódiai jellemzőkkel jut kifejezésre a beszédben a nyelvi tartalommal összefonódva. Ilyenek például a szomorúság, izgatottság, idegesség, vidámság stb. vagy akár az egyetértés és az egyet nem értés prozódiai jellemzőkkel való kifejezése.
2. A nyelvi tartalomtól elhatárolt, attól független hangi események, amelyek további csoportokra bonthatók:
   - 2.1. jelentést kifejező hangjelenségek – ezek a hanggesztusok. Ilyenek például a sírás, a nevetés, a különböző érzelmet kifejező felkiáltások.
   - 2.2. jelentéssel nem rendelkező hangjelenségek:
     - 2.2.1. Kitöltött szünetek
     - 2.2.2. Egyéb hangjelenségek, mint pl. levegővétel, hangos nyelés, a krákogás, köhögés, egyéb testi hangok stb.

Mindezen hangesemények jelen vannak a spontán beszédben, és szerepük van az információátadásban. Megismerésük elengedhetetlen a természetes gépi beszédelőállítás és a gépi spontánbeszéd-felismerés megvalósításához.

Ebben a cikkben összefoglaljuk azokat a vizsgálatokat, amelyek a nyelvi tartalommal együtt megjelenő érzelmi, hangulati tartalomra vonatkoznak, azokra, amelyek prozódiai jellemzőkkel jutnak kifejezésre a beszédben a nyelvi tartalommal összefonódva. Továbbá csoportosítva tárgyaljuk azokat a nyelvi tartalomtól elhatárolt, attól független hangi jelenségeket, amelyek a spontán beszédben előfordulnak, és bemutatjuk az általunk létrehozott hanggesztustárat. Mindezen vizsgálatokhoz igen nagy mennyiségű spontán hanganyag gyűjtésére és feldolgozására volt szükség.

2 Módszer, adatbázisok

Vizsgálataink során 5 különböző spontán vagy közel spontánbeszéd-adatbázist dolgoztunk fel, amelyeket magunk vettünk fel, vagy médiából gyűjtöttünk. Ezek az alábbiak:

Magyar Telefonos Ügyfélszolgálati Beszéd Adatbázis (MTÜBA)

Ügyfél és diszpécser beszélgetése került rögzítésre, az adatbázis 1100 ilyen felvételből áll.

Maptask adatbázis

Az adatbázis 1113 „rövid” .wav fájlt tartalmaz, 10 különböző személlyel rögzített spontán beszéd útkeresés témában.

Balázs-show felvételek

135 percnyi műsoridő. 44 női és 99 férfibeszélő hanganyaga került feldolgozásra.

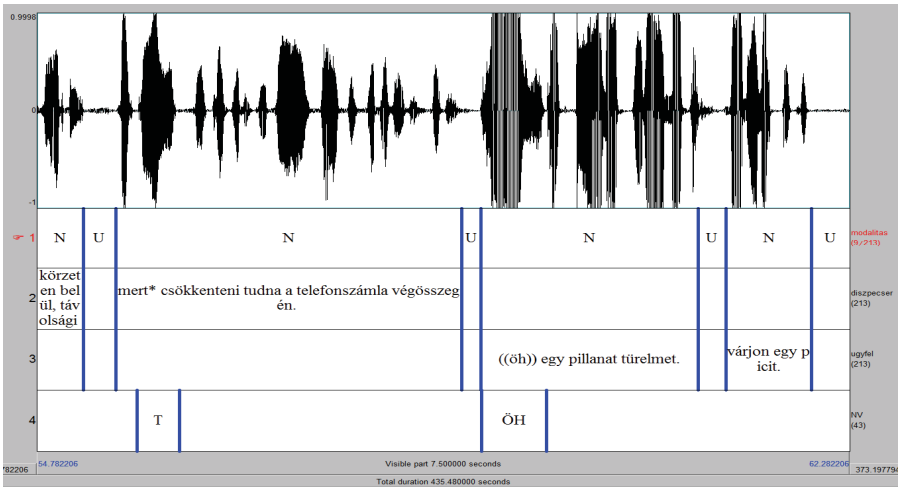
Joshi Bharat-felvételek

Szintén egy beszélgetős műsor, 61 percnyi műsoridővel.

Mozi

Végül pedig egy spanyol „Torrente” című 3 részes akcióvígjátékból gesztusok és egyéb nem verbális hangesemények kerültek kigyűjtésre.

A hanganyagok feldolgozása frázisegységenként [2] több szinten történt (lásd 2. ábra). Első szinten frázisonként bejelölésre került az adott frázisban kifejezésre jutó érzelem. A következő szinten/eken a nyelvi tartalom került bejegyzésre beszélőnként külön-külön. Az utolsó szinten a szövegben már csillaggal jelzett helyeknél lévő hangi események időtartama és típusa lett bejelölve.



2.ábra. Az adatbázisok többszintű feldolgozása. 1. Frázisonkénti érzelembejelölés (N: semleges, U:szünet); 2.3. Nyelvi tartalom bejelölése; 4. Nem verbális hangi események (T:kitöltött szünet ’t’ hang után, ÖH kitöltött szünet ’öh’-t ejtve)

Ezen adatbázisok vizsgálatával a társalgás során előforduló különböző nem verbális hangjelenségeket gyűjtöttük, amelyeket csoportosítottunk, és akusztikailag elemeztünk.

3 Nyelvi tartalommal együtt megjelenő érzelmi tartalom

Csak néhány éve kezdődött meg a beszéd különböző, nem verbális tartalmának, főként a hangulat kifejezésének, az érzelemnek a vizsgálata. Már korábban is érdekelte ez a kifejezési forma a kutatókat, de vizsgálataik során számos nehézségbe ütköztek, mivel a probléma igen összetett. A beszédben kifejezésre kerülő érzelmek vizsgálatának számos nehézsége van, melyek közül a leglényegesebbeket az alábbiakban soroljuk fel.

Statisztikai feldolgozásra, elegendő érzelmet kifejező spontán beszédanyag gyűjtése nehéz. Az irodalomban található ugyan néhány kutatási leírás, amely a beszéd emóciótartalmának vizsgálatával és az emóció automatikus, gépi felismerésével foglalkozik, de ezek az eredmények mind laboratóriumi körülmények között elhangzó tiszta beszédre vonatkoznak [3, 4, 5, 6]. A publikációk legtöbbjében szimulált emóciótartalmú beszédet használnak, leggyakrabban művészek bemondásmintáit. A valós szituációkban elhangzó, spontán beszédre jellemző adatok jelentősen különböznek a színészek által produkált beszédtől [7]. A beszédtechnológiai alkalmazásokban a valóságos spontán beszéd feldolgozása szükséges. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával [8] és információtartalmainak felismerésével [9] foglalkozik.

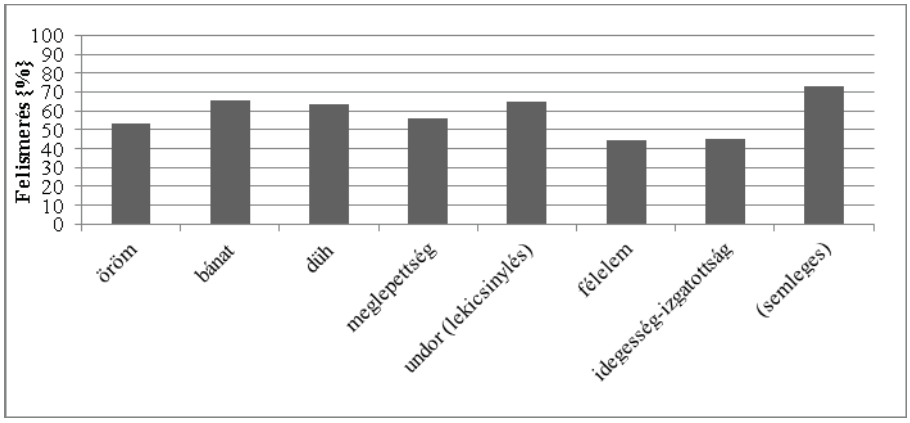
Problémát jelent továbbá az érzelmi kategóriák változatos megjelenése a beszédben. Az emóció jellemzésére a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban hagyományos emóciókategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban [10] e kategóriákat az arcmimika jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták. A beszédtechnológiai szakemberek ezeket a kategóriákat vették át a beszédben rejlő érzelem vizsgálatára is. Ha ezt összevetjük a valós helyzettel, az látszik, hogy a spontán beszédben sokkal változatosabb az érzelmi kategóriák tárháza, és ezek a téma szerint erősen változhatnak is. Kutatási céllal a spontán beszédben leggyakrabban előforduló érzelmi kategóriákat gyűjtöttek ki a PHYSTA 2001 adatbázisból [11]. Ez az adatbázis spontán társalgást, televíziós beszélgetőműsorok, és különböző vallási műsorok gyűjteményét tartalmazza (298 egység, 1 egység 10-60 s hosszú). A kiválasztott leggyakoribb érzelem és azok gyakorisága a 1 sz. táblázatban látható.

1. táblázat: Érzelmek csoportosítása és gyakoriságuk a PHYSTA 2001 spontán audiovizuális adatbázisban.

Címke	Használati gyakoriság	Csoport
Semleges	273	Nem erősen érzelemvezérelt
Dühös	114	Erősen negatív
Szomorú	94	Erősen negatív

Örvendező	44	Nem orientáltan pozitív
Boldog	37	Nem orientáltan pozitív
Jókedélyű	26	Nem orientáltan pozitív
Aggódó	19	Erősen negatív
Csalódott	17	Nem erősen érzelemvezérelt
Izgatott	17	Orientáltan pozitív
Félelem	13	Erősen negatív
Magabiztos	13	Nem erősen érzelemvezérelt
Érdeklődő	12	Nem erősen érzelemvezérelt
Gyengéd	10	Orientáltan pozitív
Elégedett	4	Nem erősen érzelemvezérelt
Szeretetteljes	3	Orientáltan pozitív

További problémát jelent a beszédben kifejezésre kerülő érzelmek vizsgálatánál, hogy a szemantikus tartalom (verbális csatorna) és a beszélő hangulatának, általános érzelmi állapotának a tükröződése (nem verbális csatorna) egyazon beszédfolyamatban valósul meg, és a szemantikus tartalom hozzájárul a beszéd emóciótartalmának a felismeréséhez is. Nyelvi tartalom nélkül az emberi emóciófelismerés sem jobb, mint 60-65%, a korábbi percepciós kutatások szerint [12]. Az említett munkában ugyanazon szemantikai tartalmú mondatok különböző érzelmekkel kerültek bemondásra két csoportban, színészekkel és átlagemberekkel (3 mondat, mondatonként 8 érzelem, 15 személlyel).



3. ábra. Az átlagemberek bemondásainak érzelmek szerinti felismerése percepciós teszttel mérve.

Ezeket a mondatokat meghallgattatták érzelem szerinti megítélésre 20 személlyel. A szubjektív lehallgatás eredményeit a 3. ábra mutatja.

A színészek és átlagemberek bemondásával kapott szubjektív lehallgatási eredmények között szignifikáns eltérés nem volt.

A helyzetet tovább bonyolítja, hogy az érzelmeinket a kommunikáció során, több érzékszervi csatornán keresztül juttatjuk el a másik félhez, e csatornák közül a legjelentősebb, a beszédhang maga, és az arcmimika (de még a testbeszéd, bőrpír és egyéb

tényezők is szerepet játszhatnak az érzelem kifejezésében). Agyunk az összes érzékszervi csatornán keresztül kapott információ együtteséről dönt [6]. Például egyes érzelmeket hallva az ember maga sem tud különbséget tenni a két érzelem között, de látva az arckifejezést, már könnyebben dönt. Az is megfigyelhető, hogy az ember érzelemfelismerési képessége csupán az arckifejezést látva meglepően jó. Az, hogy a hangi információ ad több információt vagy pedig a kép az érzelem felismeréséhez, az attól függ, hogy a hangi információban a nyelvi tartalom is benne van, vagy nincs. Amennyiben a hangi információ nyelvi tartalmat is ad, akkor csak hangi információ alapján lényegesen jobb a felismerés, mint csak az arckifejezés alapján. Ha viszont a hangi információ nyelvi tartalmat nem ad, pl. idegen nyelv esetén, akkor az arckifejezés alapján lesz jobb felismerés [13].

A hang- és képinformációt kombinálva javul a legjobban a felismerés minősége, eddig az automatikus felismerésben a kutatóknak megközelítőleg 80% körüli felismerést sikerült elérniük a kombinált információ felhasználásával [5].

Továbbiakban célunk csak a hang alapján történő érzelem kifejezés jellemző paramétereinek a vizsgálata. A fenti felsorolt nehézségek talán magyarázatul szolgálnak arra, hogy az eddig elért kutatások, kizárólag hang alapján, 60% körüli gépi felismerést értek el legjobb esetben is [1, 3, 6, 12]).

3.1 Beszédérzelmek jellemző vektorai a szakirodalomban

A gépi érzelem-felismerés során a meglévő hanganyagból jellemzővektorokat nyerünk ki, és ezeket használjuk fel az automatikus felismerő tanításához, majd ezekkel hajtjuk végre a felismerést. Ehhez persze tudni kell, hogy mik azok a jellemzők, amelyek jól leírják az emberi beszéd érzelmi tartalmát. Tehát először a beszédérzelem jellemzőit kell definiálni, kategorizálni.

A beszéd semleges érzelem kifejezésekor is rendkívül változatos, két különböző személy ugyanazt a mondatot másképp ejti ki, továbbá ugyanazt a mondatot, ugyanaz a személy sem ejti kétszer ugyanúgy. A kiejtett hangok fizikai paraméterei függhetnek a beszélő egészségi, fizikai állapotától is (megfázás, stressz, fáradtság, torokbetegségek). Mindezekhez hozzájárul még az a tény, hogy a beszélő a szándékától, érzelmi állapotától függően is változtathat egy mondat hangzásán, ezzel is kifejezve érzelmi állapotát. A beszédhang fizikai jellemzői tehát ugyanannál a szemantikai tartalomnál is sokfélék lehetnek.

Ez megnehezíti az érzelem gépi felismerését, hiszen meg kell tudnunk mondani, hogy mely változások játszanak fontos szerepet az érzelemkifejezésben, és melyek nem. A mai napig az ide vonatkozó szakirodalom egyik fő kérdése, hogy az automatikus érzelemfelismeréshez milyen jellemzőket kell kigyűjteni, amelyek alapján majd a felismerés működni fog.

Az irodalomban összefoglalóan az alábbi érzelmekre jellemző fizikai paraméterekkel találkozhatunk [14, 15]:

Alapszintű adatok a jellemzővektorokban

Az úgynevezett alapszintű jellemzők közé tartoznak a keretenkénti alaphangfrekvenciaértékek, a hangintenzitás-értékek, valamint a beszédhangok időtartama.

Az alaphang erősen beszélőfüggő, személyenként és időben változó érték. Mégis az irodalomban érzelmet tükröző alapszintű jellemzőnek tekintik.

A beszédhangok intenzitása és annak deriváltja is fontos paraméter, kifejezi a nyomatékokat, a hangsúlyokat. A témával foglalkozó cikkek mind besorolják a vizsgálandó paraméterek közé.

A harmadik alacsony szintű jellemző a szótagok, beszédhangok időtartama. Ezek meghatározzák a beszéd tempóját, ritmusváltásait.

Származtatott adatok a jellemzővektorokban

A származtatott jellemzőket az alapszintűekből képezzük, azok valamilyen változását, statisztikáját tekintve, melyet jellemzően egy mondatnyi hosszúságú beszédre számítanak ki. A cikkek szerint ezek a származtatott jellemzők meghatározzák az egyén beszédének prozódiai jegyeit. Információt hordoznak az intonációról, a tempóról és a hangerőről. Ilyen származtatott jellemzők az alaphang és az intenzitás maximuma, minimuma, átlagértéke, deriváltja, értéktartománya egy hosszabb közlésre, például egy mondatra.

Újabban már a színképi jellemzőket például a mel skálás frekvenciatartomány együtthatóit (MFCC-együtthatók) is besorolják az érzelmek jellemző paraméterei közé [12].

A származtatott jellemzők, amelyet az irodalomban mondategységekre számítottak ki, folyamatos spontán beszédben nem igazán vezettek eredményre, mivel a hosszabb összetett mondat szerkezete függvényében a mondat más-más részében jelenik meg az érzelem kifejezése.

Éppen ezért, a legújabb kutatások szerint [2] az érzelem kifejezésének alapegységeként a frázist tekintjük. Amennyiben frázisonként vizsgáljuk az érzelmek kifejezését, akkor nagyobb részben már ki tudjuk küszöbölni a mondat szerkezetétől való függést, ugyanakkor a frázis már elég hosszú beszédegység ahhoz, hogy érzelmet tükrözhessen.

A kérdés tehát az, hogy milyen fizikai paraméterek és azok milyen kombinációi tükrözik az egyes érzelmeket a frázisokban.

3.2 Beszédérzelmek jellemző vektorai frázisokban

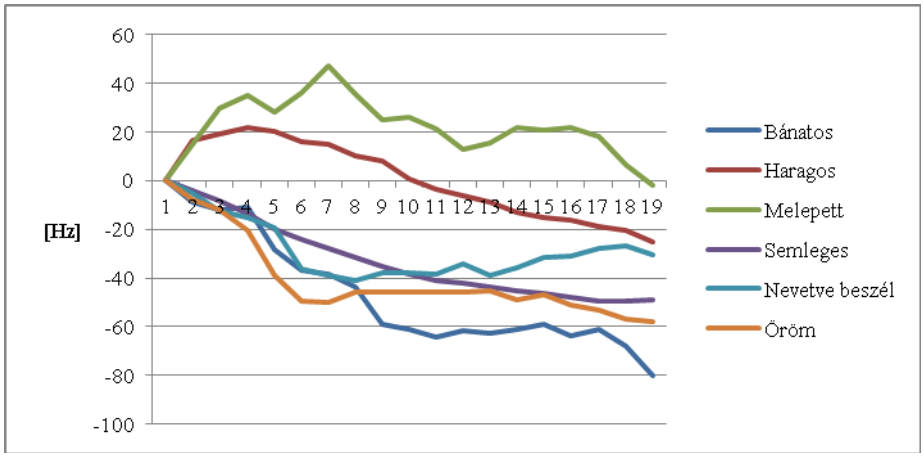
Jellemző vektorok vizsgálatát az összegyűjtött 5 spontán adatbázis felhasználásával végeztük el. Ezeknek az adatbázisoknak a feldolgozása során már kiderült, hogy csupán a tiszta érzelmek jelölése sem egyértelmű feladat, és rendszerint az annotátort a döntésben a szövegkörnyezet nagymértékben befolyásolja. Amennyiben azokat a prozódiai jellemzővektorokat akarjuk meghatározni, amelyek az érzelmi, hangulati tartalmat hordozzák a beszédben a nyelvi tartalom nélkül, akkor olyan mintákat kell elemeznünk, amelyek biztosan hordoznak ilyen információt. Az elemzéshez szükséges minták kiválasztása a szövegtartalomból kiragadott frázisok szubjektív lehallgatásával történt. (20 egyetemi hallgató, férfiak, nők vegyesen). Azokat a frázisokat tartottuk meg a további vizsgálatokhoz, amelyek esetében a hallgatók legalább 70%-a egy adott érzelemre ítélt. Így spontán 43 beszélő 1000 frázisát választottuk ki és osz-

tottuk be 6 különböző érzelmi kategóriába, amelyek a semleges, bánatos, haragos-ideges, meglepett, nevetve beszélő, örömöt kifejező.

Az alapszintű jellemzőket vizsgálva a kiválasztott hanganyagon, az volt a tapasztalat, hogy az alapfrekvencia és az intenzitás időbeli változása egy frázison belül jellemző a különböző érzelmekre.

A vizsgálati anyagban a különböző hosszúságú frázisok lineárisan vetemítésre kerültek úgy, hogy mindegyik minta „n"hosszúságú lett, majd a mért adatokat normáltuk a frázisban mért első átlagadat értékére úgy, hogy a mintavételezési pontoknál mért adatokból az első minta értéke levonásra került. Végül az érzelem szerinti csoportok frázisonkénti értékei átlagolásra kerültek, vagyis minden érzelemre elkészült az adott "**érzelemre jellemző átlagos hangminta-dinamika**" mind alapfrekvenciában, mind összintenzitásban.

Az alapfrekvencia dinamikája n=19 értékek esetén a 4. ábrán láthatók, ahol az alapfrekvencia szórás értékei 5-10 Hz közötti értékeknek adódtak. Az alapfrekvencia dinamika érzelem szerint szépen elkülönül az alábbiak szerint.



4. ábra. A különböző érzelmek átlagos alapfrekvencia-dinamikája. Vízszintes tengelyen a mintavételezési pontok láthatók.

Bánatos:

Alapfrekvencia folyamatos és nagymértékű csökkenését figyelhetjük meg. Majd körülbelül a frázis felénél, 60Hz-es csökkenés után egy stagnálást, majd a végén újabb csökkenést.

Haragos:

Az elején nő az alapfrekvencia, majd folyamatosan csökken.

Meglepett:

Az elején nagymértékű alapfrekvencia-növekedés látható, majd valamelyes csökkenés. Ennél az érzelemkategóriánál figyelhető meg leginkább az alapfrekvencia növekedése.

Semleges:

Az alapfrekvencia folyamatos szabályos csökkenése figyelhető meg, bár annak mértéke nem igazán jelentős.

Nevetve beszél:

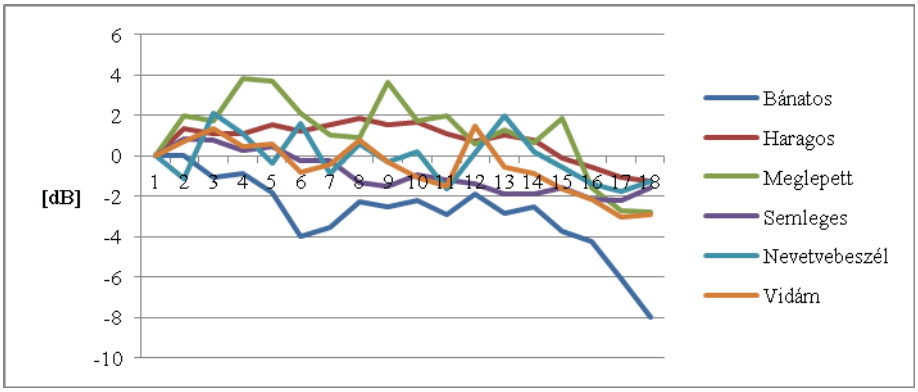
Az alapfrekvencia csökkenése, majd körülbelül a frázis felétől alacsony növekedése jellemzi.

Öröm:

Elején az alapfrekvencia lényeges csökkenése figyelhető meg a frázis felétől körülbelül 50Hz, majd utána stagnál, igen hasonlóan a nevetve beszél kategóriához.

Tehát a kísérlet alapján kijelenthető, hogy egy frázison belül az alapfrekvencia dinamikája jól jellemzi az érzelmeket.

A kísérlet tanulsága szerint alapvetően az egyes érzelemkategóriák intenzitásának dinamikái nem különülnek el olyan szépen, mint az alapfrekvencia változásának esetében, amint ez az 5. ábra alapján látható. Itt az értékek nem az első mintavételezési helytől kerültek ábrázolásra, hanem a másodiktól, emiatt az utolsó mintavételezési hely sorszáma a 18-as.



5. ábra. A különböző érzelmek átlagos intenzitásdinamikája. Vízszintes tengelyen a mintavételezési pontok láthatók.

A szórásértékek körülbelül 3dB értékűek voltak. Ez itt relatíve magas érték. Amit érdemes megfigyelni az az, hogy a „bánatos” érzelemnél jól látható és a többi érzelemtől elkülönült az intenzitás csökkenése, stagnálása, majd újabb csökkenése, illetve a „haragos” érzelemnél az intenzitás növekedése körülbelül a frázis feléig. A „semleges” érzelemnél az elején kicsi növekedés figyelhető meg, majd az érték folyamatos csökkenése. A „nevetve beszél” és a „vidám” érzelmeknél az intenzitás folyamatos változása figyelhető meg.

Az intenzitásértékek kevésbé tükrözik a különböző érzelmeket, bár azért jellemző dinamikajegyek az intenzitásnál is fellelhetők.

Érzelmekre jellemző lényeges színképi változás az idő függvényében a frázison belül nem tapasztalható, ugyanakkor egy frázisra átlagolt színképi paraméterek már érzelemre jellemző eltéréseket mutatnak.

Összefoglalva, a 43 beszélő 6 különböző spontán beszédben felvett érzelmi kategóriáinak statisztikai vizsgálata alapján elmondható, hogy az alapfrekvencia és az intenzitás frázison belüli időbeli változása, valamint egy frázis egészére átlagolt színképi paraméterek együttesen jellemzik a különböző érzelmeket. Az, hogy meg tudjuk mondani, melyik paraméter mikor és milyen súllyal járul hozzá a komplex érzelmi jellemzés kialakításához, még további kutatást igényel. Ezen jellemző vektorok alapján végzett automatikus érzelem-felismerési kísérletekről jelen kötetben egy másik cikk fog beszámolni.

4 A nyelvi tartalomtól független hangi események

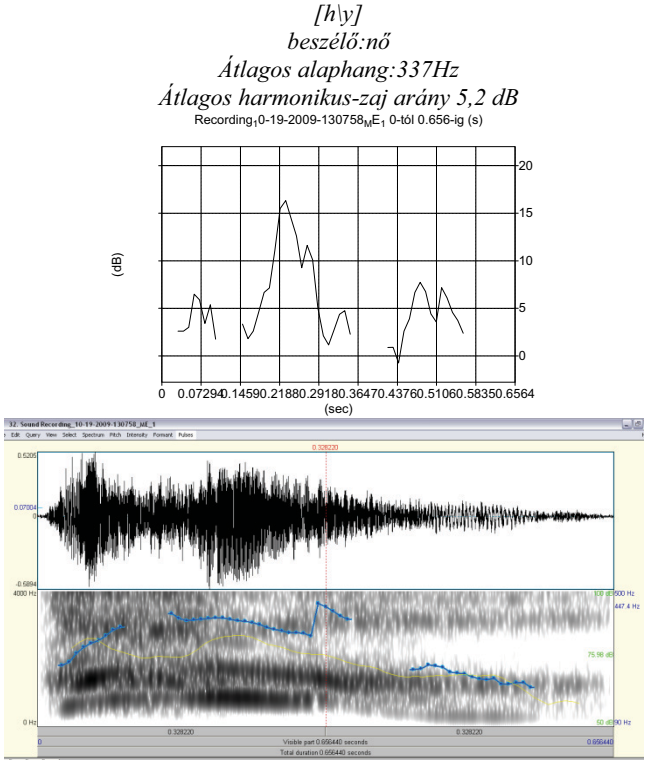
Az 5 felsorolt adatbázisban jelölésre kerültek azok a hangesemények is, amelyek a nyelvi tartalomtól elhatároltan, attól függetlenül jelentek meg. Ezek a jelentést kifejező hangjelenségek, vagyis a hanggesztusok, valamint a jelentéssel nem rendelkező hangjelenségek, kitöltött szünetek, testhangok. Bejelölésre kerültek még olyan a beszélgető partnerektől származó hangok, amelyek nem vokális eredetűek, mint például a csók vagy taps. Az 5 adatbázisban előforduló hangjelenségeket a 2. táblázatban soroljuk fel.

2. táblázat: A nyelvi tartalomtól elhatárolt, attól független hangi események.

Hanggesztusok	**Kitöltött szünetek**
L – nevetés(15)	A: – [A:] (25)
S –sírás(0)	d'2 – [(ho)d'2], [(pEdi)g2], [(ho)d'2], [(E)d'2],(72)
[jO], [jOj] (15)	h2 – [hA:t 2:] (47)
[nO]! (16)	ER – er… (18)
[(h\)u:h\O] (4)	k2**:** – [(ki:vA:no)k2:],[(tSO)k2:], [(mond'u)k2:] (7)
[h\A:t] (31)	2: – [2:] (66)
[h\y]?, [h\yF]? (8)	2:x: – [2:x:] (15)
yep (67)	2F: – [2F:], [2h\F:], [2yFh\:] (30)
[F:] (9)	2: – extrém hosszú[2:] (67)
h\F – hum! (csukott szájjal) (76)	r2: – [(Omiko)r2:] (7)
[ps]! (1)	t2: – [(mEr)t2:], [(tEh\A:)t2:] (72)
egyéb – (hahaha, há, fú, hóhó, phöhö, éé, hoppá, ú, húha, ajaja, háhá, nya, aő, au) (36)	

Nem vokális eredetű hang	**Testhangok**
KISS – csók hangja (3)	B – böfögés (2)
SLAP – tapsolás (2)	CO – köhögés (32)
	MO – csámcsogás (2)
	HIC –csuklás (4)
	BR – lélegzés (7)
	S – szipogás (11)
	SN – trüsszentés (1)

A kijelölt hangeseményeket kivágtuk és csoportokba gyűjtöttük. Megadtuk a csoportonként jellemző akusztikai jellemzőket. Így hoztunk létre egy ún. HANGGESZTUSTÁRAT, amelybe a hanggesztusokon kívül a 2. táblázat összes hangeseményét feltüntettük. A tárban a kigyűjtött hangesemények gyűjteménye található, az akusztikai jellegzetességeikkel együtt, továbbá egy-egy jellemző minta hangképe (spektrogram, alaphang, intenzitás, dinamika, harmonikus-zörej arány dB-ben), amint az a 6.ábrán látható. A tár alapja elkészült és azóta is folyamatosan bővül.



6. ábra. A [h\y] meglepődésgesztus adatai a hanggesztustárban. Balra: harmonikus-zörej arány dB-ben az idő függvényében, jobbra: amplitúdó-idő függvény, alatta spektrogram.

Távlati cél, annyi hanggesztus példa összegyűjtése egy-egy fajtából, hogy alkalmas legyen az adott hanggesztus akusztikai modelljének a felépítésére, ami majd az automatikus spontánbeszéd-felismerést fogja segíteni.

Köszönetnyilvánítás

Ez a kutatás a Jedlik OM-00102/2007 számú "TELEAUTO" projekt és a TÁMOP-4.2.2-08/1/KMR-2008-0007 projekt keretein belül készült.

Bibliográfia

1. Burkhardt, F., Paeschke A. et al.: A Database of German Emotional Speech. In: Proc. of Interspeech2005 (2005) 1517–1520
2. Vicsi, K., Sztahó, D.: Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In: Tanács A., Szauter D., Vincze V. (szerk.):VI. Magyar Számítógépes Nyelvészeti Konferencia. JATEPress, Szeged (2009) 217–225
3. Campbell, N.: Getting to the Heart of the Matter. Keynote Speech. In Proc. Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal (2004)
4. Campbell, N.: Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. In COST Action 2102 International Conference on Verbal and Nonverbal Features. Patras, Greece (2007) 107–120
5. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional Speech: Towards a New Generation of Databases. Speech Communication Vol. 40 (2003) 33–60
6. Hozian, V., Kacic, Z.: Context-Independent Multilingual Emotion Recognition from Speech Signals. International Journal of Speech Technology Vol. 6 (2003) 311–320
7. Kostoulas, T., Ganchev, T., Fakotakis, N.. Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data. In: COST Action 2102 International Conference on Verbal and Nonverbal Features. Patras, Greece (2007) 235–242
8. Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features. In: Building Corpora for Emotional TTS. IEEE transactions on audio, speech, and language processing Vol. 14, No. 4 (2006)
9. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence Vol. 2 No. 12 (1995) 1137–1143
10. MPEG-4 (1999): ISO/IEC 14496 standard. http://www.iec.ch

Nogueiras, A., Moreno, A., Bonafonte, A., Marino, J. B.: Speech Emotion Recognition Using Hidden Markov Models. In: Eurospeech (2001)

11. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G.,Kollias, S., Fellenz, W.,Taylor, J. Emotion Recognition in Human–Computer Interaction. IEEE Signal Process. Mag. Vol. 18 No.1 (2001) 32–80
12. Tóth, Sz. L., Sztahó, D., Vicsi, K.: Speech Emotion Perception by Human and Machine. In: Proceedings of COST Action 2102 International Conference. Patras, Greece, October 29-31, 2007. Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction 2008. ISBN: 978-3-540-70871-1. Springer LNCS (2008) 213–224
13. Esposito, A.: The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. Cogn. Comput DOI 10.1007/s12559-009-9017-8. Springer Science+Business Media, LLC (2009)
14. Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., Garay, N.: A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms. In: TSD LNAI 4629 (2007) 423–430
15. Seppänen, T., Väyrynen, E., Tovanen J.: Prosody-based classification of emotions in spoken Finnish. In: Eurospeech (2003)

Érzelmek automatikus osztályozása spontán beszédben

Sztahó Dávid, Imre Viktor, Vicsi Klára

Budapest Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
1111 Budapest, Stoczek utca 2.
sztaho@tmit.bme.hu, imreviktor.bmevik@gmail.com,
vicsi@tmit.bme.hu

Kivonat: A Budapesti Műszaki és Gazdaságtudományi Egyetem Beszédakusztikai Laboratóriumában automatikus érzelemfelismerésre, valamint automatikus beszéddetekcióra, illetve beszédszegmentálásra irányuló vizsgálatok folynak. A cikk ismerteti az érzelem felismerése során felhasznált különböző akusztikai jellemzőkkel kapott eredményeket, valamint a szupport vektor gép alapú gépi tanulási eljáráshoz használt spontán beszédet tartalmazó adatbázisokat. A beszéddetektálás, illetve beszédszegmentálás eredményeinek bemutatása során ismertetjük a rejtett Markov-modelleken alapuló felismerési eljárást, valamint a felhasznált telefonos adatbázist. Célunk egy olyan detektáló eljárás kidolgozása, amelyet alkalmazva, a szegmentált beszéden a fentebb említett érzelmi osztályozást el tudjuk végezni.

1 Bevezetés

Az automatikus érzelemfelismerés összetett probléma. Ahhoz, hogy valós időben meg lehessen valósítani, magán az érzelemfelismerésen kívül a beszéd valós idejű detektálásával és szegmentálásával is szembe kell nézni. Ennek a problémának a megoldása szintén kritikus fontosságú, ugyanis az előre elkészített és megfelelő beszédegységekkel betanított érzelemfelismerő működése e nélkül nem megvalósítható.

Ezért a Budapesti Műszaki és Gazdaságtudományi Egyetem Beszédakusztikai Laboratóriumában automatikus érzelemfelismerésre, valamint automatikus beszéddetekcióra, illetve beszédszegmentálásra irányuló vizsgálatokat végzünk. Adatbázisokat hoztunk létre, szegmentáltunk, illetve annotáltunk, amelyekkel a fenti feladatok elvégzésére alkalmas rendszereket kísérleteztünk ki.

Az emberek érzelemfelismerési képessége nyolc érzelem esetén (hét érzelem + semleges) 60-65%-ra adódik abban az esetben, amikor a nyelvi tartalom a döntésben nem játszik közre [1]. Ennél jobb felismerési eredményt egy géptől sem várhatunk el. További kérdés, hogy a felismerésben milyen akusztikai jellemzők játszanak közre. A cikkben az irodalomban [2, 3] megtalálható alapvető jellemzőkön kívül egyéb spektrális jellemzőket is felhasználunk. A beszédfelismerésben leggyakrabban alkalmazott alapegység a szavak, illetve a mondatok szintje. Az általunk választott alapvető időtartam azonban a korábbi eredményeink alapján [4] a frázis. Ezen belül kívánjuk az

érzelmeket felismerni. Ennek megfelelően az automatikus beszéddetektáló, illetve -szegmentáló eljárásnál is ekkora egységet tekintünk a felismerés alapegységének.

2 Beszéddetektálás

A valós idejű érzelemfelismerés problémája több összetevőből áll. Az audiojelben a spontán beszéd detektálása, valamint annak tagolása kiemelt tényező. Az általunk használt felismerési egység a frázis. Ebben a fejezetben bemutatjuk az automatikus beszéddetektáló eljárását, valamint a felhasznált adatbázist.

2.1 Telefonsávú felvételek beszéddetektáláshoz

A beszéddetektálási rendszer betanításához, teszteléséhez olyan beszédadatbázisra volt szükség, amely a felhasználási körülményekhez hasonló hanganyagot tartalmaz. A felhasznált adatbázist a BME Távközlési és Médiainformatikai Tanszék Beszédtechnológiai Laboratóriumának dolgozói és hallgatói készítették mobiltelefonnal. A felvételeket három különböző zajszintre lehet osztani. Vannak tiszta beszédjelet tartalmazó, nagyjából zajmentes környezetben készült felvételek. A zajjal terhelt beszélgetések további két részre bonthatóak: közepesen zajos, ahol a beszéd még jól érthető, de különböző háttérzajok fordulnak elő (autózaj, utcai zajos, háttérbeszéd); az erősen zajos felvételekben a beszéd már nehezen érthető.

1. táblázat: Felvételek száma osztályok szerint.

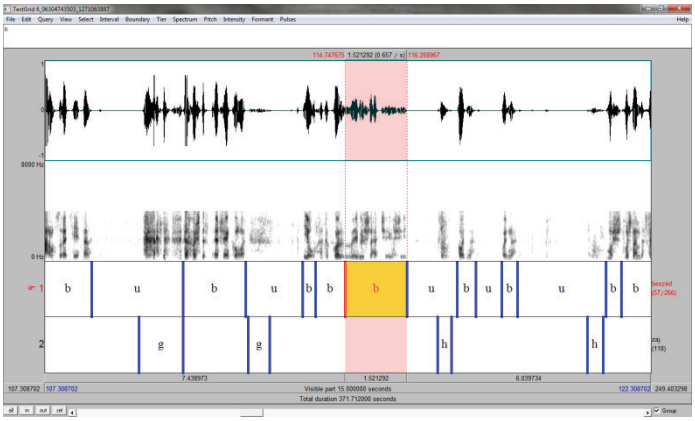
Zajszint	Felvételek száma
Alacsony	9
Közepes	16
Magas	6

2. táblázat: Alkalmazott jelölések az adatbázis annotálása során.

Sor neve	Hangtípus	Jelölés
beszéd	beszéd	b
	nem beszéd	u
zaj	gépjárműzaj	a
	gesztusok	g
	beszéd a háttérben	k
	szélzaj	s
	telefonhang	t
	recsegés	r
	sziréna	i
	ütés	h
	papírzörej	p
	levegővétel	l

A felvételek a felhasználás alapján is két csoportra oszthatóak: a kötött beszédet tartalmazó felvételek időben jól elkülönülő különálló mondatokat, míg a célzottan beszéddetektálásra készült felvételek egybefüggő, spontán beszédet tartalmaznak.

A felvételek annotálása során a frázisszintű címkézést a Praat szoftver felhasználásával végeztük el [5], amelyre egy mintát az 1. ábrán mutatunk be. A címkefájl két sort tartalmaz, a „beszéd” és „zaj” sávot. A beszédsávban a beszéd-nem beszéd részeket, és azok határait jelöltük. A zajsávban a különböző háttérzajokat és azok határait adtuk meg. A megkülönböztetett zajtípusokat a 2. táblázat tartalmazza.



1. ábra. Példa a kézi szegmentálásra.

2.2 Beszéddetektálási eljárás

Az automatikus felismerés rejtett Markov-modellek segítségével történt. Ehhez a HTK Toolkit-et [6] alkalmaztuk, amely egy beszédfelismerő keretrendszer, rejtett Markov-modell megvalósítással.

Az eljárás lényege, hogy a különböző zajtípusokra, valamint a beszéd (frázis) szakaszokra külön Markov-modelleket építünk, a 2.1. részben bemutatott adatbázis segítségével, amelyhez először egy akusztikai előfeldolgozást kell végezni. A beszéddetekció során, szintén egy akusztikai előfeldolgozás után, az egymás utáni időszakaszokra kapott legvalószínűbb Markov-modellek alapján lehetséges a beszéd szakaszok határainak bejelölése. Az eljárás erőssége az, hogy a felismert időszakasz hossza nem előre meghatározott, hanem változó hosszúságú lehet.

Az akusztikai előfeldolgozás során a következő jellemzőket használtuk fel a 3. táblázatban megadott számítási paraméterekkel. Ezután a kiszámított jellemzőket 50 ms-os ablakot alkalmazva kétszer deriváltuk. A végső tanítóvektorba az alapjellemzők, valamint az első, illetve második deriváltak kerültek.

A Markov-modellek építése során különböző hosszúságú (állapotszámú) modelleket alkalmaztunk a beszédre, valamint a zajokra. Előkísérletek alapján beszéd esetén 11 állapotú Markov-modelleket, zaj esetén 5 állapotú Markov-modelleket, valamint

csend esetén 3 állapotú Markov-modellek lettek elkészítve. Így a beszédrészeket az automatikus felismerő nem darabolja fel apró részekre, valamint a kevesebb állapotszámú zajmodellek segítségével a rövidebb időtartamú zajok is detektálhatóak.

3. táblázat: Felhasznált akusztikai jellemzők.

Jellemző	Időablak	Lépésköz
Alaphang	75 ms	10 ms
Intenzitás	250 ms	10 ms
Mel-frekvenciás kepsztrális együtthatók (MFCC)	500 és 250 ms	10 ms

A tanításra és tesztelésre következetesen elkülönített minták kerültek felhasználásra. Ez azt jelenti, hogy minden tesztet ugyanazon mintacsoporton végeztünk el, amelynek mintáit véletlenszerűen, de a változatosságot figyelembe véve válogattuk ki. Így extrém zajos, valamint normál minőségű, enyhén zajos (felhúzott ablak, kocsiban, nem kihangosítóval készült) minták is szerepeltek a tanító adatbázisban, valamint a tesztelő adatbázisban is.

A minőség kiértékelésére egy egyszerű, a döntést meggyorsító indexet használtunk. Két mátrixot számoltunk, melyekben beszúrási és tévesztési statisztikák szerepelnek. A beszúrási mátrix sorai azt mondják meg, hogy az eredetileg adott akusztikai osztálynak jelölt időintervallumok alatt hány darab jelölés található meg, tehát egy eredeti szakaszhoz mennyi felismert szakasz tartozik. A tévesztési mátrix sorai ehhez hasonlóan: az eredeti akusztikai osztály egyes intervallumaihoz mint (változó hosszúságú) időegységhez vesszük az ezen intervallumok alatt lévő jelölések időtartamát, tehát az eredeti szakaszokhoz időarányosan mennyi felismert időintervallum tartozik.

Ezek a mátrixok azonban bizonyos esetekben elég nagyok lehetnek, például sok címketípus esetén. Ez azzal a következménnyel jár, hogy nehezen átláthatóak, sok ideig tart, míg megállapítja valaki, hogy első közelítésben mennyire jó a felismerés. Ennek a kiküszöbölésére, az átláthatóság kedvéért egy egyszerű indexszámítást vezettünk be. Ez két részből áll: egyrészt az úgynevezett beszédindex, másrészt a zajindex. Ezeknek súlyozott összegéből adódik az összesített index, melyben a zajindex csak negyed súllyal szerepel. Ennek értelme az, hogy a végső felismerés céljából elhanyagolható, hogy a zajt milyen arányban találjuk el helyesen, ha a beszédet viszont annál jobban, mivel az automatikus felismerés végső célja a beszéd detektálása.

A beszédindex két összetevőből áll össze: beszúrási arány, valamint a tévesztési arány.

$$\text{beszúrási arány} = \frac{\text{az osztályban jól beszúrt intervallumok száma}}{\text{az osztály eredeti intervallumainak száma}}$$

$$\text{tévesztési arány} = \frac{\text{lefedett időintervallum száma}}{\text{az osztály eredeti intervallumainak száma}}$$

Látható, hogy a tévesztési arány maximuma 1, míg a beszúrási arány lényegében akármekkora lehet, így a beszédindexnek sem 100 a maximuma. Ahhoz, hogy legyen maximum, 100-nál törést kellett bevezetni, vagyis ha a beszúrási arány 1-nél nagyobb, akkor a beszédindexet maximalizáljuk. Az eredmények értékelésekor látható, hogy ez a változtatás a kiértékelhetőséget nem rontja. 80-as beszédindex körül már elfogadható felismerés adódik.

A későbbiekben egy, a zajos beszéd jelölésére szolgáló osztály ezt a számítási módot a következőképpen módosította: nem számít, hogy zajos beszéd és beszéd között mit döntünk, így ezeket ezután egyben kezeltük.

A zajindex az előzőekben elmondottakkal azonosan kerül kiszámításra az egyes zajokra, majd a végső index pedig ezeknek az átlaga. Az összesített index pedig:

$$\text{összindex} = \frac{3}{4} * \text{beszédindex} + \frac{1}{4} * \text{zajindex}$$

2.3 Eredmények

A tesztsorozat megkezdésekor a következő osztályok voltak felvéve tanításra: b (beszéd), u (csend/szünet), a (autózaj), g (gesztus), k (háttérbeszéd), s (szélzaj), t (telefonos jelzés), r (recsegés), i (sziréna).

A szirénahangot az első teszteléskor rögtön eltávolítottuk a tanított osztályok közül, mivel összesen egyetlen hangfájlban szerepelt, és abban is rövid ideig. A p (papírzörgés) és h (ütés/ütődés) hangokat a recsegéshez vontuk, elégtelen mennyiségű minta miatt, valamint a hangok akusztikai hasonlósága miatt. A tesztek során bevezettünk egy légzés címkét is, amely a telefonban jól hallhatóan a beszélőtől származó belégzési zörejeket fogja össze. Az 1. tesztsorozatban 100, 250, 500 és 750 milliszekundumos ablakokkal számolt mel-frekvenciás kepsztrális együtthatók, az intenzitás és az alaphang értékek szerepeltek, valamint ezek első, illetve második deriváltja. A legjobb eredményeket az 500 ms-os ablakmérettel számolt MFCC paraméterek esetén kaptuk (5. táblázat).

4. táblázat: Osztályokhoz rendelt Markov-modellek hossza.

Állapotszám	Címkék (osztályok)
11 állapotú modell	*b*, *k*
5 állapotú modell	*a*, *g*, *s*, *t*, *r*, *u*, *l*

A legrosszabb minőségű hangfájlok esetében (autóban, kihangosítóval) az osztályozási eredmények is rossz minőségűek lettek. Szinte egyáltalán nem ismert fel beszédet a rendszer ezekben a fájlokban. Ennek javítása érdekében bevezettünk egy zajos beszéd osztályt ("z" címkével jelölve). Az így kapott eredmények és az eredeti modellekkel kapott eredmények az 5. táblázatban láthatóak.

Az osztályozás további javításának érdekében többféle megközelítés szerint igyekeztünk módosítani a modelleket. A vélelmezett bonyolultság (akusztikai osztály összetettsége), az osztályozás alapján hibásnak vélt címkék, valamint az egyes hangminták átlagos hossza alapján hoztuk létre a modellek különböző csoportjait, ame-

lyekhez ezután különböző állapotszámú Markov-modelleket rendeltünk. Az így kapott osztálycsoportokat, valamint a hozzájuk tartozó felismerés eredményét a 6. és 7. táblázat mutatja.

5. táblázat: A legjobb, 500 ms-os időablakkal kapott osztályozási eredmények a különböző indexek szerint [%]-ban.

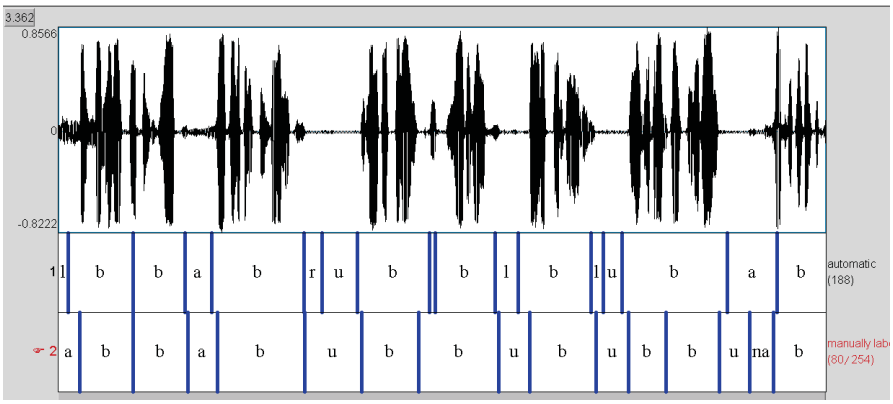
	Eredeti modellek esetén			Zajos beszédmodell bevezetése után		
Hangfelvétel-azonosító	Beszédindex	Zajindex	Összindex	Beszédindex	Zajindex	Összindex
01	0,69	63,95	16,51	46,81	63,3	50,93
02	11,36	24,29	14,59	32,74	24,29	30,6
03	100	33,7	83,42	100	35,58	83,89
04	83,62	29,39	70,07	68,43	29,07	58,59
05	100	15,34	78,84	82,64	9,8	64,43
06	98,75	22,9	79,79	98,88	23,34	79,99
07	67,22	33,4	58,76	76,8	33,28	65,92
08	83,61	33,1	70,98	84,22	32,71	71,34
09	76,31	0,46	57,35	80,06	0,58	60,19
10	84,55	36,79	72,61	88,82	38,24	76,17

6. táblázat: A módosított osztálycsoportosítás eredménye.

Állapotszám	Osztályok
14	b, z, k
11	s, a, u
5	g, r
4	l, t

7. táblázat: A módosított osztálycsoportokkal kapott felismerési eredmény [%]-ban.

Hangfelvétel azonosító	Beszédindex	Zajindex	Összindex
01	49,65	57,96	51,73
02	16,75	28,95	19,79
03	100	38,34	84,58
04	87,23	17,75	69,86
05	82,64	8,61	64,13
06	100	29,2	82,3
07	65,2	30,09	56,42
08	86,91	37,24	74,49
09	83,24	0,58	62,57
10	88,1	36,89	75,3



2. ábra. Példa az automatikus osztályozás eredményére.

3 Érzelemfelismerés

3.1 Érzelmi adatbázis

Az érzelemfelismerés megvalósításához folyamatos beszélgetéseket tartalmazó spontán telefonos felvételek, különböző talkshow-k felvételei kerültek összegyűjtésre, valamint annotálásra. A folyamatos beszéd frázisegységekre lett feltagolva, a frázisok pedig érzelem szerint lettek annotálva, mely során a legjellemzőbb érzelmi minták kerültek bejelölésre. A folyamatos feldolgozás során az derült ki, hogy a szövegkörnyezet figyelembevétele nélkül a frázis egységek érzelmi osztályozása számos esetben nem egyértelmű. Ezért a bejelölést végző személyeknek ezután csupán az érzelemmel töltött részeket kellett megjelölni, azok osztályozását külön szubjektív tesztsorozat során több lehallgató végezte el. Így végül 2540 érzelmes szakasz szubjektív lehallgatását 30 személy végezte el, amelyek után végül 43 beszélőtől, összesen 985 érzelmes szakasz lett kiválasztva, 6 érzelem szerint. A kiválasztás során csupán azokat a hangmintákat válogattuk ki, amelyeknél a szubjektív lehallgatás során 70%-os egyezés volt a döntésekben. Az érzelmek az alábbiak voltak: semleges, szomorú, meglepett, dühös/ideges, nevetés beszéd közben, valamint boldog. A kategóriák közötti eloszlást a 8. táblázat mutatja.

8. táblázat: A 30 lehallgató személy által kiválasztott érzelmes minták száma.

Érzelemtípus	**Frázisok száma (a lehallgatók döntéseinek 70%-os egyezése)**
Semleges	517
Dühös/ideges	290
Boldog	39
Nevetve beszél	42
Szomorú	54
Meglepett	43

3.2 Érzelemfelismerési eljárás

Az érzelemfelismerési kísérletek során végül 4 érzelmet használtunk fel, mivel ezekhez volt elegendő hangminta, amellyel tanítani lehetett. A 10. táblázat alapján ezek a következők: semleges, harag/ideges, öröm és nevetve beszél együtt, szomorú. Az automatikus osztályozáshoz szupport vektor gépeket alkalmaztunk, amelyhez az SVMLib [7] szabadon letölthető C# programozási nyelvű könyvtárcsomagját használtuk. A kísérletek célja az volt, hogy megvizsgáljuk, milyen akusztikai jellemzők szükségesek az érzelem felismeréséhez.

A következő jellemzőket vizsgáltuk meg:

- az alaphang átlaga, maximuma, tartománya és szórása (jelölés: F0)
- az alaphang deriváltjának átlaga, maximuma, tartománya és szórása (jelölés: ΔF0)
- az intenzitás átlaga, maximuma, tartománya és szórása (jelölés: EN)
- az intenzitás deriváltjának átlaga, maximuma, tartománya és szórása (jelölés: ΔEN)
- 12 mel-frekvenciás kepsztrális együttható átlaga, maximuma, tartománya és szórása (jelölés: $MFCC_i$)
- harmonicity értékek átlaga, maximuma, tartománya és szórása (jelölés: HARM)

Minden jellemzőt 10 ms-os lépésközzel nyertünk ki, majd frázisonként számoltuk ki a megfelelő statisztikai jellemzőt. Így egy frázisra egy ilyen érték adódott, ezekből állt végül elő a hangmintához tartozó jellemzővektor.

3.3 Eredmények

A tesztek során a következő osztályjelölések szerepelnek: harag/ideges: A, boldog: J, semleges: N, szomorú: S. A 9. táblázat(csoport) négy kísérleti összeállítás eredményeit tartalmazza.

9. táblázat: Automatikus felismerési eredmények [%]-ban négy jellemzővektor-összeállítás esetén.

jellemzővektor: F0, ΔF0, EN, ΔEN				
	A	**J**	**N**	**S**
A	51	15	5	4
J	18	32	17	2
N	6	9	57	3
S	15	4	13	7
			Felismerési eredmény: 56,98	

jellemzővektor: F0, ΔF0, EN, ΔEN, HARM,				
	A	**J**	**N**	**S**
A	46	13	10	6
J	17	30	16	6
N	7	8	56	4
S	12	7	12	8
			Felismerési eredmény: 54,26	

jellemzővektor: F0, ΔF0, EN, ΔEN, $MFCC_i$				
	A	**J**	**N**	**S**
A	57	13	4	1
J	12	37	13	7
N	4	12	55	4
S	5	17	5	12
			Felismerési eredmény: 62,40	

jellemzővektor: F0, ΔF0, EN, ΔEN, HARM, $MFCC_i$				
	A	**J**	**N**	**S**
A	61	9	4	1
J	11	41	11	6
N	3	12	56	4
S	5	16	5	13
			Felismerési eredmény: 66,27	

A felismerési eredmények azt mutatják, hogy az alapjellemzőkön kívül (alaphang, intenzitás) a mel-frekvenciás mel-kepsztrum jellemzők nagy szerepet játszanak az automatikus felismerésben. A harmonicity értékek ezt még javítani tudják. Ám mivel a minták száma jelenleg még nem kielégítő, ezért ahhoz, hogy ezeket az eredményeket megbízhatóbbá tegyük, folyamatos adatbázisgyűjtés és -feldolgozás szükséges.

Annak ellenére, hogy a tesztek során az alaphang és intenzitás értékek normáltan szerepeltek a jellemzővektorban, érdemes megnézni az eredményeket akkor, ha a hangmintákat külön válogatjuk női, illetve férfi mintákra. Ennek eredménye látható a 10. táblázatban. Habár a felismerés enyhe javulást mutat, a hangminták nem kielégítő száma miatt ez csupán pár hangmintaeltérést jelent.

10. táblázat: Automatikus felismerés eredménye [%]-ban női és férfi hangminták esetén a legjobb felismerési teljesítményt adott jellemzővektor esetén.

férfi beszélők

	A	**J**	**N**	**S**
A	17	0	4	1
J	1	7	2	7
N	2	2	18	0
S	1	5	0	14

Felismerési eredmény: 69,14

női beszélők

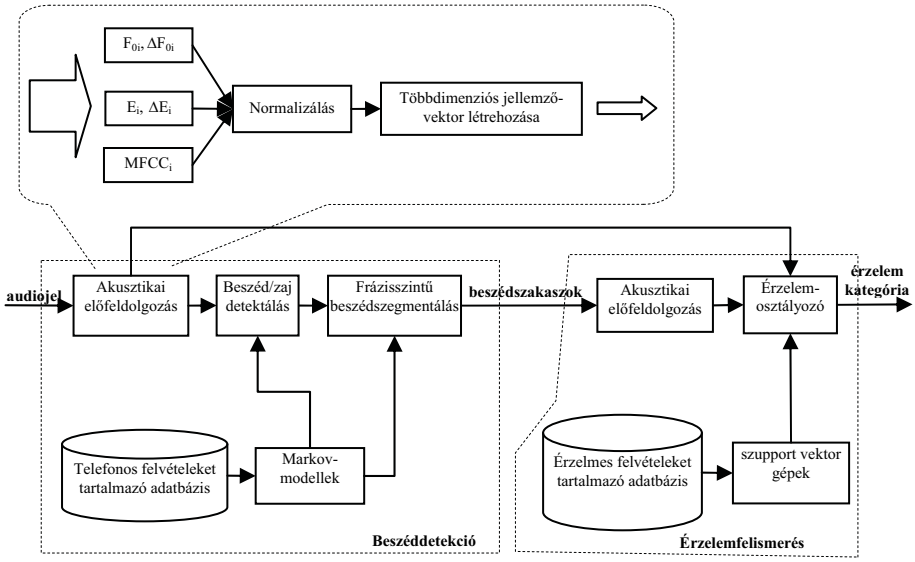
	A	**J**	**N**	**S**
A	46	6	1	0
J	9	31	11	1
N	1	9	40	3
S	3	8	6	2

Felismerési eredmény: 67,23

4 Kvázi valós idejű beszédfelismerési eljárás terve spontán beszédben

Beszédkommunikáció közben, főként hosszú beszélgetés esetén, a beszélő személy érzelmi állapota folyamatosan változik. Annak érdekében, hogy a beszélő mentális állapotát követni tudjuk, a folyamatos beszélgetést szakaszokra kell tagolnunk. Jelen esetünkben a frázist választottuk a szegmentálás alapegységének.

Az automatikus frázisszintű szegmentálást a megvalósítandó valós idejű felismerőben a már fentebb bemutatott beszéddetektáló végzi. Az egybeépített automatikus felismerő blokkvázlata a 3. ábrán látható. Az ábrán a fentebb bemutatott két különálló felismerő akusztikai feldolgozása külön szerepel, mivel azokat két különálló modul végzi. A végső szoftverben azonban sebességoptimalizálási célból ezt egyetlen modul fogja végezni.



3. ábra. Az automatikus érzelemfelismerő blokk vázlata spontán beszéd esetén.

5 Összefoglalás

A cikkben bemutatásra került egy olyan automatikus érzelemfelismerési eljárás, amely spontán zajos környezetű beszédben, valós időben képes érzelmek felismerésére kizárólag a beszéd prozódiai jellemzői alapján.

Ehhez kifejlesztettünk egy olyan rejtett Markov-modelleken alapuló eljárást, amely a hanganyagot frázisegységekre szegmentálja, és osztályozza beszédosztályra, valamint egyéb akusztikai környezeti zajosztályokra. Így oldva meg a beszéd-nem beszéd detektálást és a frázisszintű szegmentálást.

A beszéddetektálási eredmények kiértékelése során megállapítható, hogy a detektáló eljárás alkalmazható spontán beszédre. A kapott beszédindex-eredmény nem kiemelkedően zajos felvételek esetén eléri a 80 %-ot, ami, ahogy az eredményeket bemutató ábrán is látható, elfogadható teljesítmény.

A detektáló, fázisszegmentáló eljárást követi az érzelemfelismerő eljárás. Négy érzelemre szubjektív lehallgatással kiválogatott hangminták betanítása esetén a szupport vektor gép alapú automatikus felismerő 66%-ban osztályozta megfelelően az érzelmes hangmintákat.

Köszönetnyilvánítás

Ez a kutatás a Jedlik OM-00102/2007 számú "TELEAUTO" projekt és a TÁMOP-4.2.2-08/1/KMR-2008-0007 projekt keretein belül készült.

Bibliográfia

1. Tóth, Sz. L., Sztahó, D., Vicsi, K.: Speech Emotion Perception by Human and Machine. In: Proceedings of COST Action 2102 International Conference. Patras, Greece, October 29-31, 2007. Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction 2008. ISBN: 978-3-540-70871-1. Springer LNCS (2008) 213–224
2. Hozjan, V., Kacic, Z.: A rule-based emotion-dependent feature extraction method for emotion analysis from speech. The Journal of the Acoustical Society of America. Vol. 119 No. 5 (2006) 3109–3120
3. Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE Transactions on Audio, Speech and Language Processing Vol. 14 No.4 (2006)
4. Vicsi K., Sztahó D.: Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2009) 217–225
5. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org
6. The Hidden Markov Model Toolkit (HTK). http://htk.eng.cam.ac.uk/
7. Chang, C.C., Lin, C-J.: LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)

VI. Morfológia, korpusz

Ismeretlen kifejezések és a szófaji egyértelműsítés

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

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{zsibrita, vinczev}@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: A jelenleg használt magyar morfológiai elemző és szófaji egyértelműsítő eszközök számos esetben nem működnek megfelelően, elsősorban az ismeretlen (szótárban nem szereplő) szavak és kifejezések kezelése miatt. Előadásunkban bemutatunk egy új (teljesen JAVA-ban implementált) szófaji egyértelműsítő rendszert („magyarlanc"), amely a morphdb.hu nyelvi erőforrásra épülő morfológiai elemzőn és számos, ismeretlen kifejezések kezelésére kidolgozott szabályon alapul.

1 Bevezetés

Ebben a munkában bemutatjuk a `magyarlanc`-nak keresztelt szegmentáló és szófaji egyértelműsítő rendszerünket. A rendszer a morphdb.hu nyelvi erőforrásra [8] épül, de számos ponton kiegészíti (alternatívája) a hunpos rendszernek [4]. A legfontosabb eltérések:

- a harmonizált KR-MSD kódrendszert használja [3], így a Szeged Korpuszon [1] közvetlenül tanítottuk,
- relatív szótöveket ad eredményül,
- teljesen JAVA nyelven implementált, így könnyen integrálható nagy (akár webszerver) alkalmazásokba,
- számos szabályt tartalmaz ismeretlen kifejezések kezelésére.

A következő fejezetekben röviden bemutatjuk az egész elemző láncot, majd az utolsó pontot tárgyaljuk részletesen.

2 Kapcsolódó munkák

Számos magyar nyelvre kidolgozott szófaji egyértelműsítő rendszer látott már napvilágot. A Szegedi Tudományegyetemen két szófaji egyértelműsítő is készült korábban: egy, a rejtett Markov-modellre épülő statisztikai módszer, illetve a szabályalapú RGLearn algoritmus [5]. A két módszert kombinálták a TnT taggerrel is: a hibrid algoritmus körülbelül 1%-os javulást eredményez a szófaji egyértelműsítésben a

Szeged Korpusz 2.0-n mérve. A BME MOKK fejlesztése a hunpos, egy ingyenes és nyílt forráskódú HMM-alapú szófaji egyértelműsítő [4], egy nyílt forráskódú implementációja a TnT-nek. Itt az elsődleges cél az ismeretlen szavak morfológiai kódjának minél pontosabb megállapítása volt. A hunpos OCaml nyelven készült, egy magasrendű nyelven, mely támogatja a tömör, könnyen érthető kódolási stílust[1]. A HuMOR morfológiai elemzőre[2] is épült egy ismeretlen szavakat elemző rendszer: a szimbolikus megszorításokon alapuló részleges elemző a Magyar Nemzeti Szövegtárból[3] nyert statisztikai információval egészül ki [7]. Ezen – kifejezetten szófaji egyértelműsítésre mint célfeladatra kidolgozott – rendszerek mellett a szófaji egyértelműsítőt mint köztes lépést használják a magasabb rendű magyar szintaktikai elemzők is, mint például az MTA Nyelvtudományi Intézetében magyarra átültetett NooJ[4] és a MorphoLogic kft. MetaMorpho MorphoParse-ja[5].

A bonyolultabb morfológiával rendelkező nyelvek esetében a HMM-alapú egyértelműsítés versenyképesnek bizonyul a többek között SVM vagy CRF módszereken alapuló tanuló algoritmusok jelenlegi generációjával szemben. A magyarban, mint más erősen ragozó nyelvekben igen fontos megőrizni a részletes morfológiai információkat a szófaji kódokban annak érdekében, hogy a magasabb rendű feldolgozási feladatokban is hasznosíthatóak legyenek. Ez az angolban használatosnál jóval nagyobb kódhalmazhoz vezet (kódrendszertől függően akár 1000 körüli is lehet a címkék száma az angol treebankekben rendszerint alkalmazott 36-hoz képest), azonban ez nem válik a tanítás és az egyértelműsítés hátrányára, noha a nem generatív modellek tanító folyamatát számítási szempontból megdrágítja.

3 magyarlanc

A `magyarlanc` programcsomag[6] magyar nyelvű szövegek alap nyelvi elemzésére szolgál. A csomag tisztán JAVA nyelvű modulokat tartalmaz, ami biztosítja a platformfüggetlenséget és a nagyobb rendszerekbe (például webszerverek) történő integrálhatóságot. A csomag magában foglal egy angol/magyar nyelvdetektort, magyar nyelvre adaptált mondat- és tokenszegmentálót[7], illetve egy szófaji elemzőt.

[1] http://mokk.bme.hu/resources/hunpos

[2] http://www.morphologic.hu/Morfologiai-elemzes.html

[3] http://corpus.nytud.hu/mnsz/

[4] http://corpus.nytud.hu/nooj/

[5] http://www.morphologic.hu/MetaMorpho-technologia

[6] A rendszer nyílt forráskódú, a Creative Commons licenc alatt szabadon hozzáférhető: http://www.inf.u-szeged.hu/rgai/magyarlanc

[7] Kiindulási alapként a morphadorner rendszer szegmentálóit használtuk: http://morphadorner.northwestern.edu/

3.1 Szófaji elemző

A szófaji elemző (lemmatizáló és POS-tagger) a Stanford POS-tagger[8] egy módosított változata, amely az ismeretlen szavakra a morfológiai elemző által adott lehetséges elemzéseket használja fel (az eredeti implementáció az ismeretlen szavakra az összes lehetséges morfológiai kódból választ). A POS-taggert a Szeged Treebanken [1] tanítottuk az automatikus morfológiai elemzéseket bemenetként felhasználva. A tanítás folyamán egy csökkentett MSD-kódhalmazt (42 kóddal) használtuk, hogy a lehetséges címkék számát kezelhető korlátok közé szorítsuk. A csökkentett kódhalmazban a szófaji alkategóriákat csak akkor vettük fel, ha a megkülönböztetés egyes szóalakok esetén szükségesnek látszott a Szeged Korpusz alapján (például megkülönböztetjük a főneveken belül a részes és birtokos esetben állókat). A kódhalmaz redukálásánál azt az irányelvet követtük, hogy a csökkentett kódkészletet használó szófaji egyértelműsítő modul kimenete egyértelműen megfeleltethető legyen az eredeti MSD-kódoknak. Tehát például az Nc-sd és Nc-sg kódok redukált alakja különbözik, míg a Nc-sd és Nc-sd---s3 ugyanarra a kódra redukálódik, mert soha nem fordulhat elő, hogy egy szóalaknak ez a két kód lehetséges elemzése (és a szófaji egyértelműsítőnek döntenie kell köztük).

3.2 Morfológiai elemző

Ahogyan az előző fejezetben bemutattuk, azon szóalakok esetén, amelyek nem szerepeltek a tanító adatbázisban, egy morfológiai elemző meghatározza a lehetséges elemzések halmazát, majd a szófaji egyértelműsítő modulnak ezen halmazból kell választania. Az alkalmazott morfológiai elemző a morphdb.hu nyelvi erőforrás [8] egy új változatára épül. Az új verzióban a KR és MSD kódrendszer harmonizált verziója található meg [3]. A nyelvi erőforrást mint bemenetet használva, Gyepesi György szoftvercsomagja egy véges állapotú (karakterátmeneteket használó) automatát állít elő. Az elemzés eredménye egy KR-kódhalmaz, mely visszaírási információkat is tartalmaz. A morfológiai kódharmonizációnak és a visszaírási információknak köszönhetően ezek a kódok egyértelműen megfeleltethetőek egy MSD-kódnak és a hozzá tartozó relatív szótőnek. A megfeleltetést végrehajtva már közvetlenül használhatjuk a morfológiai elemzőt a szófaji elemző tanítására és kiértékelésére a Szeged Korpuszon.

Természetesen egyetlen nyelvi erőforrás sem lehet tökéletes fedésű. A következő fejezetben bemutatunk néhány egyszerű megoldást azoknak az eseteknek a kezelésére, amelyekre a morphdb.hu erőforrásra épülő automata nem ad egyetlen morfológiai elemzést sem.

[8] http://nlp.stanford.edu/software/tagger.shtml

4 Ismeretlen szóalakok kezelése

Ismeretlen szóalakok kezelésére kidolgoztunk néhány egyszerű megoldást (amelyek a `magyarlanc`-ba beépítésre kerültek). A Szeged Korpusz 2.5-ben 143612 különböző szóalak fordul elő. A morphdb.hu jelen verzióira épült automata ezeknek nagyságrendileg (l. következő alfejezet) 75%-ára ad legalább egy elemzést. A fejezetben bemutatásra kerülő egyszerű módszerek segítségével az ismeretlen szavak (amelyekre az eredeti automata nem ad elemzést) háromnegyedére kapunk elemzést.

4.1 Tulajdonnév gazetteer a morfológiai elemzéshez

Első lépésben megvizsgáltuk azt is, hogy milyen hatásai vannak az alap nyelvi erőforrás (morphdb.hu) tulajdonnevekkel történő felbővítésének, ugyanis az ismeretlen szavak nagy része tulajdonnév. Az alábbi táblázatban láthatóak a Szeged Korpuszon tanított és kiértékelt POS-tagger eredményei, amelyek csak a morfológiai elemzőhöz felhasznált tulajdonnév gazetteerben térnek el egymástól (a kiértékelési módszertan pontos leírását l. az 5.3 fejezetben).

1. táblázat: Különböző méretű tulajdonnév gazetteerek eredményei.

#tulajdonnév	Ismeretlen szavak	főnév (P/R/F)	összes szófaj (P/R/F)
498	24,79%	70,50/85,53/77,29	77,04/79,11/78,06
339133	19,47%	72,89/88,87/80,09	79,43/79,65/79,54

A felbővített alapszótárral 111199 szóalakra kapunk legalább egy elemzést (80,53% az ismert szavak aránya) és a szófaji egyértelműsítő rendszereknek mind a pontosságát, mind fedését javította. Az alább bemutatásra kerülő kísérleteink során minden esetben ezt a felbővített alapszótárból kiinduló morfológiai elemzőt használtuk.

4.2 Arab és római számok

Az ismeretlen esetek egy jelentős részét az arab és római számok képezték. Ezek nyílt tokenosztályt alkotnak. A véges állapotú automata kiegészíthető lenne speciális állapotokkal és átmeneti szabályokkal ezeknek a felismerésére, ami tulajdonképpen egy független automatát jelentene. A `magyarlanc`-ban egyszerű reguláris kifejezésekkel ismerjük fel ezeket (megjegyezzük, hogy kifejezéseink nem kiterjesztettek, így reguláris nyelvet generálnak, azaz ekvivalensek egy determinisztikus véges állapotú automatával). A kidolgozott reguláris kifejezések megkülönböztetik a sorszámneveket, a tőszámneveket, a törtszámneveket és osztószámneveket, valamint ezek nyelvtani eseteit is, és összesen 5708 szóalakra adnak elemzést (az ismeretlen szóalakok 21,23%-a).

4.3 Összetett szavak

Az összetett szavak szótárban történő felsorolása soha nem lesz tökéletes fedésű, míg az összetétel tagjai általában ismertek (pl. *szárny+fesz+táv*). Elemzésükkor ki lehet ezt használni, oly módon, hogy ismert összetevőkre bontjuk azt és ellenőrizzük, hogy érvényes összetételről van-e szó (például a *vírusgazda* szó *vírus*ra és *gazdá*ra történő felbontása után mindkét összetevő értelmes, de a *futár fut+ár* felbontása nem értelmes). A balról jobbra haladó véges állapotú automatás morfológiai elemzők is alkalmassá tehetők az összetett szavak elemzésére, például ha megsokszorozzuk az állapotokat és megkülönböztetjük a *táv* elemzéseit aszerint, hogy a szó elején vagyunk vagy már egy elemzett főnév vagy ige megelőzi azt.

Az általunk javasolt eljárás ennél jóval egyszerűbb és hatékonyabb. Amennyiben egy szóalakra nincs elemzésünk, megvizsgáljuk, hogy az összetett szó-e. Ehhez megkeressük a szó minden lehetséges (legfeljebb háromtagú) felbontását. Azokat a felbontásokat tekintjük lehetségesnek, ahol minden egyes összetevőnek van legalább egy elemzése az eredeti automata szerint. Vannak azonban olyan pszeudoösszetételek, amelyek nem érvényesek. Ezek kiszűrésére szakértői szabályokat adtunk meg, mint például: ha az első összetevőnek csak igei elemzése van, és a másodiknak nincs igei elemzése, akkor nem érvényes az összetétel. Az eljárás végén minden érvényes összetételt lehetséges elemzésként ajánlunk fel az utolsó összetevő morfológiai kódjával, illetve az utolsó összetevőt lemmatizáljuk (például a *részrehajlónak* szóalak esetén a lemma *részrehajló*).

A Szeged Korpuszon ezzel a módszerrel 12012 olyan szó helyes elemzését kaptuk meg a lehetséges elemzések között, amelyet az eredeti automata nem elemzett (az ismeretlen szavak 44,67%-a), és mindössze 1654 szóra (ismeretlen szavak 6,15%-a) ad a módszer helytelenül összetett szavas elemzést.

4.4 Kötőjelet tartalmazó tulajdonnevek

Az összetételek egy speciális esete, amikor kötőjellel képzünk egy ismeretlen szóból (általában tulajdonnév) és ismert köznévből álló összetételt (például *Bush-kormány*), ahol tehát már nem is szükséges minden összetevő „ismerete”. Egy utófeldolgozó lépésben minden olyan szót megvizsgálunk, amely tartalmaz kötőjelet. Amennyiben a kötőjel utáni rész egy főnév, feltehetjük, hogy ez egy tulajdonnév-köznév összetétel, és főnévnek jelöljük a köznév morfológiai kódjaival és relatív lemmájával (a *Telenor-csoporttal*-nak *Telenor-csoport* lesz a lemmája).

Hasonló módon, ha a kötőjel után egy lehetséges főnévi toldalék áll, akkor feltesszük, hogy a kötőjel előtti rész egy tulajdonnév, és főnévnek jelöljük a toldalék által megadott esettel és a kötőjel előtti résszel mint szótő (például a *Vodafone-nak* szóalak lemmája *Vodafone*). Mivel az összes lehetséges főnévi toldalékot nem akartuk felsorolni, más módszerhez folyamodtunk: a különböző morfofonológiai osztályokra választottunk egy-egy főnevet (*lány*, *némber*, *sün*, *fal*, *holló*, *felhő*, *kalap*, *kert*, *köd, néni*) és ellenőrizzük, hogy a toldalékot a mintafőnév után írva főnévi elemzést kapunk-e. Előfordulhatnak azonban pszeudotoldalékok is a kötőjel után (például *Ray-Ban*). Ezek nagy része morfofonológiai és hangrendi összeférhetetlenségi szabályok alapján kiszűrhető.

A kötőjeles esetek vizsgálatával a Szeged Korpuszon 1085 esetben kapunk helyes elemzést (az ismeretlen szavak 3,17%-a).

5 Szófaji egyértelműsítés és többszavas kifejezések

A szófaji egyértelműsítés kapcsán egy érdekes kérdés, hogy mi az elvárt elemzése a többszavas kifejezéseknek. Szemléletes példa a *Magyar Nemzeti Bank* frázis, amely egy darab főnévként szerepel a korpuszban, Np-sn MSD-kóddal. Ha az ilyen és ehhez hasonló kifejezések szavait külön-külön vizsgálnánk, akkor a frázis minden egyes szavához tartozna egy-egy lemma és az ahhoz tartozó szófaji kód. E példát vizsgálva a *Magyar* és a *Nemzeti* szavakra egyaránt melléknévi elemzést kapnánk (Afp-sn), míg a *Bank* egy főnévi (köznév, Nc-sn) szófaji kóddal lenne ellátva. A jelenlegi nyelvi elemző megoldások azt a stratégiát követik, hogy első lépésben minden tokenre meghatározzák annak morfológiai elemzését (a POS-tagger kimenete), majd egy későbbi (általában független) lépés feladata a frázisok azonosítása.

Mivel korábbi névelem-felismerési kísérleteinkből [2] azt tapasztaltuk, hogy a szófaji kódok hozzáadott információtartalma a névelem-felismeréshez elhanyagolhatóan kicsi, ezért egy újszerű megközelítést javaslunk: első lépésben egy modul vonja össze a kifejezéseket, majd ezeken végezzük el a szófaji egyértelműsítést. Ily módon a *Magyar Nemzeti Bank* kifejezésről mint egyetlen egységről kell döntést hoznia egy szófaji egyértelműsítőnek, ami intuitíve kézenfekvőbbnek látszik (ez a szintaktikai egység ugyanúgy viselkedik, mint bármely más főnév).

5.1 Frázishatárok azonosítása

Megvizsgáltuk, hogy ha a nyelvi elemző első lépésben meghatározza a frázisokat, majd ezeken hajtja végre a szófaji egyértelműsítést, jobb eredményeket érhetünk-e el, mint a hagyományos megközelítéssel.

A frázisok azonosításához szekvenciális tanulást (CRF, Conditional Random Fields [6]) használtunk. A rendszer a Szeged Korpuszban jelölt frázisokon (olyan termek, amelyek tartalmaznak szóközt) tanult[9]. A frázisok esetünkben a több tokenből álló tulajdonnevek, de a módszer tetszőlegesen kiterjeszthető (a tanító adatbázis módosításával), bármely, egy logikai egységet alkotó tokensorozat összevonására, mint például mennyiségek (*3 millió Ft*) vagy dátumok (*2012. december 21.*).

A frázishatár-jelölő tanuló algoritmus egyszerű jellemzők halmazát (kb. 100 ezer dimenzió) használta fel. A felhasznált jellemzőcsoportok az alábbiak voltak (részletesen l. [2]):

- felszíni jellemzők (a szóalakra mint betűsorozatra vonatkozó információk)
- környezeti jellemzők
- gyakorisági adatok

[9] A Szeged Korpusz 2.0-ban a több tokenből álló tulajdonnevek egyetlen tokenként vannak jelölve, és a lehetséges morfológiai kódok és lemmák is frázisszinten lettek meghatározva.

- tulajdonnévszótárak
- egyértelmű tulajdonnevek listája

Ezen egyszerű jegyeknek felhasználásával már 90% körüli pontosságú eredmény érhető el. Az így kapott modell segítségével ismeretlen (korábban nem látott) szövegekből tudjuk detektálni az összevonandó frázisokat.

5.2 Szófaji egyértelműsítés a frázisokon

Ha már ismertük az összevonandó frázisokat, minden frázist lecseréltünk annak utolsó szavára, tehát a *Magyar Nemzeti Banknak*-ot egyszerűen *Banknak*-ra cseréltük. Ezt követte a szófaji egyértelműsítés és a lemmák meghatározása.

Vegyük az alábbi példamondatot: *Levélben fordult az Országos Magyar Méhészeti Egyesülethez*. Egy egyszerű elemzés során az eredmény: `[levél/N, fordul/V, az/Tf, országos/A, magyar/A, méhészeti/A, egyesület/N, ./.]` lenne, melyben ugyan ha külön-külön vesszük a szavakat, akkor valóban helyes az elemzés, de a valamely szervezetre utaló jelentéstartalom teljesen elvész.

A fent ismertetett módszer alapján, ha sikerült helyesen felismerni fráziskénti az *Országos Magyar Méhészeti Egyesülethez* tokensorozatot, akkor az elemzés eredménye: `[levél/N, fordul/V, az/Tf, Országos Magyar Méhészeti Egyesület/N, ./.]`, ahol a szervezetre való utalás nem vész el, illetve a szervezetet jelölő valamennyi token egy egységet alkot, és főnévi kóddal kerül az elemzés eredményébe.

5.3 Kiértékelési módszertan

Ahhoz, hogy a standard megközelítéssel összevethető legyen a módszer, először automatikusan lemmatizáltuk a Szeged Korpuszt (`magyarlanc` felhasználásával), és a szótöveken tanítottunk egy frázishatár-felismerő CRF rendszert, minden egyéb paraméterében a korábban bemutatott módszerrel megegyező módon. Az így – immár lemmákon – tanult modell lemmatizált szövegek frázishatárainak meghatározására lesz alkalmas.

Ebben a megközelítésben tehát először szófajilag egyértelműsítjük a mondatokat, majd ennek eredményét felhasználva célozzuk meg a frázisok azonosítását (intuitíve a szótárakon alapuló frázishatár-felismerőnek jobban kell teljesítenie a szótövek ismeretében). Az előző példa szerint, ha a rendszernek sikerül detektálnia az `országos/A, magyar/A, méhészeti/A, egyesület/N` lemmasorozatot mint egy négy szóból álló frázist, akkor a tokensorozat a második lépésben összevonódik, így az a későbbiekben egy frázist fog alkotni. A frázis utolsó szava lemmatizált formában fog szerepelni a frázisban, a többi token azonban az eredeti formában kerül be, szófaji kódként pedig a frázis utolsó tokenjének szófaji kódja kerül az elemzésbe: `Országos Magyar Méhészeti Egyesület/N`.

Tehát a nyelvi elemzés kimeneteként mindkét módszernél szófajilag elemzett és frázishatárokkal annotált mondatot várunk el. A szófaji egyértelműsítőt és a frázisha-

tár-felismerőt is a Szeged Korpusz egy véletlenül választott 80%-án tanítottuk, majd a kiértékelést a maradék 20%-on végeztük el. A kétfajta megközelítést két különböző módon értékeltük ki. Az egyik esetben a névelem-felismerésben használatos frázisszintű pontosság/fedés/F-mértéket számoltuk ki. Ebben az esetben ha egy frázishatár nem jól lett meghatározva vagy annak típusa nem egyezett, azt mind hibás illesztésnek tekintettük. A másik kiértékelés tokenalapon történt, itt az egy egységként azonosított (és szófajilag egyértelműsített) többszavas frázisokat tokenekre bontottuk, és minden token a frázis szófaji kódját kapta meg (ezt a szétbontást az etalon és a predikált halmazon is végrehajtottuk).

5.4 Eredmények

Az alábbi táblázat tartalmazza a kétfajta frázishatár- és szófaji egyértelműsítő módszer eredményeit, valamint a 4. fejezetben tárgyalt utófeldolgozási lépések hozzáadott értékét.

2. táblázat: Szófaji egyértelműsítő rendszerek eredményei.

		frázisalapú kiértékelés P/R/F	tokenalapú kiértékelés P/R/F
1. POS-tagger 2. frázishatár	N	83.50/92.72/87.87	90.41/95.45/92.87
	A	93.92/89.66/91.74	94.04/89.67/91.81
	összesen	88.40/89.61/89.01	90.93/90.64/90.79
1. frázishatár 2. POS-tagger	N	89.00/95.07/91.93	90.49/95.75/93.04
	A	95.07/89.58/92.24	95.11/89.58/92.26
	összesen	90.38/90.27/90.33	91.06/90.79/90.93
1. frázishatár 2. POS-tagger +utófeldolgozás	N	88.96/95.19/91.97	90.50/96.04/93.19
	A	95.05/89.61/92.25	95.10/89.62/92.28
	összesen	92.25/90.31/90.36	91.15/90.88/91.02

Az eredmények alapján mindkét kiértékelő módszer szerint a frázishatárok előzetes detektálása, majd a frázisok egy egységként történő kezelése szignifikánsan jobb eredményt ér el (McNemar-teszt alapján), mint a klasszikus megközelítés. Ez elsősorban a főnevek és melléknevek pontosságának javulásának köszönhető, ami arra enged következtetni, hogy a frázisösszevonásokkal sok tévesen főnévnek/melléknévnek jelölt tokent javítani tudunk (például a *Magyar Nemzeti Ban*k esetében a két melléknévi token helyett – ha a frázishatárokat sikerül azonosítani és a frázist főnévnek jelölni – két főnévi jelölésünk lesz).

Az ismeretlen szavak elemzésére adott utófeldolgozási megoldásaink hozzáadott értéke a végső rendszerhez a tokenalapú kiértékelés alapján szignifikáns. A főnevek és a melléknevek esetén ezek alkalmazásával a fedés nő, míg a pontosság tulajdonképpen nem változik. Előbbi természetesen annak a következménye, hogy több főnevet és melléknevet azonosítunk utófeldolgozással, mint a nélkül.

6 Konklúzió

Ebben a munkában bemutattuk a `magyarlanc` nyelvi elemző rendszert. Ennek jellegzetességei, hogy JAVA nyelven implementálódott, szabadon hozzáférhető, MSD-kód és relatív szótő alapú, számos utófeldolgozási lépést tartalmaz ismeretlen szavak kezelésére, a frázishatárok felismerését is elvégzi (méghozzá a szófaji egyértelműsítés előtt).

A végső rendszer a klasszikus szófaji egyértelműsítő modulnál 1,3%-kal jobb F-mértéket ér el a Szeged Korpuszon.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD). LNAI series Vol. 3658 (2005) 123–-131
2. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
3. Farkas R., Szeredi D., Varga D., Vincze V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 354–357
4. Halácsy P., Kornai A., Oravecz Cs.: HunPos — an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007) 209–-212
5. Kuba A., Bakota T., Hócza A., Oravecz Cs.: A magyar nyelv néhány szófaji elemzőjének összevetése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 16–22
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)
7. Novák A., Nagy V., Oravecz Cs.: Magyar ismeretlenszó-elemző program fejlesztése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 45–54
8. Trón V., Halácsy P., Rebrus P., Rung A., Vajda P., Simon E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation (2006)

Obi-ugor morfológiai elemzők és korpuszok

Fejes László[1], Novák Attila[2]

[1]MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr utca 33.
fejes@nytud.hu
[2]MorphoLogic
1116 Budapest, Kardhegy utca 5.
novak@morphologic.hu

Kivonat: Cikkünkben a végéhez közeledő OTKA NF 71707 projekt keretein belül létrehozott obi-ugor számítógépes morfológiákat, annotált korpuszokat, a használatukat lehetővé tevő webfelületet és azokat a problémákat mutatjuk be, amelyek a fejlesztés során felmerültek.

1 Bevezetés

A kisebb uráli nyelvek veszélyeztetettek, ezért dokumentálásuk nemzetközi jelentőségű feladat. A magyarországi uralisztika ezen a területen jelentős hagyományokkal rendelkezik: a 19. század közepétől kezdve magyar kutatók rendszeresen gyűjtöttek szövegeket, szótári anyagokat, és ezek alapján készítettek grammatikai vázlatokat is. Végéhez közeledő projektünkben (OTKA NF 71707) a korábban gyűjtött obi-ugor szövegek számítógépes feldolgozásával morfológiailag annotált korpuszokat hoztunk létre.

A projekt a két obi-ugor nyelv három nyelvjárását öleli fel, és az alábbi négy fő modulra oszlik:

Vogul (manysi) északi nyelvjárás: Kálmán Béla gyűjtése (WT) [6]
Vogul (manysi) északi nyelvjárás: Munkácsi Bernát gyűjtése (VNGY) [7]
Osztják (hanti) szinjai nyelvjárás: Ruttkay-Miklián Eszter gyűjtése
Osztják (hanti) kazimi nyelvjárás: különböző gyűjtések [12, 15, 14]

A modulokban egy-egy gyűjtés, illetve nyelvjárás feldolgozására vállalkoztunk. Ezt az indokolja, hogy a számítógépes elemzés megköveteli a lehető legegységesebb korpuszok használatát: a sokszínű korpuszokhoz megengedőbb elemzőt kellene építeni, ami viszont óhatatlanul a téves elemzések megszaporodásával járna együtt. Éppen ezért minden egyes tér- és időbeli nyelvváltozathoz önálló elemzőt építettünk.

Az elkészült elemzők és korpuszok egy része már online hozzáférhető, és folyamatosan tesszük közzé az újabb elkészült erőforrásokat [16].

2 Az elemzők építése

A hanti nyelvjárások közötti igen jelentős különbségek miatt a két hanti elemzőt egymástól függetlenül, az alapoktól építettük fel. A két manysi gyűjtés esetében ugyanazon nyelvjárás két időben eltérő nyelvállapotát két igen eltérő transzkripcióval rögzítették: ez indokolta, hogy itt is két külön morfológiát hoztunk létre. A Kálmán jelen projekt keretében feldolgozott szövegeihez készült elemző esetében támaszkodhattunk egy korábbi projekt keretében Kálmán által máshol [5] publikált szövegekhez készült elemzőnkre. Munkácsi szövegei esetében azonban ismét az alapoktól kellett kezdenünk a munkát.

A manysi elemzők tőtárát az adott kiadványokhoz készült szójegyzék [6], illetve szótár [9] alapján készítettük el. A hanti tőtárak alapjául elsősorban Steinitz szótára [13] szolgált. A szövegek feldolgozása során az egyik legfőbb problémát a szövegek belső inkonzisztenciája és a szótárak pontatlansága okozta. A másik probléma a nyelvtanok ([5, 6, 8, 12, 13]) vázlatszerűsége és felületessége volt: ezek ritkán adtak elég támpontot a morfofonológiai jelenségeknek a számítógépes implementációhoz szükséges pontos leírásához. Cikkünkben bővebben kitérünk néhány olyan nyelvtani problémára, amelyek megoldása jelentős kihívást jelentett.

A tőtárak és a szövegek nagy részének digitalizálása begépelés útján történt, a Munkácsi–Kálmán szótárat [9] pedig (ez a Munkácsi által gyűjtött és publikált szövegek szóanyagát fedi le) OCR-rel digitalizáltuk. A Munkácsi–Kálmán szótárban alkalmazott manysi átírás számtalan szokatlan karaktert tartalmaz (magánhangzóbetűket több különböző ékezettel, felső indexben álló gammákat stb.), ezért az OCR programot egyedileg kellett betanítani a feladatra. Ráadásul a szótárban szereplő dőlt betűs cirill karakterek egy része (*a, c, e, m, n, o, p, x, y*) megkülönböztethetetlen a manysi címszavakban álló dőlt betűs latin karakterektől, ezért ennek a megkülönböztetésnek a felismerését nem bíztuk az OCR programra, hanem az összes ilyen karaktert cirillként ismertettük fel a programmal, és utólag automatikusan konvertáltuk manysi részekben álló karaktereket. Konverzió után az OCR-hibákat kézzel javítottuk. A szótárban a tipográfia alapján programmal azonosítottuk a címszavakat és a magyar, német, illetve helyenként orosz nyelvű fordításokat, a nyelvjárásra vonatkozó adatokat, így képezte a szótár a Munkácsi-szövegek feldolgozására készülő manysi elemző tőtárának alapját. A szótár és Munkácsi szövegkiadásai más manysi nyelvjárások szóanyagát is tartalmazzák. Jelen projektben azonban csak a legbővebben adatolt északi nyelvjárás feldolgozására vállalkoztunk.

3 A morfológiai elemzők jellemzői

A projekt keretében elkészült morfológiai elemzők mindegyike a MorphoLogic *Humor* elemzőmotorjára épül. A morfológiai adatbázisok létrehozására a korábban már számos más nyelv (elsőként a magyar) számítógépes morfológiájának létrehozásához használt morfológiaiadatbázis-leíró keretrendszert használtuk ([10, 11]). A Humor elemző morfémaallomorfok felszíni alakjainak egy véges állapotú automata által leírt szónyelvtannak és a lokális szomszédossági megszorításoknak is megfelelő sorozatait

ismeri fel a bemenetén kapott szóalakban, és az ezeknek megfelelő morfémasorozatokat jeleníti meg elemzésként. A rendszert kiegészítettük egy olyan mechanizmussal, amely az eredetileg a morfológia forrástőtárában tárolt különböző nyelvű glosszákat is az elemzésekhez csatolja, így a rendszer egyben szemantikai címkézést is végez. Az elemzőkhöz készített webes felületen így az elemzések magyar és angol (illetve a manysi elemzők esetében emellett még német) nyelvű glosszákkal együtt jelennek meg. Ez lehetővé teszi, hogy a szövegeket a nyelvet nem beszélő kutatók is értelmezni, illetve egyértelműsíteni tudják.

A keretrendszerben a tövek és a toldalékok leírására különböző formalizmus szolgál, de mindkettőben általában csak morfémák és megjósolhatatlan lexikai jegyek redundanciamentes leírása szerepel. Az elemző által használt allomorfokat és a morfok szomszédossági megszorításait leíró teljes jegyegyütteseket az elemző lexikonának kompilálásakor a keretrendszer állítja elő a morfológia forrásának részét képező szabályrendszer felhasználásával. A forráslexikonban allomorfok, illetve toldalékolt alakok csak akkor szerepelnek, ha olyan mértékben rendhagyóak, hogy szabállyal való előállításuknak nem láttuk értelmét.

A jelen projekt keretében feldolgozott nyelvek és korpuszok esetében azonban jóval gyakoribb eset volt, hogy a lexikonba allomorfokat, írásváltozatokat kellett felvennünk, mint például a sztenderd mai magyar szövegek elemzésére készített elemzőnk esetében, mert itt nagyságrendekkel több a lejegyzési következetlenség, illetve a nyelvek kevésbé sztenderdizált voltából adódóan is jóval nagyobb a változatosság. Az alábbi táblázat ezt szemlélteti.

elemző	lexikálisan megadott allomorffal vagy toldalékolt alakkal rendelkező tövek aránya	
mai magyar:	274/139859	0.20%
manysi WT	475/4209	11.29%
manysi VNGY	3705/16526	22.42%
szinjai hanti	314/2606	12.05%
kazimi hanti	301/1958	15.37%

A következő táblázatban összefoglaltuk az egyes nyelveken, nyelvváltozatokon rendelkezésünkre álló, illetve feldolgozott korpuszok és az elkészült elemzők mennyiségi jellemzőit.

Nyelv	korpusz	tőlexikon				toldaléklexikon	
	szó	lemma (*jelentés)		allomorf		mögöttes alak	allomorf(sorozat)
		zárt	nyílt	zárt	nyílt		
manysi WT	10659	387	3822	622	5483	376	5285
manysi VNGY	81717 (1026)	909	15617	1900	34665	297	2944
szinjai hanti	151500 (6539)	256	2350	615	7894	140	813
kazimi hanti	19228	209	1749	689	6756	150	1491

A táblázatban külön oszlopban soroltuk fel a nyílt (főnév, melléknév, ige, határozószó) és a zárt szófajosztályokba (többi szófaji kategória) tartozó tövek számát. A tövek elkülönült jelentései külön tételként jelennek meg a tőtárakban. A táblázatból kitűnik, hogy a hanti elemzők esetében az egyes morfémáknak átlagosan több mint 3 allomorfja van, ami az alább részletezett szótagszerkezeti megszorításokból, az azok megvalósítására a beszélők által alkalmazott stratégiák változatosságából, valamint a lejegyzésekben tapasztalható ingadozásából adódik. A megadott korpuszméreteknél néhol szereplő zárójeles szám egy olyan alaposabban ellenőrzött részkorpusz méretére utal, amelyeken belül igyekeztünk minden lejegyzési hibát kijavítani, és az elemző által teljes lefedést biztosítani.

A toldaléklexikonokban toldalékkapcsolatok is szerepelnek, illetve az inflexióstoldalék-sorozatok nagy részét a keretrendszer offline kigenerálja, így az elemző gyorsabban működik, mert a teljes sorozatot egy lépésben találja meg elemzéskor a lexikonában. Ebből adódik a mögöttes toldalékok és az elemző kigenerált allomorflexikonjának mérete közötti sokszoros különbség.

4 A morfológiai elemzők jelentősége

A morfológiai elemzők használatával előállítható, morfológiailag annotált korpuszok jelentőségéről itt nem kívánunk szólni, ezek haszna minden szakmabeli számára nyilvánvaló. Azt azonban jeleznünk kell, hogy a projekt sajnos még nem foglalta magában egy komplex korpuszkezelő fejlesztését, így az ilyesféle lehetőségek – például kifinomult keresőrendszer hiányában – korlátozottak.

Fontosnak érezzük azonban szólni a morfológiai elemző fejlesztése során nyert tapasztalatok jelentőségéről.

Az obi-ugor nyelvek kutatásának lehetőségei – bár ma is élő nyelvekről van szó – nagyjából a holt nyelvek kutatásának lehetőségeihez hasonlíthatók. Élő nyelvhez hasonlóan csak az éppen terepen levő nyelvész kutathatja, ilyen jellegű munkára azonban ritkán nyílik alkalom, s mivel a terepmunkás is tisztában van az alkalom különleges voltával, idejét leginkább nyelvi anyag (szövegek) rögzítésére fordítja. Maga a nyelvészeti kutatás elsősorban ezekre a szövegekre épül, azaz az obi-ugor nyelvészet szorosan összefonódik az obi-ugor filológiával. Mivel egy-egy nyelvjárásról, illetve annak időbeli állapotáról mindig igen korlátozott adatunk van, és az egyik nyelvjárásban vagy állapotban megfigyelt szabályszerűségeket nem vetíthetjük át automatikusan más nyelvjárásokra és állapotokra, az adatok kezelése nagy óvatosságot és pontosságot igényel.

A számítástechnika előtti korszakban az adatok gyűjtése, kezelése, feldolgozása rengeteg hibalehetőséget rejtett magában. Nem csupán az adatok rögzítésekor kerülhetett hiba a rendszerbe, az adatokat is kézzel másolták, a sajátos jelek kezelése a nyomda számára is nehézséget jelentett. A hibákat nehéz volt kiszűrni, hiszen a kiadott szövegekben, a szótárakban és a nyelvtanokban szereplő adatok többé nem „találkoztak” egymással. Egy lexikai jellegű tanulmány már nyilvánvalóan a szótárra épült, nem ment vissza a szövegekhez. Azok a hibák, melyek a szövegek feldolgozásakor és a szótár készítésekor keletkeztek, torzították a nyelvről alkotott képet.

A számítógépes morfológiai elemzők nagy előnye, hogy a korpuszban és a tőtárakban levő adatok, illetve az explicit módon, képletszerűen megfogalmazott morfofonetikai és morfológiai szabályok interakcióban vannak egymással, a közöttük levő ellentmondások az esetek nagy részében szükségszerűen nyilvánvalóvá válnak.

Az általunk épített manysi elemzőkben mindig az adott szövegkorpuszhoz kiadott szójegyzékeket, illetve szótárat használtuk. Mindhárom esetben kiderült, hogy a szójegyzékek, illetve a szótár hibásak, illetve hiányosak. A szavak nem ugyanabban az alakban szerepelnek a szótárban, mint a szövegekben (jellemző például a magánhangzók hosszúságának eltérő jelölése, de gyakori a puszta helyesírási következetlenség, pl. a kötőjel használatában való ingadozás is), vagy nem szerepel a szövegben előforduló összes alakváltozat. Egyes szavak teljesen hiányoznak, különösen gyakori ez a képzett szavak esetében (olyanoknál, melyek alapszava szótározva van), illetve a tulajdonneveknél. Vannak esetek, amikor a szótár szerint a szó nem dokumentált az általunk vizsgált északi nyelvjárásban, szövegeinkben azonban mégis szerepel. Az összetett szavak szótározása is meglehetősen rapszodikus: egyes transzparens összetételek szerepelnek a szótárban, miközben sajátos jelentésű összetételek hiányoznak. Az elemzők fejlesztése során véletlenül bukkanunk olyan esetekre, amikor a szó ugyan szótározva van, de nem minden, a szövegekben dokumentált jelentésében. Az ilyen esetek módszeres felderítésére majd a teljes korpuszok egyértelműsítése fog lehetőséget teremteni.

A hanti korpuszok esetében a feldolgozott szövegekhez nem készültek szójegyzékek, ezeket mi magunk hozzuk létre. A Steinitz-féle szótárral [13] való egybevetés ugyan fontos szerepet játszott, de mindkét korpuszunkban jócskán találtunk olyan töveket, melyek Steinitznél nem, vagy más alakban szerepeltek.

Pusztán az a tény, hogy a szövegek digitalizálva vannak, lehetőséget teremt a lejegyzés egyenetlenségeinek korrigálására. Így például az alakváltozatok megjelenésének aránya utalhat arra, hogy mikor lehet szó valódi alakváltozatokról, és mikor valószínűbb, hogy egyes írott „alakváltozatok” csupán sajtóhiba eredményei. A szöveg feldolgozásának később stádiumában más lejegyzési egyenetlenségek kiküszöbölésére is sor kerülhet, így például a hol külön, hol összetett szóként leírt szószekvenciák lejegyzése egységesíthető. A szövegek digitalizálásának köszönhető, hogy felfedeztük: a Munkácsi–Kálmán szótárban [9] olyan szóalakok is szerepelnek, amelyek a szövegben [7] nem – ezek feltehetően Munkácsi kéziratos cédulaanyagából kerültek a szótárba. Ennél azonban sokkal érdekesebb, hogy a szótárban olyan példamondatok is vannak, melyek a kiadott szövegekben nem lelhetők fel. Ennek alapján azt gyanítjuk, hogy Munkácsi cédulái jelentős korpuszt, ha nem is szövegeket, de elszigetelt mondatokat tartalmaz. Okkal feltételezhetjük, hogy ezen példamondatoknak töredéke került csak be a szótárba. Sikerült tehát (újra)felfedeznünk egy olyan 19. századi manysi forrásanyagot, mely időközben kiesett a kutatás látóköréből, és a morfológiai elemzés fejlesztése nélkül talán örökké „elveszett” volna. A cédulaanyag ilyen típusú feldolgozására egy további projekt folyamán kerülhet sor, mindenesetre ezt is feladataink között tarjuk számon.

A morfológiai elemző építése során erősen támaszkodtunk a szóban forgó nyelvjárásokat leíró nyelvtani vázlatokra. Ezek – érthető módon – nem olyan egzakt leírásokat tartalmaznak, melyek azonnal alkalmasak szabályokba kódolásra, de mindenesetre jó kiindulópontul szolgálnak. A morfofonológiai váltakozások közül az obi-ugor

nyelvekben a legjelentősebbek a jól formált szótagok építését célzó szabályok. Ezt minden nyelvváltozatban vegyes stratégiával érik el: részben mássalhangzók törlésével, részben magánhangzók (elsősorban svá) betoldásával. A helyzetet nagyban bonyolítja, hogy a szonoránsok előtti svábetoldás helyett gyakran a szonoráns válik szótagalkotóvá – legalábbis a lejegyzésben ez szerepel. Vannak azonban helyzetek, amikor a lejegyzés sem a svá betoldását, sem a szonoráns szótagalkotóvá válását nem jelzi. Kezdetben azt feltételeztük, hogy ezekben az esetekben egyszerűen a lejegyzés pontatlanságáról van szó.

Későbbi megfigyeléseink azonban ezt megkérdőjelezik. Nem egy esetben svá előtt a tőnek az az alakváltozata jelenik meg, amelynek szabályszerűen magánhangzóval kezdődő toldalék előtt nem lenne szabad megjelennie. Úgy tűnik, a rosszul formált szótagszerkezet kiküszöbölésére két szabály is aktivizálódik, holott az egyik bőven elegendő lenne. Ennek megfelelően az elemzőben azokat a toldalékokat, amelyek szonoránssal kezdődnek, akár svá-betoldásos alakjukban, akár svá nélkül jelennek meg, sem magán-, sem mássalhangzós kezdetűként nem jelöljük meg, így mindkét tőalakváltozathoz kapcsolódhatnak.

Más esetekben viszont nem toldódik be svá, a tőnek mégis az az alakváltozata jelenik meg, amelyet csak magánhangzós toldalékok előtt várnánk. Nehéz eldönteni, hogy ilyen esetekben nem egyszerű sajtóhibáról van-e szó. Amióta azonban felfigyeltünk a problémára, több független forrást is felfedeztünk, melyek azt a benyomásunkat erősítik meg, hogy ez igenis előfordulhat. Pillanatnyilag azt a megoldást követjük, hogy a mássalhangzóval kezdődő toldalékok előtti mássalhangzókapcsolat-egyszerűsödések fakultatívak: az elemzéskor nem várjuk el a svá-betoldást, ám a szóalak-generátor a svát mindig betoldja.

Elképzelhető azonban, hogy szabályaink túlságosan megengedőek. Előfordulhat például, hogy az általunk homonimként kezelt toldalékok a morfofonológiai váltakozásokban eltérő viselkedést mutatnak. Ezt azonban csak a kutatás egy későbbi szakaszában, az egyértelműsítés elvégzése után lehet vizsgálni: az, hogy valójában melyik morfémának milyen allomorfjai jelenhetnek meg a különböző környezetekben, csak a már egyértelműsített szövegeken vizsgálható. Ám ekkor sem lesz könnyű elkülöníteni a sajtóhibákat a valódi alternánsoktól.

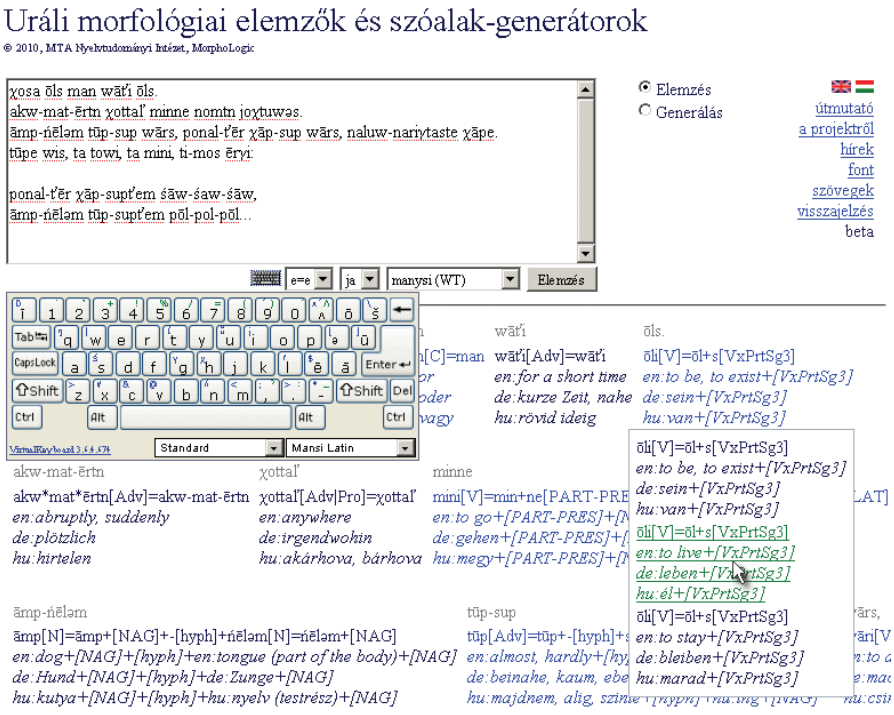
Az első obi-ugor elemző készítése során elsősorban a nem első szótagban található magánhangzók minősége, illetve a svá betoldása és be nem toldása kapcsán vetett fel kérdéseket. A problémák megoldása céljából több kutatás indult el, köztük akusztikai vizsgálatok is: ezekről több előadás és cikk is született ([1, 2, 3, 4]). A jelenlegi problémák inkább fonológiaelméleti kérdéseket állítanak a központba: hogyan lehetséges az, hogy miközben egy nyelv radikális váltakozásokat vezet be a rosszul formált szótagok kiküszöbölésére, ezzel egy időben nagyfokú toleranciát is mutat ezen rosszul formált alakokkal szemben. E kérdéssel kapcsolatban is újabb tanulmányok sora várható.

5 Online morfológiák

A projekt keretében készült morfológiák és a korpuszt alkotó szövegek a projekt végére webes felületen keresztül válnak elérhetővé [16]. Az elemzők esetében a kivá-

lasztott szöveget a megfelelő ablakba másolva a felhasználó megkapja a szövegben szereplő szavak lehetséges morfológiai elemzéseit és az elemzésekben szereplő tőmorfémák jelentését. Virtuális billentyűzet segítségével maga is gépelhet be szövegeket. Az elemzéseket megjelenítő webes felület egyben kézi egyértelműsítő eszközként is szolgál: a többértelmű szavak elemzései pop-up ablakban jelennek meg, ha az egeret egy többértelmű szó fölé mozgatjuk, ezek közül egérrel választhatunk. Az elkészült elemzések, illetve azok egyértelműsített változata elmenthető, az elmentett változatot a böngészőbe betöltve, az esetlegesen félbehagyott egyértelműsítő munka később folytatható.

A webes felületen keresztül nemcsak morfológiai elemzők, hanem szóalak-generátorok is elérhetők az egyes nyelvekhez. Az alábbi képernyőképek illusztrálják a szövegbeírás, a virtuális billentyűzet, az egyértelműsítő felület és a szóalak-generátor használatát. Ha egy adott morfémasorozat több formában is megjelenhet, akkor a generátor kimenete az elemző többértelmű kimenetének megjelenítéséhez hasonlóan jelenik meg a webes felületen, a lehetséges szóalakváltozatok itt is az egérmutatót a generált szóalak fölé mozgatva megjelenő pop-up ablakban láthatóak.



Bibliográfia

1. Bakró-Nagy M., Fejes L.: Schwa or not schwa? Synchronic and diachronic speculations on an Ob-Ugric vowel. FUSAC, Vancouver. 2008. június 8.
2. Fejes L.: A vogul morfológiai elemző(k) felé. Fonológiai és morfológiai megfigyelések. Obi-ugorok a 21. században (CD-ROM). MTA Nyelvtudományi Intézet, Budapest (2006) http://fgroszt.nytud.hu/publikaciok/obi-ugorok/text/nyelv2.html
3. Fejes L.: Az északi-manysi vokalizmus néhány kérdése. MTA Nyelvtudományi Intézet, 2008. május 8. http://nytud.hu/~fejes/pdf/manysiV/manysi_V-k_ea.pdf
4. Fejes L.: On the acoustics of the Northern Mansi Vowel System. Poszterelőadás a 17. Manchesteri Fonológiai Találkozón. 2009. május 29. http://fgrtort.nytud.hu/images/stories/fejes/fejes_manchester_poster.pdf
5. Kálmán B.: Chrestomathia Vogulica. Tankönyvkiadó, Budapest (1989)
6. Kálmán B.: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest (1976)
7. Munkácsi B.: Vogul népköltési gyűjtemény. 1–4. Budapest (1892–1921)
8. Munkácsi B.: A vogul nyelvjárások szóragozásukban ismertetve. Budapest (1894)
9. Munkácsi B., Kálmán B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)
10. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, Szeged (2003) 138–145
11. Prószéky G., Novák A.: Computational Morphologies for Small Uralic Languages. In: Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H,. Yli-Jyrä, A. (szerk.): Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford (2005) 116–125
12. Rédei K.: Nord-ostjakische Texte (Kazym Dialekt) mit Skizze der Grammatik. Vandenhoeck and Ruprecht, Göttingen (1968)
13. Steinitz, W.: Dialektologisches und Etymologisches Wörterbuch des Ostjakischen Sprache. Akademie-Verlag, Berlin (1966)
14. Steinitz, W.: Ostjakologische Arbeiten. Beiträge zur Sprachwissenschaft und Ethnographie. Herausgegeben von Gert Sauer und Renate Steinitz. Bd. I–IV. Akadémiai Kiadó – Akademie-Verlag, Budapest – Berlin (1980)
15. Хомляк, Л. Р. (ред.): Арєм-моньщєм ел ки мăнл... (Если моя сказка-песня дальше пойдёт...) Фольклорное творчество Пелагеи Алексеевны Гришкиной из деревни Тугияны. ГРУПП »Полиграфист«, Ханты-Мансийск (2002)
16. http://www.morphologic.hu/urali/index.php

A magyar frazeológiai adatbázis létrehozása és az ebből generált szinonim frazémaszótár munkálatai

Bárdosi Vilmos[1], Kiss Gábor[2]

[1] ELTE BTK, Francia Tanszék, 1068 Budapest, Múzeum krt. 4-6.
bardosi.vilmos@btk.elte.hu
[2] TINTA Könyvkiadó, 1116 Budapest, Kondorosi út 17.
kissgabo@tintakiado.hu

Kivonat: A munka során létrehozott magyar frazeológiai adatbázis lineáris elrendezésű frazémagyűjtemény, a definícióikkal együtt tartalmazza a magyar nyelv leggyakoribb kollokációit, szólásait, közmondásait és helyzetmondatait. E gyűjteményből a definíciók elemzésével, kulcsszavak felhasználásával gépi transzformációval állítottuk elő a frazémák szinonimikus elrendezését. A transzformáció során első lépésként a frazémadefiníciókat tanulmányoztuk, és definiáltunk 850 kulcsfogalomsort, melyek mindegyike 2-8 szinonim szó alkotott. Az egyes kulcsfogalomsorok alá a frazémák besorolását úgy végeztük, hogy a kulcsfogalomsor szinonim szavait illesztettük az egyes frazémák definíciójára. Az automatikus elrendezést manuálisan javítottuk és finomítottuk (pl. az aspektusbeli jellemzők figyelembe vételével).

1 Bevezetés

Korábbi cikkünkben bemutattuk a magyar szavak jelentéshasonlóságának automatikus feltérképezési lehetőségeit [3]. Jelen cikkünkben frazémák hasonló osztályozására tett kísérletünket vázoljuk fel. Az automatikus szövegfeldolgozás nem nélkülözheti a szavaknál nagyobb lexikai egységek, a frazémák helyes kezelését. Frazéma gyűjtőfogalom alatt értünk minden olyan többelemű szósort, melynek szavai együttesen mást jelentenek, mint várnánk a szavak puszta kompozíciójából. A régebbi magyar szakirodalom terminológiája nem volt egységes, számos elnevezés volt használatos: *közmondás, példabeszéd, közbeszéd, közmondat, közszólás, közhasonlat, közpéldabeszéd, közpéldaszó, példaszó, példázat, aggszó, rólabeszéd, szólásmód, szójárás, szólásmondás, szokásmondás, mondás*, ahogy erre Tolnai Vilmos felhívja figyelmünket [10]. Voigt Vilmos A magyar folklór című könyvben a Kisepikai műfajok között tárgyalja a szavaknál nagyobb nyelvi egységeket, és a néprajzi szakirodalomban bevett proverbium terminust használja [11].

Az ezredforduló után szerencsés módon szinte egy időben több magyar frazémaszótár is napvilágot látott [2, 4, 5, 9, 8]. A puszta leltározást többirányú feldolgozásnak kell követnie. Az egyik lehetséges feldolgozás a magyar frazémakincs szinonimikus elrendezése. Erre a feladatra Tolnai Vilmos már 1935-ben felhívta a figyelmet ezekkel a szavakkal: „A szóláskutatás feladata … a hasonlóértelmű szólá-

sok csoportosítása, a szólás-szinonimika, melyben nem az alak, hanem az alkalmazás jelentése az osztályozás alapja, pl.: *halálos beteg: nem sokáig viszi, hálni jár belé a lélek, nem lesz többé ember belőle, ütött az utolsó órája.*'' [10]. A szerző még a frazémák fogalomköri elrendezését is sürgeti ezekkel a szavakkal: „Feladatunk a fogalmi körök szerint való csoportosítás. Pl.: szólások háziállatainkról; az év napjaihoz fűződő mondások; az időjárás jóslatai; a művelődéstörténet tárgyi csoportjai.'' [01]. Az ilyen jellegű hiányt az elméleti háttér feltárása után [1] pótolták a Bárdosi Vilmos szótáraiban található fogalomköri mutatók [2, 4].

2 A frazéma-adatbázis létrehozása és jellemzője

A magyar frazeológiai adatbázis több mint 25.000 frazémát tartalmaz, két nagy gyűjtemény [2, 4, 7] frazémáinak összefésülésével hoztuk létre. E lineáris elrendezésű gyűjtemény a definícióikkal együtt tartalmazza a magyar nyelv leggyakoribb kollokációit, szólásait, közmondásait és helyzetmondatait.

A magyar frazéma-adatbázis szerkezete

1. a frazéma; **2.** a frazéma típusa [egyszavas ekvivalenssel rendelkező szókapcsolat | kollokáció | helyzetmondat | szólás | közmondás]; **3.** a stílusminősítés [bizalmas | durva | gúnyos | hivatalos | népi | régi | ritka | rosszalló | szépítő | szleng | tájnyelvi | tréfás | választékos]; **4.** a definíció

A frazémák típusának jelzése

A frazéma-adatbázisban megjelöltük a frazéma típusát. Az alábbi kategóriákat határoztuk meg, és megadjuk, hogy az adott kategóriából kerekítve hány darab található az adatbázisunkban.

EK – Szókapcsolat, melynek van egyetlen szóból álló ekvivalense (260 darab)
Pl.: *gazdasági növény: haszonnövény*
költöző madár: vándormadár
Krisztus földi helytartója: pápa

KO – Kollokáció (1650 darab)
Pl.: *elveszti a fogadást*
kezet fog <vkivel>
szoros eredmény

HMO – Helyzetmondat (4300 darab)
Pl.: *Eszed tokja!*
Meglesz ennek még a böjtje!
Sokra megyek vele!

HMI – Helyzetmondat alesete, mely az írott nyelvhasználatra jellemző (15 darab)
Pl.: *Isten dicsőségére állíttatta:*
Kihajolni veszélyes!
Minden külön értesítés helyett.

HMM – Helyzetmondat alesete: népmesékre jellemző fordulat (10 darab)
Pl.: *Boldogan éltek, míg meg nem haltak.*
Hol volt, hol nem volt.
Szerencséd, hogy öreganyádnak szólítottál.

MSZ – Helyzetmondat alesete: megszólítás, köszönés (60 darab)
Pl.: *A vérfagyasztó viszontlátásra!* (biz)
Jó szerencsét!
Kezét csókolom!

SZO – Szólás (18850 darab)
Pl.: *a szavakon lovagol*
egy húron pendül <vkivel>
szélmalomharcot folytat

SZH – Szólás alesete: szóláshasonlat (750 darab)
Pl: *Áll, mintha gyökeret vert volna a lába.*
Úgy hasonlítanak egymásra, mint két tojás.
Olyan a bőre, mint a rinocérosznak.

KM – Közmondás (950 darab)
Pl: *A rest kétszer fárad!*
Ahol nincs, ott ne keress!
Reggeli vendég nem sokáig marad (népi)

A stilisztikai minősítések

A frazéma-adatbázisban, ahol szükséges volt, a frazémákat stilisztikai minősítéssel láttuk el. Az alábbiakban közöljük a leggyakoribb minősítéseket, példákkal, és feltüntetjük, hogy az adott stílusminősítéssel hány frazémát láttunk el.

BIZ – a bizalmas nyelvhasználatra jellemző (3050 darab)
Pl.: *Akárcsak a falnak beszélnék!*
csinálja a cirkuszt
nincs egy megveszekedett vasa sem

DUR – a durva, trágár nyelvhasználatra jellemző (350 darab)
Pl.: *Hallgat, mint szar a fűben.*
kitapossa a belét <vkinek>
vén szatyor

GÚNY – gúnyos jelentésű, jelentésárnyalatú frazéma (350 darab)
Pl: *egy kézlegyintéssel intéz el* <vmit>
feltalálta a spanyolviaszt
Nagy az Isten állatkertje, és sok bolond lakik benne.

HIV – a hivatalos, sajtónyelvi nyelv jellegzetes frazémája (50 darab)
Pl.: *benyújtja a lemondását*
halottá nyilvánít <vkit>
késedelmet szenved

NÉP – a népi és a tájnyelv jellegzetes frazémája (1250 darab)
Pl.: *Guba gubához, suba, subához*
Hencidától Boncidáig folyt a sárga lé.
Mennél koszosabb a malac, annál jobban vakaródzik.

RÉG – a régi nyelvben használatos, ma már elavult frazéma (950 darab)
Pl.: *Gyepre legény!*
tetézve adja vissza a csapott vékát
Üres gyomornak nem kell prédikáció.

RIT – ritkán használt, nem gyakori frazéma (800 darab)
Pl.: *magával hord* <vmit>*, mint a csiga a házát*
ráfordítja a kulcsot <vkire>
zászlós bajusz

ROSSZ – rosszalló stílusértékű frazéma (80 darab)
Pl.: *a színfalak mögött*
egy követ fúj <vkivel>
tartja a markát

SZÉP – szépítő, eufemisztikus frazéma (100 darab)
Pl.: *a szerelem papnője*
magához szólít <vkit> *az Úr*
testi szükséglet

SZLE – a szlengben használatos frazéma (1150 darab)
Pl.: *eldobja az agyát*
lemegy hídba
tökig nyomja a gázt

TRÉF – a tréfás stílusértékű frazéma (650 darab)
Pl.: *Ha a hegy nem megy Mohamedhez, Mohamed megy a hegyhez.*
Kismise, nagymise, utoljára semmise.
Semmi baja, csak a torka véres.

VÁL – a választékos, az emelkedettebb nyelvhasználatban, az irodalom nyelvében előforduló frazéma (1500 darab)
Pl.: *búnak ereszti a fejét*
kiissza a méregpoharat
pusztába kiáltott szó

A többjelentésű frazémák feldolgozása

A nyelvben nem csak a szavak, hanem a frazémák is sok esetben többjelentésűek. Éppen ezért minden frazéma mellett pontosan feltüntetjük minden jelentését. Így vált lehetségessé, hogy a frazémát az adott jelentésének megfelelő szinonimikus csoportba soroljuk be. Mintaképpen bemutatunk néhány többjelentésű frazémát és közreadjuk az egyes jelentésekhez tartozó definíciókat.

1. *benne van a buliban* = előnyös munka, vállalkozás részese
2. *benne van a buliban* = együtt szórakozik a társasággal

1. *előkészíti a talajt* = a szükséges mezőgazdasági munkálatokat a talajon előre elvégzi
2. *előkészíti a talajt* = (átvitt értelemben) tevékenységével v. hatásával vkit, vmit fogékonnyá, alkalmassá tesz vmire

1. *felszaggatja a sebet* = a rajta levő kötést és vart letépi
2. *felszaggatja a sebet* = (átvitt értelemben, választékos) vmely már-már elfelejtett lelki fájdalmat felújít, felszínre hoz

1. *Isten malmai lassan őrölnek.* = {ha későn is, de utoléri a gonosz embert a méltó büntetés}
2. *Isten malmai lassan őrölnek.* = (tréf) {vmely hivatalban lassan intézik az ügyeket}

1. *jártatja a pofáját* = dicsekszik
2. *jártatja a pofáját* = (gyakran badarságokat fecsegve) feleslegesen és sokat beszél

1. *jó füle van* = élesen hall; kifogástalan zenei hallása van
2. *jó füle van* = (átvitt értelemben) azt is meghallja, amit nem neki szántak

1. *kitér a hitéből* = régi vallását elhagyja és esetleg más felekezet tagja lesz
2. *kitér a hitéből* = (tréf) (vminek hallatán) nagyon bosszús, dühös lesz v. nagyon megbotránkozik

1. *kitöri a nyakát* = baleset (különösen lezuhanás) következtében szörnyethal
2. *kitöri a nyakát* = (nagyra törő v. hatalmon levő személy) vállalkozására csúnyán ráfizet

1. *komoly szándékai vannak* = komolyan gondol a házasságkötésre
2. *komoly szándékai vannak* = nagyon szeretne vmit, erősen törekszik vmire

1. *kosarat ad* = <nő> a férfi házassági ajánlatát visszautasítja
2. *kosarat ad* = vmely ajánlatot visszautasít, elhárít magától

1. *Néha még a kapanyél is elsül.* = {veszélyre való figyelmeztetés kifejezése: a veszélytelennek látszó dolgok is veszélyessé válhatnak}
2. *Néha még a kapanyél is elsül.* = {bizonytalanság kifejezése: semmi sem lehetetlen}

1. *nem a mai világba való* <vki> v. <vmi> = élhetetlen, gyámoltalan, ügyetlen személy
2. *nem a mai világba való* <vki> v. <vmi> = túlhaladott, korszerűtlen dolog

1. *nem fér a fejébe* <vmi> = nem tud megérteni, fölfogni, megtanulni <vmit>
2. *nem fér a fejébe* <vmi> = nem tud elfogadni, elképzelni <vmit>

3 A kulcsfogalmak rendszere és a kulcsfogalmak szinonimáinak illesztése a definíciókra

A frazéma-adatbázisban található definíciók szemantikai elemzése után 850 kulcsfogalomsort határoztunk meg. A kulcsfogalomsorok szinonim szavakból épülnek föl. Példa kulcsfogalomsorokra (feltüntetve a sorszámukat is):

456: HARAGSZIK, MEGHARAGSZIK, MEGSÉRTŐDIK, NEHEZTEL, MEGNEHEZTEL, DUZZOG

518: MENTEGETŐZIK, MAGYARÁZKODIK

774: RAVASZ, AGYAFÚRT, FORTÉLYOS, CSELES, HUNCUT

812: SZÍVESEN, ÖRÖMMEL, KÉSZSÉGESEN

Minta a magyar szinonimikus frazémaszótárból

Az automatikus illesztés után manuális szerkesztés következett, melyek során az illesztés hibáit javítottuk. Ugyancsak kézzel dolgoztuk fel és bontottuk szét igei aspektus szerint a frazémákat. Egy-egy kulcsfogalomsor összetartozó, aspektusbeli aleseteit megkülönböztetjük. Ezeket az alábbi mintában csillaggal (*) jelöljük.

480 IHLET
a múzsa csókja
isteni sugallat
* IHLETET KAP
homlokon csókol a múzsa <vkit> (tréf)
megszáll a szentlélek <vkit> (tréf)

593 MUNKANÉLKÜLI, ÁLLÁSTALAN
állás nélküli
maga ura
senkinek sem parancsol
munka nélküli
pénztáros az ingyen uszodában (szleng)
* MUNKANÉLKÜLIVÉ VÁLIK, ELBOCSÁTJÁK, KIRÚGJÁK, FELMONDANAK <VKINEK>
kirepül az állásából (biz)
lapátra kerül (biz)
repül az állásából (biz)

801 SÁPADT, FALFEHÉR
annyi vére sincs, amennyivel egy szúnyog jóllakhatna
fehér, mint a meszelt fal
fehér, mint a kísértet (nép)
fehér, mint a lárva (nép)
minden csepp vér kiszaladt az arcából
nincs rajta emberi szín
olyan a képe, mintha megette volna a szappant
olyan a színe, mint a fagyos ing (nép)
olyan fehér, mint akit a meszesgödörből húztak ki
olyan, mint a napon sült málé (rég)
pápista a színe
sápadt, mint a holdvilág
sápadt, mint a hulla
szép, mint a tejbetök
* ELSÁPAD, ELFEHÉREDIK
kifut az arcából a vér
kiszalad az arcából a vér

821 SZABAD, ÖNÁLLÓ, FÜGGETLEN
a maga embere
a maga ura
kirepült már a fészekből (biz)
maga gazdája
maga kenyerén van (nép)
megáll a maga lábán
nem szorul senkire
senki nem parancsol neki

szabad, mint a madár
* ÖNÁLLÓSODIK
szárnyra kel (vál)
* FELSZABADUL, MEGSZABADUL
széttöri az igát
letépi a láncait
szétzúzza a láncokat

Meggyőződésünk, hogy a munka során létrehozott frazéma-adatbázis és a belőle generált frazéma-szinonimaszótár a magyar frazeológiai kincs mélyebb összefüggéseinek feltárása mellett magyar nyelvű szövegek automatikus feldolgozásához, szemantikai elemzéséhez is hatékonyan használható fel.

Bibliográfia

1. Bárdosi, V.: Problèmes posés par le traitement lexicographique des figés dans les dictionnaires français. In: Fremdsprachen Lehren und Lernen, Jg. 21 (1992) 104–116
2. Bárdosi V. (főszerk.): Magyar szólástár. Szólások, helyzetmondatok, közmondások értelmező és fogalomköri szótára. Tinta Könyvkiadó, Budapest (2003)
3. Bárdosi V., Kiss G., Kiss M., Rapcsák T.: Kísérlet magyar szavak jelentéshasonlóságának meghatározására a Magyar szókincstár segítségével. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Juhász Nyomda (2004 27–37
4. Bárdosi V.: Magyar szólások, közmondások értelmező és fogalomköri szótára. TINTA Könyvkiadó, Budapest (2009)
5. Forgács T.: Magyar szólások és közmondások szótára. Mai nyelvünk állandósult szókapcsolatai példákkal szemléltetve. TINTA Könyvkiadó, Budapest (2003)
6. Kiss G.: Magyar szókincstár. Rokon értelmű szavak, szólások és ellentétek szótára. Tinta Könyvkiadó, Budapest (1998)
7. Országh L. (főszerk.): A magyar nyelv értelmező szótára (I–VII.). Akadémiai Kiadó, Budapest (1959–1961)
8. Szemerkényi Á.: Szólások és közmondások. Osiris Kiadó, Budapest (2009)
9. T. Litovkina A.: Magyar közmondástár. Közmondások értelmező szótára példákkal szemléltetve. TINTA Könyvkiadó, Budapest (2005)
10. Tolnai L.: A magyarság szellemi néprajza. Harmadik kötet. Királyi Magyar Nyomda, Budapest (1933–1938) 397–433
11. Voigt V. (szerk.): A magyar folklór. Osiris Kiadó, Budapest (1998)

Nyelvtechnológiai módszerek a Budapesti Szociolingvisztikai Interjú lexikai és szintaktikai vizsgálatában

Váradi Tamás, Peredy Márta, Oravecz Csaba

MTA Nyelvtudományi Intézet
e-mail:{varadi,mperedy,oravecz}@nytud.hu

Kivonat A dolgozat célja a Budapesti Szociolingvisztikai Interjú társalgási moduljainak lexikai és szintaktikai elemzése nyelvtechnológiai módszerekkel. Az elemzés a gépi eljárással annotált szövegeket elsősorban statisztikai módszerekkel vizsgálja. A BUSZI társalgási nyelvhasználatát a Magyar Nemzeti Szövegtárból vett minta segítségével az írott nyelvhasználat jellemzőivel veti össze. Ahol erre mód nyílik, a BUSZI2 által vizsgált társadalmi csoportok közötti lexikális és mondatszerkesztésbeli nyelvhasználati különbségeket is vizsgálunk.

Kulcsszavak: beszélt nyelv, korpusz-összehasonlítás, korpuszhomogenitás, jellemzőszó-vizsgálat, mondatszerkesztés, szintaktikai elemzés

1. Bevezetés

A BUSZI 2 [7] öt foglalkozás szerinti társadalmi csoport nyelvhasználatát vizsgálja a szociolingvisztikai interjú Labov által kidolgozott módszerével. Ennek fontos eleme az irányított társalgás, melynek során a gondosan kiképzett terepmunkások kötelező, illetve tetszőlegesen választott témákat beszéltek meg az adatközlőkkel. A magnóra felvett anyag lejegyzése alapjában véve a helyesírási szabályokat követte, de a BUSZI vizsgálati kérdéseit tartalmazó szociolingvisztikai változók, illetve a beszéd prozódiai és paralingvisztikai kisérőjelenségei gondos megörökítését is. Az eredetileg házi norma szerint kidolgozott annotáció az anyag tartalmi felülvizsgálata után XML-szabványos alakra lett átalakítva.

A tanulmány két fő részre tagolódik. A 2. részben a lexikai vizsgálatok eredményeit mutatjuk be. A szokásos gyakorisági listák mellett, kísérletet teszünk a szövegváltozat egyedi jellemzőit tükröző lexikai mintázatok feltárására, valamint azok korszerű módszerrel történő vizualizációjára is. A 3. rész a szintaktikai elemzéseket tartalmazza, melyekhez az adatbázis reguláris lekérdező nyelvén definiált lokális grammatikákat használtunk fel. A szófajok és a felszíni szerkezeti minták statisztikai síkon megragadható jellegzetességeit az írott nyelvhasználattal, illetve a BUSZI2 adatközlő csoport egymás közötti összehasonlításával mutatjuk be. Rövid összefoglalás zárja a dolgozatot a 4. részben.

2. Lexikai vizsgálatok

2.1. Szókincs-gazdagsági vizsgálatok

A lexikai vizsgálatok legegyszerűbb változata a szövegek szókincsére irányul. Számos lehetséges mérőszám alkalmazható (pl. típus/token arány, hapax-gyakoriság, dislegomenon-gyakoriság), melyek több alkalmazásban is gyakran használatosak, például szerző-, illetve műfaj-azonosításban [5], de megbízhatóságuk éppen az egyszerűségük miatt alacsony. Az egyes szövegtípusok szókincsére vonatkozó néhány szembetűnő különbség azért kiolvasható belőlük. A mérőszámok közül néhány nagyon egyszerű statisztikát foglal össze a 2.1. táblázat. A kvóták kódjai az alábbi adatközlőcsoportokra vonatkoznak: KV1: tanárok; KV2: egyetemi hallgatók; KV3: bolti eladók; KV4: gyári munkások; KV5: szakmunkástanulók.

1. táblázat. Szóstatisztikai adatok a különböző szövegeken.

Jellemző	Korpusz						
	MNSz	Buszi	KV1	KV2	KV3	KV4	KV5
1. szóalak	224128	173331	36846	29278	40994	37116	29097
2. szótípus	52876	26449	8971	6776	8639	8601	6560
3. típus/token	0.236	0.1526	0.2435	0.2314	0.2107	0.2317	0.2255
4. normált szóalak	25000						
5. szótípus	10140		6704	6048	5866	6283	5935
6. típus/token	.4056		.2682	.2419	.2346	.2513	.2374
7. főnév	6813		4109	3535	3299	3808	3070
8. ige	3904		4231	3869	4544	4362	4396
9. Fn/Ige	1.7451		.97116	.91367	.7260	.8729	.6983
10. Hapax	4402		2416	2082	1912	2189	2021
11. Dislegomenon	1014		564	538	522	577	548

A 3. sor magasabb típus/token aránya abszolút mértékben gazdagabb szókincset tükröz (többféle szó fordul elő adott nagyságú szövegben), viszont a korpusz növelésével a típusok száma nem nő arányosan, ezért a normált korpuszméretből (25 ezer szó) számított érték (6. sor) mutatja pontosan az írott és beszéd változat közötti eltérést ezen mutató tekintetében. Jól látható, hogy az MNSz szövegeit szignifikánsan magasabb érték jellemzi. Szembetűnő az eltérés a főnév/ige használatban is, itt a beszélt nyelvi szövegre mutatható ki egyértelműen az igék használatának magasabb aránya a főnevekhez képest. Az egyszer, illetve kétszer használatos szótövek (10., 11. sor) előfordulási gyakoriságának különbsége is egyértelműen jelzi a írott változat nagyobb lexikális gazdagságát.

Fontos megjegyezni, hogy ugyan a kvóták között is jelentkezik különbség a mérőszámokban, megbízható eredményekhez azonban részletesebb vizsgálatokra, illetve nagyobb mennyiségű szövegre lenne szükség.

2.2. Jellemzőszó-vizsgálatok eredményei

Számos lehetséges módszer közül (l. pl. [3]) az alábbiakban egy olyan eljárás eredményeit mutatjuk be, amely az egyes korpuszok gyakorisági profiljainak összehasonlításával határozza meg az adott szövegre jellemző lexikai elemeket. Ebben az összehasonlításban azok a nyelvi elemek szerepelnek a rangsor elején, amelyek a két összehasonlított korpuszban jellegzetesek, mindig a másikhoz viszonyítva, vagyis az eljárás egy közös listát generál, melyet utána kvalitatív vizsgálatnak lehet alávetni.

A vizsgálatban először a két korpusz nyers gyakorisági listáit állítjuk elő, majd minden, a listában szereplő szóra log-likelihood statisztikát számolunk [4]. Az így kapott eredmények szerint rendezzük újra a gyakorisági listát, így a lista elején megkapjuk az egyik vagy másik korpuszra jellemző szavak halmazát.

Az alábbi táblázatokban szereplő listákban az első oszlop a számított súlyérték, a második a szó(tő), harmadik az egyik (C1), illetve másik (C2) korpuszbeli gyakorisági érték.

1. MNSz vs. teljes Buszi

6143.08381829616	hát	C1: 107	C2: 4232
3341.37775731983	igen	C1: 118	C2: 2512
3273.87396188346	én	C1: 307	C2: 2991
2688.3186152523	nem	C1: 2873	C2: 6672
2277.40930049444	van	C1: 2274	C2: 5438
1962.18484387733	a	C1: 21240	C2: 9687
1435.46798174574	szóval	C1: 21	C2: 973

2. KV1 vs. KV5

326.999757436891	hát	C1: 376	C2: 1044
91.7885470809332	akko	C1: 0.5	C2: 70
84.3856052479682	meg	C1: 146	C2: 347
68.5496070494782	szóval	C1: 46	C2: 162

3. KV5 vs. KV1

88.3354097204393	gyerek	C1: 108	C2: 12
53.6296019468509	ugye	C1: 59	C2: 5
49.7205434485945	a	C1: 1550	C2: 1182
47.4355156652772	gimnázium	C1: 40	C2: 1
42.7362191531778	tanít	C1: 50	C2: 5
42.0983293946033	tanár	C1: 36	C2: 1

Az eljárás eredményei elnagyolva, de nagyon szemléletesen ábrázolhatók „szófelhők” formájában, melyet az 1. ábra illusztrál.

1. ábra. Teljes Buszi vs. MNSz.

2.3. Korpuszok homogenitásának és hasonlóságának vizsgálata

Korpuszok hasonlóságának vizsgálata, illetve mérése leginkább nyelvtechnológiai alkalmazások kontextusában merül fel és arra vonatkozóan adhat hasznos információt, hogy meghatározott típusú szövegekre fejlesztett alkalmazások milyen ráfordítással alakíthatók át újabb szövegek kezelésére. Ha a szövegek hasonlóak, a ráfordítás feltehetően kisebb. Általánosságban ennél többet nem is nagyon lehet állítani, egyrészt a lehetséges mértékek is nagyon sokfélék, másrészt megbízhatóságuk is jelentősen függ az adott szövegektől, illetve alkalmazásoktól.

A hasonlóság kérdése szervesen összefügg egy adott korpusz homogenitásának kérdésével, hiszen nyilván nem egyértelmű, mit fejezne ki egy ilyen mérték abban az esetben, ha egy homogén korpuszt hasonlítunk össze egy általános, nagy változatosságot mutató heterogén korpusszal (a kérdés részletes tárgyalását l. [3]). Ideális esetben ugyanazt a mértéket lehet használni korpuszon belüli és korpuszok közötti „távolságok” mérésére, melyek így közvetlenül összehasonlíthatók. A jelen vizsgálatokban alkalmazott perplexitásmérték használható ilyen módon, de természetesen, mint minden más egydimenziós mérték, csak durva közelítésként értelmezhető egy alapvetően többdimenziós összehasonlítási feladatban.

A perplexitás annak mérésére szolgáló érték, hogy mennyire pontosan tudunk modellezni egy ismeretlen valószínűségi eloszlást a rendelkezésre álló tanító adatok alapján. Minél magasabb értéket kapunk a perplexitásra, annál bizonytalanabbak vagyunk az ismeretlen eloszlás tekintetében, annál kevésbé tudjuk megjósolni az ismeretlen eloszlás által generált adatokat. Szövegek vizsgálatára az alábbi módon tudjuk felhasználni: a szöveg egy meghatározott részét (pl. 90%) használjuk fel tanító adatként, melyből például egy trigram (szóhármas) alapú statisztikai nyelvi modellt építünk (vagyis szóhármasok előfordulási gyakorisága alapján számoljuk ki a mondatok valószínűségét). Ezek után lemérjük, hogy a készült modell alapján milyen pontosan tudjuk megjósolni a tesztadatként (a maradék 10%) szereplő szövegrész mondatait. Minél nagyobb perplexitásértéket kapunk, annál távolabb áll a tanító és tesztszöveg egymástól, minél kisebbet,

annál homogénebb és annál jobban hasonlít a két szövegrész egymásra. Informálisan, a perplexitásra kapott számérték annak a szóhalmaznak a nagyságát határozza meg, amelyből (trigram nyelvmodell esetén) a megelőző két szó ismeretében a következő szót választhatjuk. Minél kisebb ez a halmaz, modellünk annál megszorítottabb [2].

2. táblázat. Perplexitásértékek a különböző szövegeken.

Tanító korpusz	Tesztkorpusz						
	MNSz	Buszi	KV1	KV2	KV3	KV4	KV5
1. MNSz	733.618	–	–	–	–	–	–
2. Buszi	–	121.52	–	–	–	–	–
3. KV1	–	–	123.835	123.462	113.633	118.273	107.597
4. KV2	–	–	122.666	115.782	–	–	–
5. KV3	–	–	124.402	–	101.542	–	–
6. KV4	–	–	130.106	–	–	108.828	–
7. KV5	–	–	127.512	117.237	110.695	116.796	89.401

Az itt végzett vizsgálatok sztenderd tízszeres keresztvalidációval készültek, a CMU-Cambridge Statistical Language Modeling Toolkit [1] segítségével, a morfológiai variábilitásból eredő eleve magas értékeket kiküszöbölendő a szokásos gyakorlatnak megfelelően szótövesített szövegekkel. A kapott eredmények a 2.3. táblázatban láthatók. Az egyes sorokban szereplő szövegekből készült a nyelvmodell, az oszlopok jelzik a tesztadatot. Abban a cellában, ahol mindkét, a sorban és oszlopban szereplő szöveg azonos, ott az adott korpusz homogenitására vonatkozó érték szerepel, a további cellákban pedig a különböző korpuszok hasonlóságát jellemző érték jelenik meg. Mivel a vizsgálat illusztratív, nem törekszik kimerítő jellemzésre, inkább a szembetűnő jellegzetességekhez kíván kvantitatív mérőszámot rendelni, ezért nem minden cellában szerepel (az egyébként minden esetben számítható) mutató. Néhány összehasonlítás a szövegek jellegéből következően nem hordoz lényeges információt, így azokat eleve nem érdemes elvégezni. Mivel az itt szereplő MNSz-minta jól láthatóan igen heterogén, nagy variábilitású szövegeket tartalmaz, a Buszi-szövegekkel való összehasonlítás nem eredményezne újabb információt azon túl, hogy az írott szöveg a beszélthez képest sokkal változatosabb, ez pedig a homogenitásadatokból is egyértelműen látszik már. A Buszi-szövegek vizsgálatában pedig informatívabb az egyes kvóták anyagát egymással összehasonlítani, mint a teljes Buszi-anyagot a kvóták anyagával; ez utóbbi esetben sem kapunk az előbbi vizsgálathoz képest új információt.

Az egyes korpuszrészek, kvóták homogenitására vonatkozó értékből kiolvasható, hogy az adott kvótához tartozó beszélőknek mennyire változatos a nyelv-

használata. A kvóták egymással történő összehasonlításából kapott értékek arra adnak választ, hogy a kvóták szövegei mennyire állnak közel egymáshoz, illetve az egyik szöveg milyen mértékig „foglalja magában" a másikat. A KV1 és KV2 korpusz például ebben az összehasonlításban viszonylag távol esik egymástól, míg ha a KV5 korpuszhoz hasonlítjuk például a KV1 korpuszt, akkor jelentős távolságot kapunk, fordított irányban pedig alacsonyat, vagyis a KV1 korpuszból épített modell „magában foglalja" a KV5 korpuszt is.

3. Szintaktikai elemzések

3.1. Mondathossz

A szintaktikai vizsgálatok alapegysége a mondat, így minden szintaktikai elemzés a mondathatárok megállapításával kell, hogy kezdődjön. Az írott nyelvi korpuszban ez nem jelent problémát, a beszélt nyelvi korpuszt tanulmányozva azonban talán a korpusz elemzésének legbizonytalanabb pontja éppen ez. A beszélők ugyanis (szemben az írott szövegek létrehozóival) nem jelzik egyértelműen, hogy hol van szerintük a mondataik vége. A BUSZI-korpusz tagolásánál a szöveget annotáló személyek anyanyelvi intuíciójuk alapján állapították meg a mondathatárokat.

A két korpusz közti első szembetűnő különbséget a 3. táblázat mutatja. Az írott nyelvi anyag átlagos mondathossza (17,1 szó) kétszerese a BUSZI-adatközlők élő beszédbeli mondatainak (8,5 szó). A BUSZI-terepmunkások megszólalásainak célja elsősorban az adatközlők beszédének terelgetése volt, így nem meglepő, hogy az ő megszólalásaik még rövidebb mondatokra tagolódnak. (A teljes BUSZI-beli átlagos mondathossz 6,5 szó.)

3. táblázat. Átlagos mondathossz.

	BUSZI		MNSZ
	terepmunkások (tm)	adatközlők (ak)	
átlagos mondathossz	4,6	8,5	17,1

Ragozott igealakok – tagmondatok. A mondatszerkezet szempontjából a legfontosabb eltérés a ragozott igealakok számában figyelhető meg. A BUSZI-ban másfélszer annyi ragozott ige van (15%), mint az MNSZ-ben (10%), l. alább 4. táblázat. Ez az adat utal arra az alább alaposan vizsgált tényre, hogy az írott nyelv több információt sűrít a főnévi csoportokba jelzős szerkezetek segítségével, míg a beszélt nyelv több alárendelt mondatot, és így több ragozott igét használ. Figyelembe véve, hogy tagmondatonként egy ragozott igével számolhatunk, megállapítható a tagmondatok átlagos hossza. A BUSZI-ban 6,7, az MNSZ-ben

10 szó adódik. Ezeket az értékeket összevetve a feljebb említett átlagos mondathosszal (BUSZI: 6,5; MNSZ: 17,1) azt kapjuk, hogy a BUSZI mondatai jellemzően egy tagmondatból állnak, hiszen az átlagos mondat- és tagmondathossz gyakorlatilag azonos, míg az MNSZ mondatai 1,7-szer hosszabbak, mint a tagmondatai, tehát a tipikus mondat két tagmondatból áll.

A bővítmények száma. Az NP-k számát a tagmondatok számához (azaz a ragozott igékhez) viszonyítva, azt látjuk, hogy míg a BUSZI-ban kettőnél kevesebb NP jut egy tagmondatra, addig az MNSZ-ben 3,5, vagyis az írott nyelv mondatai több bővítményt tartalmaznak. (L. alább 3.1. pont és a 4. táblázat.)

3.2. Szófajstatisztika

Már a legdurvább statisztikai elemzés, a különböző szófajú szavak számának összevetése is sokat elárul a beszélt nyelvi és az írott nyelvi korpusz mondatszerkezeti különbségeiről. A 4. táblázat a különböző szófajú szavak megoszlását mutatja a két korpuszban. Láthatjuk, hogy a legtöbb esetben az adatközlők és a terepmunkások szófajarányai közel azonosak még a diskurzusban betöltött eltérő szerepek ellenére is, míg az írott nyelvi szófajmegoszlás jelentősen eltér. Megjegyezzük, hogy a dolgozatban alább közölt statisztikai eltérések, ha külön nem jelezzük, akkor 5%-os szignifikanciaszint mellett mindig szignifikánsak.

4. táblázat. Szófajok.

Szófaj	BUSZI				MNSZ	
	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
N-ek száma	12,5	11904	14,0	24345	29,0	87479
Pro-k száma	13,7	13013	12,9	22388	5,5	16667
számnév	2,3	2200	3,7	6435	3,6	10861
egy-ek száma	1,0	917	1,4	2457	0,6	1930
Det-ek száma	6,4	6094	7,0	12058	12,3	37135
A-k száma	5,6	5315	5,6	9649	10,7	32149
Adv-ok száma	20,5	19533	20,0	34722	7,6	22879
finit V-k száma	15,2	14477	15,3	26570	9,9	29943
mn-i igenevek	0,5	512	0,5	946	3,3	9840
fn-i igenevek	1,7	1616	1,7	3010	1,0	3096
hat-i igenevek	0,2	146	0,2	331	0,3	896
kötőszók	10,9	10393	12,2	21086	7,5	22651
névutó	0,7	634	0,9	1567	1,6	4778
indulatszó	1,5	1461	0,4	644	0,0	116
egyéb	7,3	6919	4,1	7120	6,9	20881

3.3. A főnévi csoport

Az alábbiakban a főnévi csoportok szerkezetével foglalkozunk részletesebben, ugyanis a közölni kívánt tartalom átadásának két véglete közül az egyik az, amikor minden egyes információdarabnak egy-egy tagmondat felel meg, míg a másik véglet a tömörített szöveg, amelyben az információ minél nagyobb részét egy mondatba kívánja foglalni a beszélő (vagy a szöveg írója), és ezért a tartalom jelentős része a mondaton belüli főnévi csoportokban jelzői szerkezetekbe sűrítve jelenik meg.

A beszélt és az írott nyelvi korpusz főnévi csoportjainak összehasonlításakor fő hipotézisünk tehát az, hogy az írott nyelvben sokkal inkább megfigyelhető az információ főnévi csoportokba tömörítése, mint a beszélt nyelvben.

A főnévi csoport feje. A főnévi csoport feje főnév vagy névmás lehet és megfordítva, minden főnévre, illetve névmásra épül egy teljes főnévi csoport. Az 5. táblázatban a főnévi csoportok számát a főnevek plusz névmások számával azonosítottam, ami annyiban pontatlan, hogy a jelzőkkel bővített főnévi csoportból olykor el van hagyva a főnévi fej, illetve a mutató névmás nem mindig alkot önálló főnévi csoportot (pl. *ezt a kutyát*). Ezekről az esetekről alább még lesz szó. A főnévi csoportok jellemzően a mondat ragozott igéjének bővítményeiként jelennek meg a mondatban, de melléknévi csoportok (pl.: *büszke a fiára*), más főnévi csoportok (pl.: *a fiú a távcsővel*, *a fiúnak a távcsöve*) és igenevek (pl.: *a kertben játszó gyerek, uszodában úszni*) bővítményei is lehetnek.

Összességében több főnévi csoport van az MNSZ-ben, mint a BUSZI-ban. A főnevek és névmások összesített aránya a teljes szószámhoz képest rendre 35%, illetve 26%. Ez az adat máris mutatja, hogy az írott nyelvi korpuszban nagyobb szerepe van a főnévi csoportoknak, mint a beszélt nyelvben, összhangban azzal a 3.1. pontban említett adattal, hogy a ragozott igék relatív száma viszont a beszélt nyelvben magasabb.

Fontos további jellemzője a beszélt nyelvi korpusznak, hogy a főnévi csoportok között sokkal nagyobb arányban vannak a névmások, mint az írott nyelvben. Míg az írott nyelvben a félreértés elkerülése végett érdemes egy teljes leírással egyértelműsíteni, hogy mire utalunk, addig a beszélt nyelv sokkal inkább támaszkodhat az egyértelműsítés nem nyelvi eszközeire is (pl. mutatás), illetve esetleges félreértés esetén lehetőség lenne visszakérdezni, így a figyelem középpontjában álló (széliens) individuumokra elegendő csupán névmással utalni. A főnévi, illetve névmási fejek aránya a BUSZI-ban közelítőleg 50-50%, míg az MNSZ-ben 84-16% a főnevek javára.

Az adatokhoz három pontosító megjegyzést kell fűznünk. Egyrészt meg kell jegyeznünk, hogy a jelzővel bővített NP főnévi feje olykor elmaradhat (pl. *a sárga tulipánból* helyett *a sárgából*), a nem alanyesetű melléknevek csak ilyen esetekben jelennek meg, ezért az esetragos melléknevek és főnevek számának összevetéséből látható, hogy milyen gyakran maradhat el a főnévi fej a főnévi csoportokból. A BUSZI-ban ez az arány 7,2%-nek adódik a terepmunkások és 6,4%-nek az adatközlők esetében, míg csupán 3,5% az MNSZ-ben. Az ellipszisek valódi száma

5. táblázat. Főnévi csoportok.

	BUSZI		MNSZ
A főnévi csoport feje	tm	ak	
főnevek aránya (%)	47,8	52,1	84,0
névmások aránya (%)	52,2	47,9	16,0
A főnévi csoportok száma			
a szószámhoz képest (%)	26,2	26,6	34,5
a finit igék számához képest (%)	1,7	1,8	3,5

azonban ennél alacsonyabb, ugyanis bizonyos főnévként és melléknévként is érthető szavak melléknévként vannak megjelölve a korpuszban, és ezért például az *a törpéket* főnévi csoportban a *törpe* esetragos melléknévként számolódik. Az ebből fakadó hiba vélhetőleg egyformán érinti a BUSZI és az MNSZ korpuszt, így ha a kapott értékek nem is pontosak, arányuk jól mutatja, hogy az MNSZ NP-i teljesebbek, nemcsak hogy ritkábban fejezhetők ki névmással, de a főnévi fej is kevésbé hagyható el belőlük.

Másrészt, mint említettük az NP-k, bár leggyakrabban a mondat ragozott igéjének bővítményei, de nem feltétlenül azok, és ezek az esetek torzítják az egy ragozott igére eső NP-k számára kapott értéket. Harmadrészt a mutató névmások (*ez*, *az*) összes előfordulásainak a BUSZI-ban mintegy 20%-a, az MNSZ-ben 27%-a nem önálló NP-ként, hanem egy határozott főnévi csoporttal együtt fordul elő, ezeket tehát le kell vonnunk az önálló főnévi csoportként elszámolt névmások közül. Ez a kis korrekció azonban a névmások és főnevek arányára kapott értékeket lényegében nem módosítja.

Jelzős szerkezetek. Feltevésünk szerint az írott nyelvi korpuszban több és összetettebb jelzős szerkezeteket találunk, mint a beszélt nyelvben. Ezt vizsgáljuk alább a névelőt is tartalmazó NP-ken a melléknévi, majd a melléknévi igeneves jelzők esetén.

Halmozott melléknévi jelzők

A BUSZI-ban a névelős főnévi csoportoknak kb. 58%-a bővítetlen, az MNSZ-ben hasonló, de ennél valamivel alacsonyabb, 54% az arány. Az egy melléknévi jelzőt tartalmazók közel kétszer annyian vannak az MNSZ-ben, mint a BUSZI-ban, a két melléknévvel bővítettek már 2,5-szer, a hárommal bővítettek négyszer annyian. Négy melléknévi jelzőt tartalmazó NP a BUSZI-ban már nem található.

Melléknévi igenevek

A melléknévi igenevek használata sokkal gyakoribb az MNSZ-ben, mint a BUSZI-ban, az adatokkal azonban óvatosnak kell lennünk, mert a melléknévi igenevek közül sok valójában már melléknévként lexikalizálódott (pl.: *elvált*), elkülönítésükre azonban az annotáció nem ad lehetőséget.

A jelző + főnév szerkezetek között a melléknévi igenévi jelző a BUSZI-ban kb. 11%-ban, míg az MNSZ-ben kétszer olyan gyakran, 22%-ban fordul elő.

6. táblázat. A bővítetlen és a melléknevekkel bővített névelős főnévi kifejezések százalékos aránya a névelők összes számához képest.

	BUSZI		MNSZ
Halmozott mn.-i jelzők	tm	ak	
névelő+főnév	60,0	57,3	54,2
ne+mn+fn	8,60	8,77	17,07
ne+2mn+fn	0,90	0,86	2,36
ne+3mn+fn	0,06	0,06	0,23
ne+4mn+fn	0,00	0,00	0,01

7. táblázat. A melléknévi igenevek százalékos aránya a szavak számához viszonyítva.

	BUSZI				MNSZ	
M. igenevek	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
folyamatos	0,3	293	0,3	600	1,9	5806
befejezett	0,2	214	0,2	340	1,3	3922
beálló	0	5	0	6	0	112

8. táblázat. A melléknévi és melléknévi igenévi jelzők aránya.

	BUSZI		MNSZ
Igenévi/melléknévi jelzők	tm	ak	
melléknévi igenév+fn	11,2	10,1	21,8
melléknév+fn	88,8	89,9	78,2

A melléknévi igenevek használata jó módja az információ NP-n belüli tömörítésének, mivel az ige nemcsak magában, hanem bővítményeivel együtt is megjelenhet így jelzőként. A 9. táblázat adatai alátámasztják azt a feltételezést, hogy az írott nyelvi szöveg sokkal inkább él ezzel a lehetőséggel, ugyanis mintegy négyszer olyan gyakran van bővítménye az igenévi jelzőnek az MNSZ-ben, mint a BUSZI-ban.

9. táblázat. A bővítményes melléknévi igenévi jelzők százalékos aránya melléknévi igenévi jelzők között.

	BUSZI				MNSZ	
Bővített mn.-i ign. jelzők	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
bővítmény+m. igenév+fn	11,5	25	10,5	43	41,3	2271

Birtokos szerkezet. A kétféle birtokos szerkezet, az alanyesetű, illetve a -nAk ragos birtokost tartalmazó, megoszlása eltér a két korpuszban. A birtokot közvetlenül megelőző, nem névmási birtokost tartalmazó szerkezeteket (pl. *a kutyának a szőre* vs. *a kutya szőre*) vizsgálva kitűnik, hogy a -nAk ragos (datívuszos) birtokos aránya a beszélt nyelvben lényegesen nagyobb. A BUSZI-adatközlők között több, mint hússzor gyakoribb, mint az MNSZ-ben (10. táblázat). Ez szintén arra utal, hogy az írott nyelv sokkal inkább a tömörségre törekszik: ha az adott szerkezet egyértelmű, akkor fölösleges a kitett raggal redundánsan megjelölni a birtokost. A kötetlen beszéd kevésbé "spórol". Továbbá egy közel kétszeres szorzókülönbség a terepmunkások és az adatközlők adatai között is mutatkozik.

10. táblázat. A birtokot közvetlenül megelőző birtokosok között a datívusz aránya.

	BUSZI				MNSZ	
Birtokos szerkezetek	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
alanyesetű birtokos	89,6	146,81	5,211	99,1	5101	
részesesetű birtokos	10,4	17,18	5,48	0,9	44	

Többszörös birtokos szerkezetek a BUSZI-ban nem fordulnak elő, míg az MNSZ-ben igen, a birtokos szerkezetek 6%-ában.

3.4. Vonatkozói mellékmondatok

A BUSZI korpusz azt bizonyítja, hogy az *amely* és a *mely* vonatkozó névmás a beszélt nyelvből mára szinte teljesen eltűnt. Az adatközlők közül a *mely*-t senki, az *amely*-t csak a tanárok és az egyetemisták használták, így az adatközlők által használt összes vonatkozó névmásnak csak 0,7%-a volt *amely*, míg az MNSZ referenciakorpuszban a *mely* és az *amely* együttesen 41%-ot tesz ki. Az *amelyik* viszont az írott nyelvből hiányzik (0,4%), míg a BUSZI-ban több, mint 3%-ot képvisel. A beszélt nyelvben ugyanis az *amelyik* nem csak kiválasztó értelemben szerepel, hanem az *ami* és az *amely* helyett is, l. (1). A vonatkozó névmások BUSZI-beli használatáról részletesen ír [6].

(1) *s az Árpád Gimnáziumnak akkor még volt egy öööö nagyon jól működő cserkész csapata, amelyik különböző rendezvényeket öööö gyártott, rendezett*

A vonatkozó névmások összes száma megadja a vonatkozói mellékmondatok számát. A vonatkozói mellékmondatok összes NP-hez képesti aránya az MNSZ-ben alacsonyabb, 2,9%, míg a BUSZI-ban a terepmunkások esetében 4,1%, az adatközlőknél 4,3%, vagyis a beszélt nyelv feltevésünknek megfelelően, valóban gyakrabban fogalmazza a mondanivalót külön tagmondatba.

11. táblázat. A vonatkozó névmások előfordulásai.

Vonatkozó névmások	BUSZI tm %	BUSZI tm ∑	BUSZI ak %	BUSZI ak ∑	MNSZ %	MNSZ ∑
aki/ami-k száma	92,4	970,96	1,1978	58,8	1799	
amely-ek száma	4,2	44,0	6,13	33,1	1011	
mely-ek száma	0,3	3,0	0,0	7,7	236	
amelyik-ek száma	3,1	33,3	3,67	0,4	12	

A vonatkozói mellékmondatok közül csak a mondatkezdő pozícióban állókat vizsgálva szintén érdekes különbségek adódtak a két korpusz között. A vonatkozói mellékmondatok topikalizációval kerülnek a mondat élére. Ha a vonatkozói mellékmondatnak névmási feje van a mondatban, akkor ez a névmás mindig a mutató névmás (*az*). Az MNSZ-ben azonban a vonatkozói mellékmondat névmási feje az esetek felében el van hagyva, pl. (2-a), míg a BUSZI-ban szinte mindig megjelenik, pl. (2-b).

(2) a. ***Aki*** *erre jár és körül akar nézni ,* ~~***azt***~~ *szívesen fogadjuk* (MNSZ)
b. ***Aki*** *hisz Istenbe,* ***az*** *hisz pap nékül is* (BUSZI)

A 12. táblázat első két sora mutatja ezt az eredményt. A terepmunkások szövegében összesen 8 olyan mondat fordult elő, amelybe a mondat eleji vonatkozói mellékmondat után beilleszthető a mellékmondat *az* névmási feje, és ebből csupán egyszer maradt el az *az*, míg az adatközlők esetében 19 esetből egyszer sem. Ezzel szemben az MNSZ-ben 43 esetből 21-ben el volt hagyva az *az*. A névmás hiányát tekinthetnénk az elhagyott hangsúlytalan személyes névmási fej esetének, ám ekkor az itteni eredmények ellentmondanának a *hogy*-os tagmondatok fejével kapcsolatban tapasztalt tendenciának, mely szerint az írott nyelv sokkal inkább a hangsúlyos *az* névmást használja, míg a beszélt nyelvben gyakrabban előfordul fejként a hangsúlytalan személyes névmás is, és ez utóbbi lenne az, ami (alany- és tárgyesetben) elhagyható. Az adatok helyes értelmezése az, hogy a vonatkozói mellékmondatot már önmagában, a névmási fej nélkül is referáló bővítményként tudjuk értelmezni, és ekkor a névmási fejre nincs szükség. A mondatkezdő vonatkozói mellékmondat után mégis gyakran és legfőképpen a beszélt nyelvben megjelenő *az* inkább topikismétlő névmásnak tekinthető.

A mondatkezdő vonatkozói mellékmondatoknak a következő csoportja a visszautaló típus, lásd a 12. táblázat 3. sorát. A BUSZI-ban ugyanis a mondat eleji vonatkozói mellékmondat gyakran nem az adott főmondat valamely frázisának bővítménye, hanem az előző mondat valamely szereplőjére utal vissza, például (3). Ezekben az esetekben a vonatkozó névmás szintén referenciális kifejezésként viselkedik, tulajdonképpen személyes névmási funkcióban jelenik meg. A terepmunkások beszédében a mondatkezdő vonatkozói mellékmondatok 27%-a, az adatközlők beszédében ezek 42%-a utalt előző mondatbeli szereplőre, míg az MNSZ-ben csupán 9%.

12. táblázat. A mondatkezdő vonatkozói mellékmondatok típusai.

Vmm kezdetű mondat	BUSZI tm		BUSZI ak		MNSZ	
	%	∑	%	∑	%	∑
vmm + *az*	21	2,7	38,0	19	31,9	22
vmm + elhagyott *az*	3,0	1	0,0	0	30,4	21
visszautaló vmm	27,3	9	42,0	21	8,7	6
kettőspontos értelmezés	3	1	0	0	13	9
egyéb	45,5	15	20	10	15,9	11

(3) *Előtte Zuglóban laktunk a nagymamáméknál. Aki most ott lakik szintén egyedül*

Végül az MNSZ-ben több példát is találunk (13%) a mondatkezdő vonatkozó mellékmondat kettőspontos értelmezésére, pl. (4), míg a BUSZI-ban összesen egy ilyen mondat szerepelt. A kettőspontos értelmezés az *Ami . . . , az az, hogy . . .* mondat rövidített változatának tekinthető, ezt a tömörítést jellemzően az írott nyelv alkalmazza.

(4) *Ami még ennél is fontosabb: a televíziók nem a háború valóságos emberi vonatkozásaira voltak kíváncsiak*

4. Összefoglalás és további feladatok

A tanulmányban az írott és beszélt nyelvhasználat néhány jellemző különbségét illusztráltuk lexikai és szintaktikai elemzés alapján. Viszonylag egyszerű eszközökkel kaptunk nem triviális eredményeket, melyek alapul szolgálhatnak további nyelvi elemzéseknek.

A beszélt nyelvi korpusz méretének és a mondatelemzés mélységének a növelésével részletesebb vizsgálatok is elvégezhetők, mint például a BUSZI kvóták közötti különbségek nyelvstatisztikai elemzése vagy a szórendre vonatkozó elemzések.

Hivatkozások

1. Clarkson, P. R. és Rosenfeld, R. Statistical language modeling using the CMU-Cambridge toolkit. In: *EUROSPEECH-97*, 1. kötet, 1997, 2707–2710.
2. Jelinek, F., Mercer, R. L., Bahl, L. R. és Baker, J. K. Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, November, 1977, 62:S63. Supplement 1.
3. Kilgarriff, Adam. Comparing Corpora. *International Journal of Corpus Linguistics*, 2001, 6(1):1–37.

4. Rayson, Paul és Garside, Roger. Comparing Corpora using Frequency Profiling. In: *Proceedings of the Workshop on Comparing Corpora*. Association for Computational Linguistics, 2000., 1–6.
5. Stamatatos, Efstathios, Fakotakis, Nikos és Kokkinakis, George. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 2000, 26 (2):471–495.
6. Szeredi, Dániel. Vonatkozó névmások használata beszélt nyelvi korpusz alapján. Szakdolgozat, ELTE, 2008.
7. Váradi, Tamás. A Budapesti Szociolingvisztikai Interjú. In: Kiefer, Ferenc és Siptár, Péter szerk. *A magyar nyelv kézikönyve*. Akadémiai Kiadó, Budapest, 2003, 339–359.

VII. Gépi tanulás

Szótáralapú névelem-felismerés szóhatárainak javítása gépi tanulási módszerrel

Móra György[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.,
gymora@inf.u-szeged.hu

[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
Szeged, Tisza Lajos krt. 103. III. lépcsoház
rfarkas@inf.u-szeged.hu

Kivonat: Cikkünkben angol biológiai és magyar nyelvű névelemeket felismerő rendszert mutatunk be. Megközelítésünk a szótáralapú és a gépi tanuló módszerek előnyeit ötvözi. A szótáralapú névelem-felismerők egy adott adatbázis alapján jelölik a szövegbeli előfordulásokat, így a névelemek előfordulásaihoz hozzárendelhetőek azok egyedi azonosítói. Az illesztett névelemek határainak korrekcióját, valamint a hibásan illesztett kifejezések kiszűrését a feltételes véletlen mezők módszerén alapuló statisztikai rendszerrel végeztük el. Módszerünk összehasonlítva más megközelítésekkel a magyar tulajdonnevek felismerésében közel azonos, a biológiai névelemek felismerésében pedig jobb eredményt ért el, mint a klasszikus névelem-felismerő módszerek.

1 Bevezetés

Jelen munkánkban egy hibrid névelem-felismerő rendszert mutatunk be, amely ötvözi a szótáralapú névelem-azonosítás előnyeit a gépi tanulási módszerek rugalmasságával. A névelem-felismerés során a szövegben megtalálható olyan elemeket azonosítjuk, amelyek valamilyen egyedi névvel rendelkező objektumot jelölnek. Ilyenek a tulajdonnevek, amelyek személyt, földrajzi helyet vagy szervezetet jelölnek. Ezeken kívül a különböző tudományterületeknek – mint például a biológiának – rendszerint saját névelemtípusai vannak.

A biológiai szövegek feldolgozásában különösen fontos a névelem-felismerés alkalmazása. A szövegben megtalálható gének, fehérjék és egyéb biológiai névelemek közötti relációk, valamint a gazdagabb információtartalommal bíró biológiai események kinyerésének alapja a névelemek azonosítása, amely a jelenlegi rendszerekben a szótárakban található entitásnevek segítségével történik.

Mind a biológiai névelemekhez, mind a magyar tulajdonnevekhez rendelkezésre állnak szótárak, amelyek a kifejezések előfordulásainak jelentős hányadát lefedik. A kizárólag szótárakat alkalmazó névelem-felismerő rendszerek jó fedést biztosítanak, ám a pontosság a szótár méretével csökken. Ennek kiküszöbölésére rendszerint szakértők által alkotott utófeldolgozási szabályokat alkalmaznak.

Bemutatunk egy módszert, amely alapvetően szótárillesztést hajt végre, de gépi tanuló rendszerek alkalmazásával a névelemek környezetének figyelembevételével kiszűrhetik a hibásan jelölt névelemeket, illetve korrigálhatjuk a névelemek határait, anélkül, hogy elveszítenénk azok pontos azonosításának lehetőségét.

Kísérleteink folyamán magyar nyelvű hírekben és angol nyelvű biológiai szövegekben jelöltünk névelemeket szótárillesztő módszer segítségével. A rendelkezésünkre álló tanító adatbázisok segítségével a névelemjelöltek és a környezetükben található szavakból jellemzőket nyertünk ki, amelyeket felhasználva a feltételes véletlen mezők módszer segítségével gépi tanuló modellt építettünk

2 Kapcsolódó munkák

A biológiai névelemek felismerésére alkalmazott szótáralapú módszerek általában valamilyen szabadon hozzáférhető adatbázist használnak a névelemek azonosításához. Ezek az adatbázisok akár több millió entitás adatait tartalmazhatják, amelyeket folyamatosan bővítenek és frissítenek. Az adatbázisok hivatkozásokkal kapcsolódnak egymáshoz. Az egyes biológiai entitásokra valamely adatbázisbeli azonosítójával egyértelműen lehet hivatkozni. Ahhoz, hogy a szövegben a névelemek kisebb írásmódbeli eltérései ne befolyásolják azok felismerését, a szavak normalizált alakjait illesztik a szótárban található szinonimákhoz, ezzel sok olyan változatát is meg lehet találni a névelemeknek, amelyek a szinonimák között pontosan nem szerepelnek.

Több szótár, illetve ontológia alkalmazásával a rendszerek fedése nő, azonban a hibás jelölések növekvő száma problémát jelent. Ennek kiküszöbölésére többnyire szabályalapú vagy gépi tanulást alkalmazó utófeldolgozást használnak [4].

A szótáralapú rendszerekkel szemben az általánosan alkalmazott statisztikai névelem-felismerő módszerek rendszerint valamilyen szekvenciaalapú gépi tanuló algoritmust alkalmaznak [1, 2]. Ezek lényege, hogy egy előzetesen kézzel jelölt szöveg mondatait tokenekre bontják, és ennek a tokenláncnak a címkéit tanulják valamilyen szekvenciatanuló algoritmus segítségével. Az egyik legelterjedtebb ilyen algoritmus a feltételes véletlen mezők módszere [2]. Más megközelítések a potenciálisan névelemeket tartalmazó mondatrészek, illetve szószerkezetek azonosítását követően hozzárendelik a megjelölt kifejezésekhez legjobban hasonlító szótárelemeket [5].

3 Szótáralapú névelem-felismerés

A szótáralapú módszerek legfontosabb előnye, hogy az illesztett névelemek szótárbeli bejegyzésük alapján hozzárendelhetőek a névelem által jelölt entitásokhoz, függetlenül attól, hogy az entitás mely szinonimája fordult elő a szövegben.

Az így azonosított egyedekről további információk nyerhetőek ki más – az adott objektumról információkat tartalmazó – adatbázisokból. Földrajzi helyek esetében ilyen adatok lehetnek a hely pontos koordinátái vagy a közelben található nevezetességek, személyek esetén az életrajzi adatok, szervezetek esetében pedig azok földrajzi

helye, illetve kapcsolatai különböző személyekkel. A lehetőségeknek csak a rendelkezésre álló adatbázisok által tárolt információk mennyisége szab határt.

A szótáralapú névelem-felismerés egy, a névelemeket és azok szinonimáit tartalmazó lista segítségével történik. A névelemek lehetséges alakjait keressük a szövegben, rendszerint valamilyen normalizációt alkalmazva. A normalizáció segítségével a ragozott vagy más írásmódú alakok is felismerhetőek.

3.1 A szótárillesztő algoritmus

A biológiai névelemeket tartalmazó adatbázisok sokszor több tízmillió szinonimát tartalmaznak. Ennyi névelem normalizált szövegre illesztése igen időigényes, ezért az illesztéshez a Lucene[1] Java-alapú kereső és indexelő keretrendszer felhasználásával fejlesztettünk normalizált szótárillesztő rendszert.

A szótárban található szinonimák normalizált formáira indexet építettünk, majd a szöveg szavainak normálalakjait ebben kerestük. A keresés mondatonként történt. A mondat minden szavához hozzárendeltük azokat a lehetséges névelemeket, amelyek normálalakjában az adott token szerepelt. A következő lépésben a névelemek hossza alapján kiszűrtük azokat a tokensorozatokat, amelyek nem elég hosszúak az adott névelemhez. Ezzel jelentősen csökkent a lehetséges jelölések száma. A megmaradt, immár kezelhető mennyiségű jelöléssorozatot illesztettük a szövegre.

3.2 Angol biológiai névelemek

A biológiai névelemek egyértelmű, egyedi azonosítása elengedhetetlen a különböző bioinformatikai alkalmazások számára. Kiterjedt adatbázisok állnak rendelkezésre, amelyek tartalmazzák az ismert gének, fehérjék, fajok és egyéb biológiai entitások neveit, valamint kapcsolatait [8]. Az Entrez Gene egy géneket és azok szinonimáit tartalmazó adatbázis, ennek az elemeit használtuk jelen munkában is névelemek azonosítására [7]. Az általunk használt adatbázis közel 6 millió gén 7,9 millió szinonimáját tartalmazza.

A biológiai entitások neveinek normalizált formáit az LVG2010 programcsomagban [7] található szövegnormalizáló segítségével határoztuk meg. A biológiai entitások listáját a szövegben jelölt névelemekkel egészítettük ki annak érdekében, hogy egy kellően nagy méretű, a génnevek mellett fehérjeneveket is tartalmazó szótárt szimuláljunk.

A statisztikai módszerek gépi tanulásához és a kiértékeléshez a BioCreative II génnév-felismerési feladatának [6] teszt- és tanító dokumentumait használtuk. Az illesztés során a szöveg minden tokenjéhez hozzárendeltük azokat a névelemeket, amelyeknek a normált alakjai az adott szövegbeli tokent normalizálva tartalmazzák. Ha több egymás utáni tokent is megjelölt az illesztés egy adott névelemmel, akkor annak az összes token részsorozatát megvizsgáltuk, hogy az entitás teljes normált alakjához illeszkedik-e a tokensorozat. Az egyező sorozatokat megjelöltük.

[1] http://lucene.apache.org

3.3 Magyar névelemek

Az általunk használt magyar nyelvű névelemszótár *243 497* elemet tartalmazott. A névelemszótárat kiegészítettük a szövegben jelölt névelemekkel. Ezek illesztéséhez a névelemeket kisbetűssé alakítottuk, a szavakról a magyarlanc szófaji egyértelműsítő [9] segítségével eltávolítottuk az esetleges jeleket és ragokat, így előálltak a névelemek normalizált alakjai. A normalizáció segítségével a szótáralapú módszer például illeszteni tudta a *magukénak tudhatják a Magyarországon évente szétosztott* szövegrészben a *Magyarország* névelemet.

A mérésekhez a HVG cikkeit tartalmazó *144 507* token méretű dokumentumhalmazt használtuk[2]. A kiértékelő halmazt a dokumentumok 30%-a képezte, a fennmaradó részt a statisztikai rendszerek tanítására használtuk. A szövegeket a biológia névelemeknél alkalmazott szótárillesztő módszer segítségével jelöltük.

4 Statisztikai névelem-felismerés

A gépi tanulást alkalmazó rendszerek teljesítménye erősen függ a tanító adatbázis jellemzőitől, és általában gyengébb eredménnyel alkalmazhatóak más stílusú vagy más részterületet lefedő szövegeken. Előnyük, hogy olyan névelemeket is felismerhetnek, amelyek a rendelkezésre álló szótárakban nem találhatók meg. A szótáralapú módszer jelöléseit jellemzőként felhasználva a klasszikus statisztikai névelem-felismerők teljesítménye javul, de az így előálló annotációk már nem rendelkeznek a szótáralapú módszerek előnyeivel. Célunk olyan rendszer megalkotása volt, amely felveszi a versenyt a szótárjellemzőket használó klasszikus névelem-felismerőkkel.

Az általunk alkalmazott statisztikai névelem-felismerő rendszer a magyar, valamint az angol nyelvű névelem-felismerésben általánosan használt felszíni és nyelvi jellemzőket használta (részletesen l. [1]). A Mallet nevű programcsomag feltételes véletlen mezők módszerét használtuk a szekvenciák tanulására és predikálására [3].

A szótáralapú és statisztikai módszerek előnyeinek ötvözéséhez egy speciális szekvenciajelölési feladatot fogalmaztunk meg. A tanítóhalmaz dokumentumait a szótáralapú módszer segítségével jelöltük, majd az így keletkezett szótárjelölések három token sugarú környezetét véve statisztikai modellt tanítottunk a véletlen mezők módszerének használatával. Minden, a szótáralapú módszerrel megjelölt kifejezés környezete egy külön szekvenciát alkotott. A szekvenciák tartalmazhatták a szomszédos szótárjelölést is, amelyekről az adott láncban nem kellett döntést hozni. Ezeket eltérő címkével jelöltük meg. A névelemet jelentő címkesorozatot a statisztikai rendszer módosíthatta, így megváltoztatva a jelölést vagy annak határait.

A szekvenciajelölési feladatnak a szótárjelölések környezetére való korlátozásával a névelemtokenek aránya nagyobb az egyes tanító példányokban, mintha az egész mondat tokenláncán tanítanánk a statisztikai névelem-felismerőt. Az így felépített modell a szótárillesztés hibáit tanulja meg kiküszöbölni, és nem tanulja meg feleslegesen az olyan mondatrészek *nem névelem*ként való címkézését, amelyek nem tartalmaznak névelemet jelölő szavakat.

[2] http://www.inf.u-szeged.hu/rgai/corpus_ne

A tanító és a kiértékelő halmaz dokumentumaiban a szótárak minden előfordulását normalizáció alkalmazásával illesztettük. A normalizációhoz a szótárak leírásánál használt magyar és angol nyelvű módszereket használtuk.

A normalizált illesztés a szavak sorrendjét nem veszi figyelembe, ennek a biológiai névelemek illesztésénél van jelentősége, ahol a több tokenből álló névelemekben a szavak sorrendje gyakran változó, például a *G-protein coupled receptor family C group 5 member D* fehérje megnevezése lehet *5 member of G-protein coupled receptor family C group* is. Mivel a szótárak nem tartalmazzák az összes lehetséges írásmódot, a tokensorozat normalizálásakor a névelem tagjait rendszerint sorba rendezik, illetve bizonyos stopszavakat nem vesznek figyelembe az illesztésnél.

5 Eredmények

A különböző névelem-felismerő rendszerek által elért eredményeket a sztenderd F-mérték metrika alkalmazásával adjuk meg frázis- és tokenszinten. A frázisszintű kiértékeléskor csak a névelem minden tokenjének egyezése számított jó jelölésnek, míg a tokenszintű kiértékelés esetén minden tokenre megvizsgáltuk, hogy az automatikus jelölés egyezik-e a kézi címkével.

5.1 Biológiai névelemek felismerése

A szótáralapú módszer fedése a szótár kibővítésének köszönhetően majdnem teljes volt, de a pontosság csak a 0,25-ös értéket érte el. A szótár által illesztett szavak és azok határainak statisztikus javítása a fedést 0,15-del csökkentette ugyan, de a pontosság 0,86-ra nőtt így összességében a névelem-felismerés pontossága 0,40-ről 0,85-re nőtt. A kiértékelést frázis- és tokenszinten is elvégeztük, az eredményeket az 1. táblázat tartalmazza. Az általunk fejlesztett névelem-felismerő rendszer eredményei a szótár+CRF oszlopban találhatóak.

1. táblázat: A szótáralapú módszer és a statisztikai javítást alkalmazó biológiai névelem-felismerő összehasonlítása.

		szótár+CRF	**szótár**
FRÁZIS	Pontosság	0,860	0,252
	Fedés	0,838	0,979
	F-mérték	0,849	0,401
TOKEN	Pontosság	0,805	0,345
	Fedés	0,835	0,969
	F-mérték	0,819	0,509

Módszerünket klasszikus statisztikai névelem-felismerő módszerekkel hasonlítottuk össze. A 2. és 4. táblázatban *statisztikai szótárral* elnevezésű mérések során a statisztikai névelem-felismerő a szótáralapú illesztés jelöléseit jellemzőként használta. Azt tapasztaltuk, hogy a szótárjelölések alkalmazása egyaránt pozitívan befolyásolta a

statisztikai jelölés pontosságát, valamint fedését a szótár jelöléseit nem alkalmazó változathoz képest.

2. táblázat: A statisztikai módszerrel javított szótárillesztést alkalmazó és a hagyományos statisztikai biológiai névelem-felismerők eredményeinek összehasonlítása.

		szótár+CRF	statisztikai	statisztikai szótárral
FRÁZIS	Pontosság	0,860	0,715	0,745
	Fedés	0,838	0,643	0,678
	F-mérték	0,849	0,677	0,710
TOKEN	Pontosság	0,805	0,636	0,662
	Fedés	0,835	0,598	0,631
	F-mérték	0,819	0,617	0,646

A szótár+CRF alapú módszer a klasszikus megközelítésnél minden tekintetben jobban teljesített, így a névelem-felismerés F-mértéke 15 százalékponttal meghaladta a szótárat használó, illetve 17 százalékponttal a szótár nélküli klasszikus statisztikai megközelítés teljesítményét.

5.2 Magyar tulajdonnevek felismerése

A biológiai névelemekhez hasonlóan a magyar tulajdonnevek esetében is magas fedést tapasztaltunk a szótáralapú névelem-felismerő használatakor, azonban itt a módszer pontossága nem volt annyira alacsony, mint a biológiai névelemek esetén. A statisztikai javítás azonban a magyar tulajdonnevek esetében is jelentősen, 17 százalékponttal javította a felismerés F-mérték szerinti teljesítményét. A frázis- és tokenszintű kiértékelés eredményeit a 3. táblázat tartalmazza.

3. táblázat: A szótáralapú és a statisztikai javítás módszerét használó magyar tulajdonnév-felismerő összehasonlítása.

		szótár+CRF	szótár
FRÁZIS	Pontosság	0,981	0,695
	Fedés	0,952	0,957
	F-mérték	0,967	0,805
TOKEN	Pontosság	0,980	0,755
	Fedés	0,908	0,907
	F-mérték	0,943	0,824

Az angol biológiai névelemektől eltérően a magyar nyelvű szövegekben található tulajdonnevek felismerésénél a statisztikai módszerrel javított szótárillesztés nem ért el jobb eredményt a szótárt mint jellemzőt alkalmazó klasszikus statisztikai módszerhez képest. Ennek oka valószínűleg az, hogy míg a biológiai szövegekben a szótárak-

ban szereplő névelemek nagy része nem névelemként is előfordul, a cikkekben szereplő magyar tulajdonnevek általában egyértelműbbek voltak, így egyszerű jellemzőként felhasználva a klasszikus feltételes véletlen mezőket alkalmazó módszer is eredményesen tudta alkalmazni, anélkül, hogy a szótárillesztés miatti fedéscsökkenés negatívan hatott volna a teljesítményre.

A 4. táblázat eredményeiből látszik, hogy bár a pontosság tekintetében 10 százalékponttal felülmúlta a klasszikus megközelítés által elért eredményt a módszerünk, F-mérték szerinti teljesítménye fél százalékponttal kisebb volt, mint a klasszikus névelem-felismerőé.

4. táblázat: A statisztikai módszerrel javított szótárillesztést alkalmazó és a hagyományos statisztikai tulajdonnév-felismerők eredményeinek összehasonlítása.

		szótár+CRF	**statisztikai**	**statisztikai szótárral**
FRÁZIS	Pontosság	0,981	0,902	0,971
	Fedés	0,952	0,865	0,973
	F-mérték	0,967	0,883	0,972
TOKEN	Pontosság	0,980	0,895	0,960
	Fedés	0,908	0,834	0,937
	F-mérték	0,943	0,863	0,949

A szótárat használó és nem használó klasszikus statisztikai módszerek teljesítménye között 9 százalékpont a különbség.

6 Konklúzió

A biológiai névelemek esetében az általunk kifejlesztett hibrid megközelítés által elért eredmények alátámasztják, hogy egy kellően nagy fedésű szótár segítségével eredményesen és kellő pontossággal ismerhetőek fel biológiai entitások nevei, anélkül, hogy lemondanánk a szótáralapú módszerek előnyeiről.

A magyar névelemek esetében a klasszikus módszeren nem javít megközelítésünk aminek oka valószínűleg az, hogy itt a tulajdonnév/köznév többértelműség elenyészően kicsi.

A jövőben további biológiai adatbázisok bevonásával és a normalizációs módszerek javításával olyan hibrid névelem-felismerőt kívánunk fejleszteni, amely egyszerre több névelemtípus jelölését is el tudja végezni.

Köszönetnyilvánítás

A kutatást – részben – a BAROSS_DA07-DA_TECH_07-2008-0028 projekt támogatta.

Hivatkozások

1. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In:IV. Magyar Számítógépes Nyelvészeti Konferencia (2006)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (2001) 282–287
3. McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
4. Corbett, P., Milward, D.: Annotating Biomedical Entities with I2E using Multiple Ontologies. In: Proceedings of the First CALBC Workshop (2010)
5. Ando, R. K.: BioCreative II Gene Mention Tagging System at IBMWatson
6. Smith, L., Tanabe, L., Ando, R., Kuo, C. J., Chung, F. I., Hsu, C. N., Lin, Y. S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H. J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Lopez, M. M., Mata, J. Wilbur, J. W.: Overview of BioCreative II gene mention recognition. Genome Biology Vol. 9 Suppl. 2. (2008)
7. The NCBI handbook. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2002) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.
8. Torii, M., Hu, Z., Wu, C. H., Liu, H.: BioTagger-GM: A Gene/Protein Name Recognition System. Journal of the American Medical Informatics Association Vol. 16 No. 2. (2009) 247–255
9. Zsibrita J., Nagy I., Farkas R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009)

Klaszterek helyett prototípusok

Kálmán László[1], Rung András[1,2]

[1] MTA Nyelvtudományi Intézet, Elméleti Nyelvészet Tanszék, Benczúr utca 33.,
1068 Budapest, Magyarország
[2] BME Fizikai Intézet, Budafoki út 8.,
1111 Budapest, Magyarország
kalman@nytud.hu
rungandras@gmail.com

Kivonat: Írásunkban bemutatjuk, hogy nyelvi elemek viselkedésének jellemzése és modellezése lehetséges klaszterekre való hivatkozás nélkül prototípusok segítségével is. Vizsgálatunkban gépileg kiválasztott prototípusok segítségével a hangkivető főnevek ingadozását modelleztük eredményesen. 282 hangkivető főnévből választottunk ki 8 prototípusnak tekinthető szót. Az egyes szavak és a hozzájuk alakjában leghasonlóbb prototípus közt mérhető távolság szignifikáns pozitív együttjárásban ($r(280) = 0{,}419$, $p < 0{,}001$) van a viszonyított szavak hangkivetési mértékével a Szószablya Gyakorisági Szótár [3] adatai alapján. Ebből láthatjuk, hogy azok a szavak, amelyek a prototípusokra jobban hasonlítanak hangalakjukban, azokhoz közelítő módon is viselkednek, azaz az egyes szavak viselkedését klaszterekre és szabályokra való hivatkozás nélkül is modellezni tudtuk.

1 Bevezetés

A statisztikai alapú számítógépes nyelvészetben (így pl. a korpusznyelvészetben és az automatikus nyelvtanindukcióban) fontos szerepe van az egymáshoz hasonlóan viselkedő egységek felfedezésének és csoportosításának, vagyis a klaszterezésnek. Vannak azonban olyan nyelvészeti feladatok, amelyeknél nem annyira magukra a klaszterekre, hanem az őket legjobban képviselő elemre van szükségünk. Ilyen például az esetalapú, példányalapú vagy általában analógiás okoskodás használata a számítógépes nyelvészetben. Ennél a fajta okoskodásnál olyan elemet keresünk, amely bizonyos szempontokból a lehető leghasonlóbb egy adatbázisban lehetőleg minél nagyobb gyakorisággal szereplő elemekhez. (Az adatbázis a beszélő korábbi nyelvi tapasztalatait kívánja ábrázolni.)

Az ilyen elven működő algoritmusokban nem annyira a korábbi elemek hasonlósági osztályai (klaszterei) játszanak szerepet, mint maguk azok az elemek, amelyek ezeket az osztályokat mind gyakoriságuk, mind tulajdonságaik alapján a legjobban képviselik, középponti szerepet játszanak bennük, vagyis prototipikusak.

A statisztikai megközelítésekben a klaszterek prototípusának fogalmát úgy szokták értelmezni, hogy az a klaszter ún. centroidjához (súlypontjához) legközelebb eső elem. Ennek a prototípus-fogalomnak azonban több hátránya is van. A legfontosabb

az, hogy a klaszter prototípusának meghatározásához először is magát a klasztert kell meghatározni, ennek a folyamatnak minden nehézségét le kell küzdeni, hiszen — ebben az esetben teljesen szükségtelenül — döntést kell hozni a klaszter határának kérdésében. A másik probléma, hogy a súlyponthoz legközelebb eső elem nem feltétlenül a klaszter legsűrűbb részére esik.

Írásunkban egy teljesen más megközelítést javasolunk: azokban a feladatokban, amelyekben a prototípusokra szükség van, de magukra a klaszterekre nem, olyan algoritmusokat is alkalmazhatunk, amelyek közvetlenül a prototípusok megtalálására irányulnak, maguknak a klasztereknek a határait pedig nem próbálják meghúzni [10]. Az általunk javasolt algoritmus a következő egyszerű alapfeltevéseken alapul:

- A klaszter prototipikus eleme legyen minél gyakoribb [1, 6].
- A klaszter prototipikus elemének közelében minél több minél gyakoribb elem legyen, vagyis sok elem hasonlítson rá.
- A klaszter prototípusa minél távolabb legyen, minél kevésbé hasonlítson más klaszterek prototípusaira.

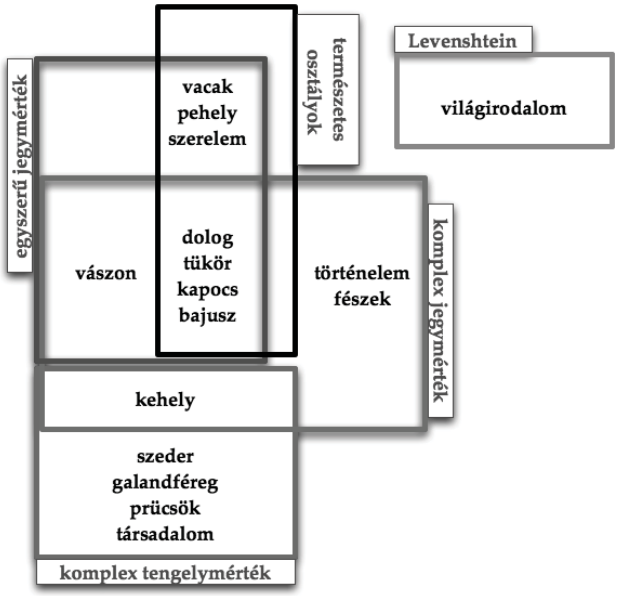
Megmutatjuk, hogy ezeknek a kritériumoknak az alapján viszonylag egyszerű algoritmusokkal hatékonyan megtalálhatóak a klaszterek prototípusai, amelyek nagyjából egybeesnek a hagyományos meghatározás szerinti prototípusokkal, de a módszer egyes nyelvészeti feladatokban talán még kedvezőbb eredményekhez is vezet, mint a hagyományos megközelítés.

2 Prototípusok gépi meghatározásának módja

A prototípusok kiválasztása során saját fejlesztésű algoritmusokkal számítjuk a hasonlóságot, amelyek a kurrens, hasonlóság mérésére használt algoritmusoknál (pl. [9]) finomabb összehasonlításokat is lehetővé tesznek. A komplex jegymérték és a komplex tengelymérték nevű algoritmusok a szavak hasonlóságát azok jobb szélétől véve számítják ki úgy, hogy a megfeleléseknek, hasonlóságoknak egyre kisebb súlyt adnak a szavak bal széle felé haladva. Így mind a két számítógépes algoritmus a *vas* és *sas* szavakat hasonlóbbnak tekinti, mint a *vas* és a *vaj* szavakat. Az algoritmusok a hasonlítást az egyes fonémák jegyei alapján végzik el, de a komplex jegymérték [7, 8] fonémákat hasonlít össze, míg a komplex tengelymérték az egyes jegyek tengelyeinek hasonlósága alapján számítja ki két szó hasonlósági értékét. Ezeket az értékeket egy 0-1 terjedő skálán adtuk meg.

További összehasonlítási eljárásunk a fonológiai természetes osztályokon alapszik, amellyel a komplex jegymértékhez hasonlóan vetettük össze a szavakat, csak két fonéma hasonlóságát annak révén határoztuk meg, hogy hány közös és hány eltérő természetes fonológiai osztályban szerepelnek ezek [2]. Összehasonlításainkban összesen 13 fonológiai jegyet vettünk figyelembe. Egyedül a komplex jegymérték egyszerűsített variánsa (egyszerű jegymérték) esetében alkalmaztunk 8 jegyet. Saját algoritmusainkat a közismert és általánosan szó-összehasonlításra is használt Levenshtein-algoritmus [5] teljesítményéhez mértük.

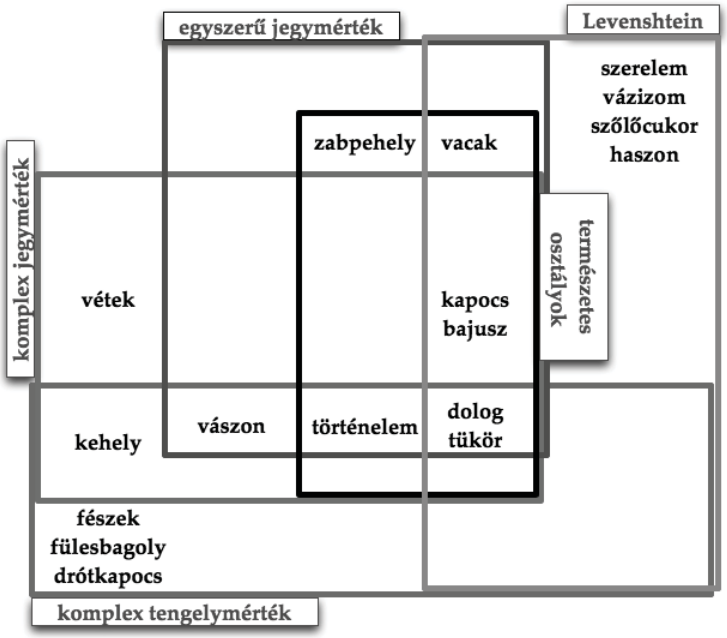
Kísérletünkben magyar hangkivető főnevek ingadozását (pl. *sátrat* : *sátort,* de *szerelmet* : **szerelemt*) modelleztük úgy, hogy ingadozásuk mértékét a legközelebbi (leghasonlóbb) prototípushoz való hasonlóság alapján határoztuk meg. Korábbi elemzések tapasztalatai alapján [4, 7, 8] a prototípusok kiválasztására egy olyan algoritmust hoztunk létre, amely a bevezetőben megadott kritériumok alapján működik. A prototípuskiválasztásban többféle küszöbértéket is megadhatunk, amelynek növelésével algoritmusunk egyre szigorúbban alkalmazza a hasonlósági szempontokat, miszerint a prototípushoz sokan hasonlítanak, de az más prototípusokra nem hasonlít. Két eltérő küszöbérték mellett kiválasztott prototipikus szavainkat a 1-2. ábrák mutatják meg.



1. ábra. 0,9-es értékkel növelt küszöbérték[1] mellett kiválasztott prototípusok[2].

[1] A teszt ismertetése szempontjából nem fontos, hogy a küszöbérték-növelések értéke pontosan hogyan járul hozzá az algoritmus működéséhez. Az ábrák értelmezéséhez csak annyit kell tudnunk, hogy minél magasabb ez az érték, az algoritmus annál szigorúbban alkalmazza a kiválasztásban a hasonlósági kritériumainkat.

[2] Az ábrákon Venn-diagramokat láthatunk az Edwards-féle módosításban, ami lehetővé teszi öt halmaz elemeinek is az összehasonlítását. A halmazok megjelenítését tartalmuk függvényében átalakítottuk a könnyebb áttekinthetőség érdekében.



2. ábra. 0,5-ös értékkel növelt küszöbérték mellett kiválasztott prototípusok.

Mielőtt áttekintenénk, hogy az egyes prototípusok mennyire jól modellezték a hangkivető szavak hangkivetésének mértékét, érdemes őket szemügyre venni. Minden mérték esetében jellemző a gyakori alakok preferenciája. Ez legszembetűnőbben a *dolog* mint prototípus választásában jelenik meg, mivel az összes hangkivető előfordulás mintegy 16,1%-át teszi ki (348 ezer egyes szám alanyesetű előfordulás a *Szószablya Gyakorisági Szótár*ban), és 2,42-szer gyakoribb, mint az őt közvetlenül követő *társadalom*. A választásokban további nagyon gyakori szavak is szerepelnek még: *szerelem* (68 ezer)*, társadalom* (144 ezer)*, történelem* (68 ezer). A *dolog*-gal együtt ezek már az összes hangkivető főnév alanyesetű előfordulásainak 29,1%-át fedik le. E kiugróan gyakori elemeken túl azonban a prototípusválasztó algoritmus inkább a hasonlósági szempontokat veszi figyelembe, hisz a következő leggyakoribb szó, a *tükör* (29., 21 ezer) már jóval elmarad ezek mögött. Az összes mértéken alapuló választásnál megfigyelhető, hogy habár a gyakori *-alom, -elem* végűek alkotják a legszámosabb alcsoportját a hangkivető szavaknak, mégis ezek vannak leginkább alulreprezentálva a prototípusok tekintetében. Általában az egyes prototípuscsoportokban csak *-elem* végű prototípus jelenik meg, ami egyaránt jól lefedi az *-alom* végűeket és a többi *-e-e* végű szót is.

A prototípusválasztó algoritmus azonban kevésbé gyakori szavakat is választ, ha azok a hasonlósági kritériumoknak jobban megfelelnek. Ezek gyakran összetett szavak, hisz hosszúságuk alapján jobban reprezentálják a zömükben összetett hangkivető szavakat, mint a példánygyakoriságban gyakoribb, de típusgyakoriságban ritkább alapszavak. Ilyen szavak a *zabpehely* (egyszerű jegymérték, természetes osztályok)*, szőlőcukor, vázizom* (Levenshtein-algoritmus)*, drótkapocs, galandféreg, fülesbagoly,*

(komplex tengelymérték). Kisebb, de jól elkülönülő szócsoportok is több esetben kapnak önálló prototípust: *vacak, bajusz* (utolsó magánhangzó nem középső nyelvállású), *vászon* (*-(á/ó)CVC* végűek), *zabpehely, kehely* (hangátvetés), *vázizom, pityer* (*-iCVC* végűek).

Az egyes hasonlósági mértékek és a két eltérő küszöbérték mentén kiválasztott prototípusokat az olyan hangkivető főnévekkel hasonlítottuk össze, amelyek legfeljebb 99,99%-ban mutattak hangkivető viselkedést (282 szó) az olyan toldalékokkal, amelyek esetében hangkivető alakokat várnánk el. Vizsgálatunkból azért zártuk ki a 100%-ban hangkivető főneveket, mert ezek esetében legfeljebb csak a kiugróan gyakoriaknál tudhatjuk, hogy az ingadozás hiányának oka következetes viselkedésük, és a 100%-ban hangkivető viselkedés nem adathiánynak tudható be. A prototípusokhoz az egyes szavakat mindig olyan mérték alapján hasonlítottuk, amilyen mértéket a prototípus kiválasztásában is alkalmaztunk. Miután minden, a vizsgálatra kiválasztott hangkivető főnevet minden prototípuscsoporttal (2 x 5 db) összehasonlítottuk, megvizsgáltuk, hogy az egyes szavak hangkivetési mértéke[3] mennyire korrelál a hozzá legközelebbi prototípushoz való hasonlóságával.

Feltételezésünk szerint egy szó minél jobban hasonlít a hozzá leghasonlóbb prototípushoz, annál nagyobb a hangkivetési mértéke is. Az együttjárások számítása során a prototípushoz való hasonlósági értéket súlyoztuk a hasonlítandó hangkivető főnév releváns toldalékos alakjai alapján meghatározott gyakoriságának 8. gyökével (pl. *dolog* esetében 5,23, a *sátor*-nál 3,34), mivel a gyakoribb főneveknél magasabb hangkivetési mértéket vártunk, de nem kívántunk ennek az értéknek túlzott súlyt sem adni. Az 1-2. ábrákon bemutatott prototípusokon túl a szavakat hasonlítottuk a *Szószablya Gyakorisági Szótár*ban az egyes szám alanyesete alapján 50 leggyakoribb hangkivető főnévhez is, mint olyan prototípusokhoz, amelyeket kizárólag gyakoriságuk alapján választottunk ki a hasonlósági szempontok figyelmen kívül hagyásával. Gyakorisági prototípusnak azért választottunk ki viszonylag több szót, mert a 10 leggyakoribb hangkivető főnévből 8 *-alom/-elem* végű volt, így ennél több szóra volt szükségünk ahhoz, hogy ne csak az *-alom/-elem* csoporthoz való hasonlóságot mérjük. A prototípusok számának növelése nem jár szükségszerűen együtt a korreláció mértékének növelésével, hisz ha az összes hasonlítandó szót felvennénk prototípusnak, akkor az önmagukhoz való hasonlóságuk 1 lenne, aminek következtében egyáltalán nem tudnánk érdemleges együttjárásokat megfigyelni a változó hangkivetési mértékek és a konstans 1-es értékek közt.

3 Kísérletünk eredményei

Az 1. táblázat alapján láthatjuk, hogy – a Levenshtein-algoritmust leszámítva – már az összes legközelebbi prototípushoz való hasonlóság közepesen korrelál a szavak hangkivetési mértékével. A komplex tengelymérték a legmagasabb együttjárást mu-

[3] Hangkivetési mérték alatt az értjük, hogy a hangkivetéssel együttjáró toldalékok (pl. tárgy, szuperesszívusz, birtokos ragok stb.) esetében mennyire stabilan jelentkezik a hangkivetés. Így ez az érték az *–alom* végű szavaknál többnyire 100%, a *sátor* esetében 81%, míg a *bajusz*-nál csak 36%.

tatja. A hangkivetés mértékét legjobban megragadó prototípusaink: a *dolog, történelem, tükör, vászon, fészek, kehely, fülesbagoly, drótkapocs*. Ez a néhány szó viszonylag jól fedi a lehetséges végmintázatokat és záró magánhangzó-szekvenciákat is, amelyek a viselkedés szempontjából a legfontosabbak lehetnek. A *fülesbagoly* és a *drótkapocs* a nagyszámú összetett szót, a *kehely* egy speciális mintát, a *vászon* pedig a mérsékelten hangkivető szavak csoportját képviseli.

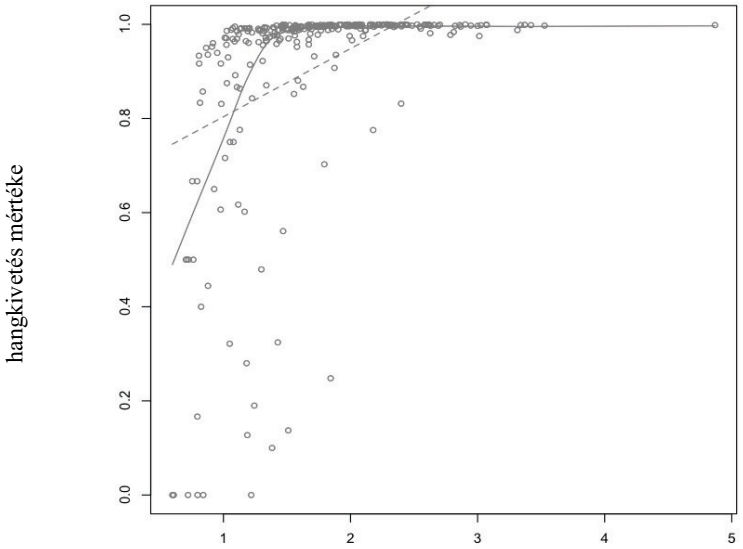
1. táblázat: A hangkivetési mérték és a prototípushoz való hasonlóság együttjárásának mértéke a felhasznált prototípusok függvényében.

	Levenshtein	**Egyszerű jegyek**	**Komplex jegyek**	**Természetes osztályok**	**Komplex tengelymérték**
0,5 küszöbérték	0,241***	0,352***	0,364***	0,371***	0,419***
0,9 küszöbérték	0,248***	0,352***	0,362***	0,370***	0,409***
gyakori szavak	0,346***	0,458***	0,461***	0,455***	0,423***

** = $p < 0,01$
*** = $p < 0,001$

Mindösszesen ennek a 8 szónak az alapján szabályoknál hatékonyabb és könnyebb módon 274 másik szó viselkedését tudjuk viszonylagos megbízhatósággal jellemezni. Az eredetileg csak viszonyítási alapnak szánt 50 leggyakoribb szóhoz való hasonlítás alapján azonban láthatjuk, hogy a gyakoriságnak van a legkiugróbb szerepe a szavak viszonyrendszerében. Ha csak a számunkra fontos szavak 20%-ához van gyors hozzáférésünk, már akkor egészen jól tudjuk leírni a maradék 80% viselkedését. Ha szavainkat a komplex jegymérték alapján azonosítható hasonlósági csoportok[4] leggyakoribb szavaihoz hasonlítjuk a komplex jegymértékkel, akkor ismét közepesen erős korrelációt tudunk kimutatni ($r(280) = 0,4$, $t = 7,31$, $p > 0,001$). Ez alapján láthatjuk, hogy ha a gyakoriságot lokálisan értelmezzük egy adott csoporton belül, akkor is képesek vagyunk az egyes szavak hangkivetési mértékével kapcsolatban együttjárásokat megfigyelni. Ha komplex jegymérték (0,5-ös küszöbérték) által kiválasztott prototípusainkból és a leggyakoribb szavakból alkotott csoporthoz hasonlítjuk hangkivető szavainkat, akkor némileg még szorosabb együttjárást ($r(280) = 0,485$, $t = 9,27$, $p < 0,001$) figyelhetünk meg a halmaz szavaiból kiválasztott leghasonlóbb prototípusok hasonlóságértéke és a hangkivetési mértékek közt. Ebből arra következtethetünk, hogy ha a prototípus kiválasztásában alkalmazott szempontjainkat még jobban optimalizálnánk, akkor a hangkivetési mértéket vagy akár bármilyen más viselkedési mutatót jobban tudnánk megragadni.

[4] Az összes hangkivető főnév viszonyait megragadó hasonlósági gráfban 50 hasonlósági csoportot tudunk azonosítani, ha csak a legszorosabb kapcsolatokat vesszük figyelembe.



3. ábra. Komplex tengelymérték leghasonlóbb prototípusaihoz való hasonlóság és a hangkivetési mérték összefüggése a *Szószablya Korpusz*ban.

Bibliográfia

1. Bybee, J. L., Eddington, D.: A usage-based approach to Spanish verbs of 'becoming.' Language Vol. 82 (2006) 323–355
2. Frisch, Stefan A.: Similarity and Frequency in Phonology. PhD-disszertáció (1996) http://www.cas.usf.edu/~frisch/Frisch96.pdf (2010.07.01.)
3. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: A Szószablya projekt. In: Alexin Z., Csendes D. (szerk.): Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). (2003)
4. Kálmán L., Rebrus P., Törkenczy M.: Lehet-e az analógiás nyelvelmélet szinkrón? A magyar nyelvészeti kutatások újabb eredményei II., Kolozsvár. 2010. április 16. http://budling.nytud.hu/~tork/KRT/bbte10_slides_print.pdf (2010.07.01.)
5. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics - Doklady Vol.10, No. 8 (1966) 707–710
6. Nosofsky, R. M. Exemplar based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition Vol. 14 (1988) 700–708
7. Rung A.: Determining word similarity in the Hungarian language. In: Kálmán L. (szerk.): Papers from the Mókus Conference. Tinta Kiadó, Budapest (2008) 112–118
8. Rung A.: Szóhasonlóság mérése analógiás megközelítésben. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2009. Szegedi Tudományegyetem, Szeged (2009) 104–113

9. Skousen, R., Lonsdale, D., Parkinson, D. B. (szerk.): Analogical Modeling. John Benjamins, Amsterdam (2002)
10. van den Bosch, A.: Expanding k-NN analogy with instance families. In: Skousen, R., Lonsdale, D., Parkinson, D. B. (szerk.): Analogical Modeling. John Benjamins, Amsterdam (2002) 209–223

Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel

Recski Gábor

MTA SZTAKI
Nyelvtechnológiai Kutatócsoport
e-mail: recski@sztaki.hu

1. Bevezetés

Cikkünkben először egy, a magyar főnévi csoportok azonosítására épített mondattani elemzőt (parsert) mutatunk be, melynek pontossága megközelíti a hasonló célú, gépi tanuláson alapuló rendszerünk (`hunchunk`, Recski et al. [8]) eredményeit, majd megmutatjuk, hogy a két eszköz egyesítésével a feladat nagyobb pontossággal végezhető, mint az önálló statisztikai alapú rendszerrel.

Először röviden bemutatjuk a megoldandó feladatot és a témában született legfontosabb eredményeket, majd a 3. fejezetben a parser építésének főbb lépéseit ismertetjük. A szabályalapú rendszer létrehozásánál kiindulási pontként Kornai környezetfüggetlen NP-nyelvtana [4,5] szolgált, melyet az NLTK nyelvtechnológiai programcsomag [2] segítségével implementáltunk. A Szeged Treebank [3] alapján készült NP-korpusz tette lehetővé, hogy a parsert kiértékeljük, a munka egyes fázisaiban a legnagyobb hibaosztályokat elkülönítsük, és a szabályrendszert ezek figyelembevételével fejlesszük – a 4. fejezet ezt a folyamatot írja le.

Végül az 5. fejezetben egy hibrid rendszert mutatunk be, mely a parser kimenetét használva gépi tanulási módszerrel végzi a magyar NP-k azonosítását, majd megmutatjuk, hogy két eszközünket egyesítve magasabb pontossággal tudjuk elvégezni a feladatot, mint a tisztán statisztikai alapú rendszerrel.

2. Előzmények

A szakirodalomban leggyakrabban *NP-chunking*-nak nevezett feladatnak Abney [1] alapján Marcus [6] adta azon definícióját, melyet felhasználva a feladat a nyelvtechnológiában használatos gépi tanuló algoritmusok egyik mércéjévé vált (l. például a CoNLL-2000 versenyt [9]). Magas pontossággal végezték a feladatot többek között a Support Vector Machine (SVM), a Conditional Random Fields (CRF) vagy a Maximum Entropy Markov Model (MEMM) módszerrel is. Az utóbbi módszertől némiképp eltérően használ maximum entrópia tanulást és rejtett Markov-modelleket saját NP-címkézőnk, a `hunchunk`, mely a feladatot magyar nyelvű szövegen is nagy pontossággal végzi ($F_2 = 94.75\%$). A főnévi csoportok azonosításával kapcsolatban egy másik feladat, a maximális NP-k azonosítása is megfogalmazható, mely egyes feladatokhoz, így különösen a frázisalapú gépi

fordításhoz fontos bemenetet szolgáltat. A `hunchunk` ezt a feladatot is $89-91\%$ közötti F-pontszámmal végzi a különböző tesztadatokon.

Bár a feladat gépi tanulási módszerekkel nagy pontossággal oldható meg, feltételeztük, hogy egy kézzel írt szabályokon alapuló mondattani elemző (parser) a magyar főnévi csoportokat hasonló vagy magasabb pontossággal is képes azonosítani. Első kísérleteink megmutatták, hogy a mondatszintű nyelvtan hiánya nem teszi lehetővé, hogy az egyébként előnyben részesített „maximális NP" feladatot egy parser magas pontossággal megoldja, így a hagyományos, minimális NP-kre irányuló feladat minél nagyobb pontosságú teljesítését tűztük ki célul, bízva abban, hogy az így születő elemzések a statisztikai rendszer teljesítményén is javíthatnak.

3. A parser építése

3.1. Formalizmus

A magyar NP-parser létrehozásához Kornai magyar NP-nyelvtanát használtuk. Az implementációhoz az NLTK programcsomagot választottuk, mivel különböző mondatelemző algoritmusok széles választékát támogatja és jegystruktúrákra hivatkozó nyelvtani szabályok használatát is lehetővé teszi. A nyelvtan szabályai ugyanis nem csupán a szavak szófajára, hanem azok számos morfológiai jegyére is képesek hivatkozni. A `hunmorph` morfológiai elemző [10] lehetővé teszi, hogy az elemezni kívánt mondatok szavairól valamennyi ilyen információt kinyerjünk. A `hunmorph` a KR-kód [7] szerinti morfológiai elemzést ad, ezek pedig maguk is jegy-struktúrák, így mechanikusan alakíthatóak át a parser bemeneti formalizmusának megfelelő formátumra – ezáltal a morfológiai elemző valamennyi lehetséges kimenete megfelel a nyelvtan egy terminális szimbólumának. A nemterminális szimbólumok egy kategóriacímkéből (`NOUN`, `ADJ` stb.) és a hozzátartozó jegy-érték struktúrából állnak.

A jegyek értéke lehet sztring, egész szám, újabb jegystruktúra, más nemterminális szimbólum és a jegy értékét egy másik jegyértékhez kötő változó. Így például egy az ige és tárgy közötti számbeli egyezést kifejező szabály `VP -> V[PL=?a] N[PL=?a]` formátumú, mely egyenértékű a megszokottabb, görög betűket használó jelöléssel: `VP -> V[`α`PL] N[`α`PL]`. A nyelvtan a projekciós szinteket nem különböző szimbólumokkal, hanem a `BAR` jeggyel kódolja; így a `NOUN[BAR=0]` például egy puszta főnevet jelöl. A morfológiai elemzés a képzett szavak esetében megadja a képzés forrását és típusát is; ezeket az `SRC` jegy kódolja, melynek értéke két további jegy, a tövet tároló `STEM` és a képzés típusát tartalmazó `DERIV`. Néhány példa az NLTK formalizmusa és a KR-kódolás közti megfeleltetésre (1)-ben látható.

(1a) `NOUN[POSS=[1=1, PLUR=1] -> NOUN<POSS<1><PLUR>>`
(1b) `NOUN[POSS=1, CAS=[SUE=1]] -> NOUN<POSS><CAS<SUE>>`
(1c) `NUM[CAS=[INS=1], SRC=[STEM=NUM, DERIV=ORD]] ->`
`-> NUM[ORD]/NUM<CAS<INS>>`
(1d) `VERB[SUBJUNC-IMP=1, PERS=[1=1], PL=1, D=1] ->`
`-> VERB<SUBJUNC-IMP><PERS<1>><PLUR><DEF>`
(1e) `NOUN[POSS=1, SRC=[STEM=VERB[SRC=[STEM=VERB, DERIV=MEDIAL]],`
`DERIV=GERUND]] -> VERB[MEDIAL]/VERB[GERUND]/NOUN<POSS>`

Noha a KR-kódok túlnyomórészt egyértelműen megfeleltethetőek ennek a reprezentációnak, néhány átalakítást mégis szükséges volt elvégezni. A KR-elemzés tartalmaz privatív jegyeket, például egy főnév egyes számát a `<PLUR>` jegy hiánya 'jelöli'. Mivel szeretnénk, hogy a nyelvtan hivatkozhasson ezen jegyek hiányára, így az NLTK formátumában ezek a jegyek binárisak: a harmadik személyt, az egyes számot és a nominatív esetet rendre a `PERS=0, CAS=0, PLUR=0` jegyek jelölik.

3.2. Főnévi csoportok azonosítása

Mivel a nyelvtan csupán a főnévi csoportokat felépítő szabályokat tartalmaz, így teljes mondatelemzést nem készíthetünk: az elemzés során bottom-up módszerrel azonosítjuk az olyan szósorozatokat a mondatban, melyet az NP-nyelvtan elfogad. A parser kimenete egy táblázat, melyből minden ilyen szósorra kiolvasható, hogy mely szabályok mely szósorokra való alkalmazásával épült fel az adott főnévi csoport. A mondatszintű nyelvtan hiánya azt jelenti, hogy nem zárhatunk ki lehetséges elemzéseket azért, mert azokból később nem építhető teljes mondatelemzés. Ehelyett néhány egyszerű szabállyal választjuk ki a végleges chunkolást, azaz a diszjunkt NP-szekvenciákat.

Kiindulópontként vesszük valamennyi, a parser által NP-ként felismert szósorozatot. Mivel minimális NP-ket keresünk, első lépésként kizárjuk azokat a sorozatokat, melyek egynél több főnevet tartalmaznak. Ezután, mivel a főnévi csoportok legmagasabb projekcióit keressük, kizárjuk azokat a jelölteket, melyeket egy másik intervallum tartalmaz. Következő lépésként az egymással átfedő jelöltek közül azokat választjuk, melyeket más kategóriájú frázisként a nyelvtan nem ismer fel – ezt az eljárást elsősorban az indokolja, hogy az elliptikus NP-ket engedélyező ún. SLASH-szabályok gyakran tévesen ismerik fel főnévi csoportként a mondat melléknévi és számnévi csoportjait. Ha ez az eljárás nem vezet sikerre – azaz mindkét vagy egyik jelölt sem áll elő más kategóriájú frázisként, akkor egyiket sem tartjuk meg.

4. A nyelvtan fejlesztése

A parsert a Szeged Treebank alapján készült NP-korpusz segítségével értékeltük ki. Az eredeti NP-nyelvtan 81.76%-os F-pontszámot ért el az 1000 mondatból álló tesztkorpuszon. A parser teljesítményét a nyelvtanon végrehajtott minden

változtatás után újra megmértük, a tévesen elemzett mondatok szemrevételezésével pedig elkülönítettük a mindenkori legnagyobb hibaosztályokat.

Az első mérés során a legtöbb hibát nem a nyelvtan hiányossága okozta, hanem az a tény, hogy melléknévi és számnévi csoportokra hivatkozik, melyekre nem adtunk nyelvtant. Így első lépésként néhány egyszerű szabályt adtunk hozzá a nyelvtanhoz:

(2a) `ADJ -> ADJ ADJ`
(2b) `ADJ -> ADV ADJ`
(2c) `NUM -> NUM NUM`
(2d) `NUM -> ADV NUM`
(2e) `NUM -> ADJ NUM`

Ezáltal a rendszer F-pontszáma 84.18%-ra nőtt.

A következő nagy hibaosztályt a névmások jelentették. Mivel a magyar névmások a főnevekhez nagyon hasonló módon viselkednek, így sem a nyelvtan, sem a `hunmorph` elemző nem tesz köztük különbséget. Így mind a KR-formalizmust, mind az NP-nyelvtant kibővítettük a `PRON` jeggyel. Egyes esetekben szükségesnek bizonyult a névmások típusára is hivatkozni, így a jegy értékeként ezt is megadtuk. A két kategória egybeejtése ugyan általában megalapozott, mégis ez a lépés tette lehetővé, hogy a nyelvtan kezelje az általános és határozatlan névmásokat tartalmazó főnévi csoportokat (l. 3).

(3a) *minden pofon*
(3b) *néhány villanykörte*

Ezeket a névmásokat a főnevek elsőszintű projekcióihoz csatoltuk, így a fenti esetek kezelésére a nyelvtant az alábbi szabályokkal bővítettük:

(4a)
```
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
-> NOUN[PRON=GEN] NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c,
CAS=?d, D=?e, PRON=?f]
```
(4b)
```
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
-> NOUN[PRON=INDEF] NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c,
CAS=?d, D=?e, PRON=?f]
```

Ez a módosítás az F-pontszámot 85.45-ra növelte.

Külön kezelést igényelt a mutató névmás egy speciális esete, melyre (5)-ben adunk példát:

(5a) *ez a pincér*
(5b) *ezek a hajók*
(5c) *attól a pasastól*

Ezeket a szerkezeteket a (6)-beli szabállyal kezeljük, ezzel újabb egy százalékpontos javulást érve el az elemző teljesítményében.

(6) `NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e] ->`
`-> NOUN[PRON=DEM, BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d]`
`ART NOUN[PRON=0, BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d,`
`D=0],`

A következő jelentős hibacsoport olyan, főnevet módosító melléknévi csoportokhoz kötődött, melyeknek feje egy igéből képzett melléknév és tartalmazzák az eredeti ige egy argumentumát is (l. 7).

(7a) *a korsónak támasztott könyvet* olvasta
(7b) *az ókori mór hódítóktól származó esküvést* hallották

Mivel a nyelvtan terminális szimbólumai kódolják a képzési információt, így lehetőség nyílt ezeket az eseteket a melléknévi csoportok nyelvtanába felvenni. A (7a) és (7b) alatti szerkezeteket rendre a (8a) és (8b) alatti szabályok kezelik.

(8a) `ADJ -> NOUN ADJ[SRC=[STEM=VERB[], DERIV='PERF_PART']]`
(8b) `ADJ -> NOUN ADJ[SRC=[STEM=VERB[], DERIV='IMPERF_PART']]`

Ez a módosítás a rendszer teljesítményét 87.87%-ra növelte. Mindezek után a parser hibáinak legnagyobb részét már a valóban kétértelmű szerkezetek okozták, azonban még számos hibát okozott az írásjelek és kötőszavak téves elemzése. Néhány, az ilyen elemek NP-beli és NP-környéki viselkedését leíró szabály a pontosságot további másfél százalékponttal növelte.

A nyelvtan végső változata, mely a cikk végén teljes egészében olvasható, a tesztadaton 89.36%-os F-pontszámot ér el. A teljesítmény javulását az egyes lépések függvényében az 1. táblázat foglalja össze. A fennmaradó hibák egy része mögött valódi szerkezeti kétértelműség áll, leggyakrabban azonban olyan hibaosztályokat találunk, melyek a korpusz sajátosságaiból fakadnak, így például gondot okoz egy-egy szövegtípusra jellemző speciális írásjelek szokatlan tokenizálása. Mivel a munka ezen fázisában egy-egy hasonló jelenség már a hibák nagyobb százalékáért felelős, mint bármelyik kezeletlen nyelvi szerkezet, úgy hisszük, hogy a pontosság további növelésének – mind a szabályalapú, mind a gépi tanuló eszköz esetében – feltétele a további tanuló- és tesztadat.

1. táblázat. A nyelvtan fejlesztésének lépései és hatásuk az elemzés pontosságára

Fejlesztés	F-pontszám
Kornai 1985	81.76%
AdjP, NumP	84.18%
Névmások	85.45%
„Ez a” szerkezet	86.68 %
Deverbális melléknevek	87.87%
Írásjelek és kötőszavak	89.36%

A fenti eredmények alapján elmondhatjuk, hogy a nyelvtanba újonnan felvett szabályok az eredeti rendszer hibáinak közel felét kiküszöbölik. Az elért eredmény jelentősen elmarad a statisztikai alapú `hunchunk`nak a minimális NP-k azonosításán elért eredményétől (94.75%), de lehetővé teszi, hogy megpróbálkozzunk hibrid rendszer kifejlesztésével.

5. Hibrid megoldás

A hibrid rendszer lényege, hogy a gépi tanulásra alapuló `hunchunk` rendszer tanításakor felhasználjuk a szabályalapú elemző kimenetét. Miután a parser a 3. fejezetben leírtak szerint elkészíti a mondat NP-chunkolását, azaz diszjunkt főnévi csoportokat jelöl meg, az egyes szavakat olyan címkékkel látja el, mint amilyenekre a statisztikai rendszert tanítottuk: a `B-NP, I-NP, E-NP, 1-NP` és `O` címkék jelölik rendre az NP elején, közepén és végén álló szavakat, az egyetlen szóból álló NP-ket, valamint az NP-n kívüli tokeneket.

A `hunchunk` rendszer teljesítményét úgy próbáltuk javítani, hogy az általa használt szószintű jegyek közé felvettük a parser által adott chunkcímkéket. Ezáltal a statisztikai modellnek lehetősége nyílik olyan súlyt rendelni a parser által adott válaszokhoz, mely a legmagasabb pontosságú címkézéshez vezet.

A minimális NP-k azonosításához elvégeztük a teljes NP-korpusz elemzését, majd az elemzett adatot tanító- és tesztadatra bontottuk ugyanúgy, mint tettük azt a `hunchunk` kiértékelésekor. Az eredmények a 2. táblázatban láthatók.

2. táblázat. A parser jegyek hatása a minimális NP-k azonosítására

	Precision	Recall	F-score
`hunchunk`	94.61%	94.88%	94.75%
`hunchunk`+parser jegyek	95.29%	95.68%	95.48%

Mint az a fenti táblázatból is látható, a parser kimenetének figyelembevétele a statisztikai rendszer hibáinak 15%-os csökkenéséhez vezetett. A parser által adott címkék ugyancsak hasznosnak bizonyultak a maximális NP-k azonosításakor (l. a 3. táblázatot).

3. táblázat. A parser jegyek hatása a maximális NP-k azonosítására

	Precision	Recall	F-score
`hunchunk`	89.34%	88.12%	88.72%
`hunchunk`+parser features	89.46%	88.76%	89.11%

6. Összefoglalás

Cikkünkben a magyar főnévi csoportok azonosításának egy szabályalapú megoldását mutattuk be. Kornai 1985-ös nyelvtanának továbbfejlesztésével a minimális NP-k azonosításán közel 90%-os F-pontszámot értünk el. Bár ez az eredmény – elvárásainkkal ellentétesen – nem közelítette meg a statisztikai rendszer pontosságát, alapjául szolgált egy hibrid rendszer megalkotásának. A maximum entrópiás Markov-modellt (MEMM) használó `hunchunk` rendszer, miután a szabályalapú rendszer kimenetét is figyelembe veszi, a korábbinál lényegesen magasabb pontosságot ér el mind a minimális, mind pedig a maximális NP-k azonosításában.

Hivatkozások

1. S. Abney. Parsing by chunks. *Principle-based parsing*, pages 257–278, 1991.
2. S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, 2009.
3. D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, pages 123–131, 2005.
4. A. Kornai. The internal structure of Noun Phrases. *Approaches to Hungarian*, 1:79–92, 1985.
5. A. Kornai. A főnévi csoport egyeztetése. *Általános Nyelvészeti Tanulmányok*, pages 183–211, 1989.
6. M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330, 1994.
7. P. Rebrus, A. Kornai, and D. Varga. Egy általános célú morfológiai annotáció. *Általános Nyelvészeti Tanulmányok*, 2010.
8. G. Recski, D. Varga, A. Zséder, and A. Kornai. Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban. *VI. Magyar Számítógépes Nyelvészeti Konferencia*, 2009.
9. E. F. Tjong Kim Sang, S. Buchholz, and Sang K. Introduction to the CoNLL-2000 shared task: Chunking, 2000.
10. V. Trón, A. Kornai, G. Gyepesi, L. Németh, P. Halácsy, and D. Varga. Hunmorph: open source word analysis. In *Proceedings of the Workshop on Software*, pages 77–85. Association for Computational Linguistics, 2005.

A. Az NP-parser nyelvtana

```
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NOUN[PRON=POS] NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d,
  D=?e, PRON=?f]
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e] ->
  NOUN[PRON=DEM, BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d]
  ART NOUN[PRON=0, BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=0]
```

```
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NOUN[PRON=GEN] NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d,
  D=?e, PRON=?f]
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NOUN[PRON=INDEF] NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d,
  D=?e, PRON=?f]
NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  ADJ NOUN[BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NOUN[BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0] ->
  ADJ NOUN[BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0]
NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0] ->
  NOUN[BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0]
NOUN[BAR=2, POSS=?a, PLUR=0, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NUM NOUN[BAR=1, POSS=?a, PLUR=0, ANP=?c, CAS=?d, D=?e, PRON=?f]
NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
NOUN[BAR=2, POSS=?a, PLUR=0, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0] ->
  NUM NOUN[BAR=1, POSS=?a, PLUR=0, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0]
NOUN[BAR=2, POSS=?b, PLUR=0, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0] ->
  NOUN[BAR=1, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e]
  /NOUN[BAR=0, PRON=?f]
NOUN[BAR=3, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f] ->
  ART[D=?e] NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, PRON=?f]
NOUN[BAR=3, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=1, PRON=?f] ->
  NOUN[BAR=0, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=1, PRON=?f]
NOUN[BAR=3, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f]
  /NOUN[BAR=0] ->
  ART[D=?e] NOUN[BAR=2, POSS=?a, PLUR=?b, ANP=?c, CAS=?d, PRON=?f]
  /NOUN[BAR=0]
NOUN[BAR=3, POSS=0, PLUR=?a, ANP=?b, CAS=?c, D=1, PRON=?f] ->
  NOUN[BAR=3, ANP=0, CAS=0] NOUN[BAR=2, POSS=1, PLUR=?a, ANP=?b,
  CAS=?c, PRON=?f]
NOUN[BAR=4, POSS=0, PLUR=?a, ANP=?b, CAS=?c, D=1, PRON=?f] ->
  NOUN[BAR=3, CAS=[DAT=1]] NOUN[BAR=3, POSS=1, PLUR=?a, ANP=?b,
  CAS=?c, D=1, PRON=?f]
NOUN[BAR=3, POSS=0, PLUR=?a, ANP=?b, CAS=?c, D=1, PRON=?f] ->
```

```
  ART[BAR=1, D=1, ME=?d, YOU=?e, PLUR=?f]
  NOUN[BAR=2, POSS=[ME=?d, YOU=?e, PLUR=?f], PLUR=?a, ANP=?b,
  CAS=?c, PRON=?f]
NOUN[BAR=3, POSS=0, PLUR=?a, ANP=?b, CAS=?c, D=1, PRON=?f] ->
  ART[BAR=0] NOUN[BAR=2, POSS=[], PLUR=?a, ANP=?b, CAS=?c, PRON=?f]
NOUN[POSS=?a, PLUR=?b, ANP=?c, CAS=?d, D=?e, PRON=?f, BAR=?g] ->
  PUNCT[TYPE='DQUOTE'] NOUN[BAR=?g, POSS=?a, PLUR=?b, ANP=?c,
  CAS=?d, D=?e, PRON=?f]
  PUNCT[TYPE='DQUOTE']
NOUN/NOUN ->

ART[BAR=1, D=1, ME=?a, YOU=?b, PLUR=?c, PRON=?f] ->
  ART[D=1] PRO[ME=?a, YOU=?b, PLUR=?c, PRON=?f]
ART[D=1] -> DET
ADJ -> ADJ ADJ
ADJ -> ADV ADJ
ADJ -> NOUN ADJ[SRC=[STEM=VERB[], DERIV='PERF_PART']]
ADJ -> NOUN ADJ[SRC=[STEM=VERB[], DERIV='IMPERF_PART']]
ADJ -> PUNCT[TYPE='DQUOTE'] ADJ PUNCT[TYPE='DQUOTE']
ADJ -> ADJ PUNCT[TYPE=COMMA] ADJ
ADJ -> ADJ PUNCT[TYPE=COMMA] CONJ ADJ
NUM -> NUM NUM
NUM -> ADV NUM
NUM -> ADJ NUM
```

VIII. Poszterek, laptopos bemutatók

Online morfológiai elemzők és szóalak-generátorok kisebb uráli nyelvekhez

Bakró-Nagy Marianne[2], Endrédy István[1], Fejes László[2], Novák Attila[1], Oszkó Beatrix[2], Prószéky Gábor[1], Szeverényi Sándor[2], Várnai Zsuzsa[2], Wagner-Nagy Beáta[3]

[1]MorphoLogic
1116 Budapest, Kardhegy utca 5.
{endredy, novak, proszeky}@morphologic.hu
[2]MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr utca 33.
{bakro, fejes, oszko, szeverenyi, varnai}@nytud.hu
[3]Universität Hamburg
Institut für Finnougristik/Uralistik
Johnsallee 35, 20148 Hamburg
beata.wagner-nagy@uni-hamburg.de

Kivonat: Cikkünkben egy olyan webhelyet mutatunk be, amelyen több korábbi projekt keretein belül számos kisebb uráli nyelvre készített morfológiai elemzőket, szóalak-generátorokat és korpuszokat tettünk elérhetővé. Az elemzések a webes felületen egy morfológiai és szemantikai egyértelműsítő eszköz formájában jelennek meg és minden nyelvhez virtuális billentyűzet segíti a szövegbevitelt.

1 Bevezetés

Az MTA Nyelvtudományi intézetének Finnugor és Nyelvtörténeti Osztálya és a MorphoLogic közötti együttműködés számos veszélyeztetett kisebb uráli nyelv morfológiájának számítógépes feldolgozására 2001-ben kezdődött a NKFP-5/135/01 számú *Komplex uráli nyelvészeti adatbázis* című projektum keretében. Ezt három OTKA projekt követte, amelyeknek keretében a finnugor nyelvcsalád permi ágához tartozó komi és udmurt nyelvre[1], az északi szamojéd nganaszan nyelvre[2] és a két obi-ugor nyelv, a manysi és a hanti három nyelvjárására[3] készített számítógépes morfológiákat sikerült olyan szintre fejleszteni, hogy azokat a tudományos közösség számára publikálhatónak éreztük.

A morfológiákat, valamint a kipróbálásukhoz használható szövegeket webes felületen keresztül tettük elérhetővé, amelyet a MorphoLogic üzemeltet, és a `http://www.morphologic.hu/urali/index.php` címen érhető el.

[1] OTKA T 048309 *Permi nyelvészeti adatbázisok*
[2] OTKA K 60807 *A nganaszan nyelv számítógépes morfológiai elemzése*
[3] OTKA NF 71707 *Obi-ugor morfológiai elemzők és korpuszok*

2 A morfológiák

A finnugor nyelvek (komi, udmurt, manysi, hanti) elemzésére a MorphoLogic *Humor* elemzőjét használjuk. A nganaszan morfológiát a Xerox *xfst* eszközének felhasználásával készítettük el.

A komi és az udmurt beszélőinek száma a többi itt bemutatott nagyon erősen veszélyeztetett nyelvvel ellentétben viszonylag jelentős, a permi nyelvek számottevő irodalommal és könyvkiadással rendelkeznek, ezért ezekre a nyelvekre olyan elemzőket készítettünk, amelyek a sztenderd cirill helyesírással írott szövegek elemzésére képesek. A kisebb, elsősorban csak beszélt nyelvként élő nyelvekre olyan elemzőket készítettünk, amelyek latin betűs fonologikus átírást használnak. Az utóbbiak esetében komoly problémát jelentett a lejegyzések következetlensége, illetve az, hogy a különböző szövegkiadásokban jelentősen különböző átírásokat használtak. A manysi esetében ugyanazon nyelvjárás (az északi) három korpuszának (ChVog[4], WT[5], VNGY[6]) feldolgozásához három különböző elemzőt kellett készítenünk.

A weboldal létrehozását elsősorban az a cél motiválta, hogy ezekkel a nyelvekkel kapcsolatos nyelvi adatokat hozzáférhetővé tegyük a tudományos kutatóközösség minél szélesebb köre számára. Ezért lehetőség szerint glosszákkal együtt jelenítjük meg az elemzéseket, hogy azok a nyelvet nem beszélő kutatók számára is értelmezhetők legyenek. A közeljövőben befejeződő obi-ugor OTKA projektben már kifejezetten cél volt az elemzők tőtárának angol glosszákkal való ellátása is. Az egyértelműsítő felület ily módon egyben szemantikai egyértelműsítésre is használható.

A weboldalon jelenleg az alábbi morfológiák érhetők el:

nyelv	glosszázás	tőtár	tőtár	toldaléktár mérete[7]
		lemma	jelentés	
nganaszan	magyar	4200	4775	310
komi (zürjén)	nincs (orosz glosszák hozzáadását tervezzük)	36000		193
udmurt	magyar	13500	18500	286
északi manysi (WT)	magyar, német, angol	3820	4200	376

Az év végéig az alábbi morfológiák kerülnek még fel az oldalra:

északi manysi (ChVog)	magyar, német, angol	1250	1530	271
északi manysi (VNGY)	magyar, német, angol	15600	16500	297

[4] Kálmán Béla: Chrestomathia Vogulica. Tankönyvkiadó, Budapest. (1989)

[5] Kálmán Béla: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest. (1976)

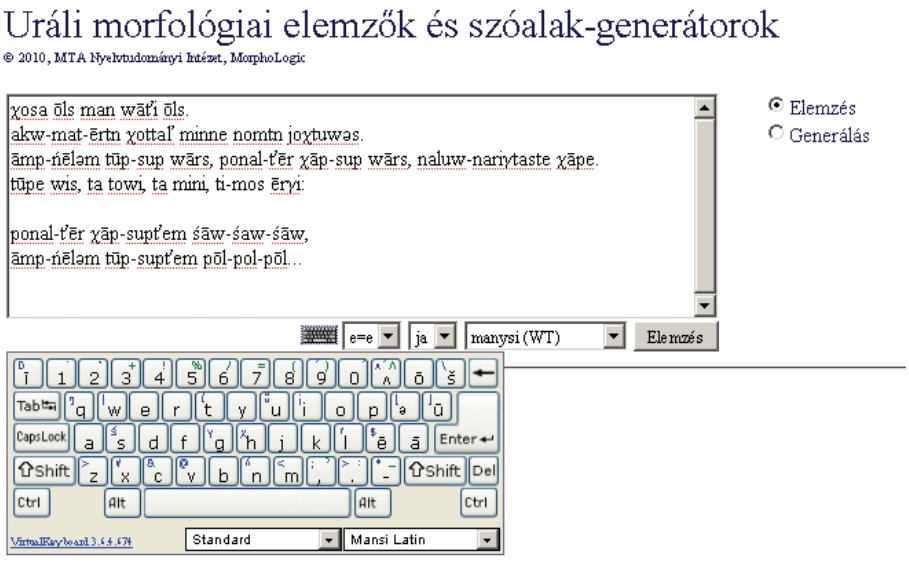
[6] Munkácsi Bernát: Vogul népköltési gyűjtemény. 1–4. Budapest. (1892–1921)

[7] morféma, ill. lexikalizált morfémakombináció

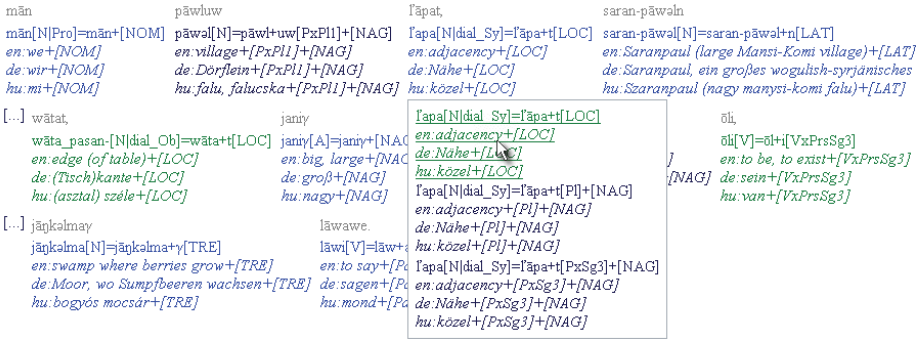
szinjai hanti	magyar és angol	2300	2500	138
kazimi hanti	magyar és angol	1750	1950	151

3 A webes felület

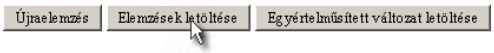
Az elemzők esetében a kiválasztott szöveget a megfelelő ablakba másolva a felhasználó megkapja a szövegben szereplő szavak lehetséges morfológiai elemzéseit és az elemzésekben szereplő tőmorfémák jelentését. A webes felületen valamennyi nyelven hozzáférhetőek olyan példaszövegek, amelyekkel az elemző kipróbálható. Virtuális billentyűzet segítségével a felhasználó maga is gépelhet be szövegeket.



Az elemzéseket megjelenítő webes felület egyben kézi egyértelműsítő eszközként is szolgál: a többértelmű szavak elemzései pop-up ablakban jelennek meg, ha az egeret egy többértelmű szó fölé mozgatjuk, ezek közül egérrel választhatunk.

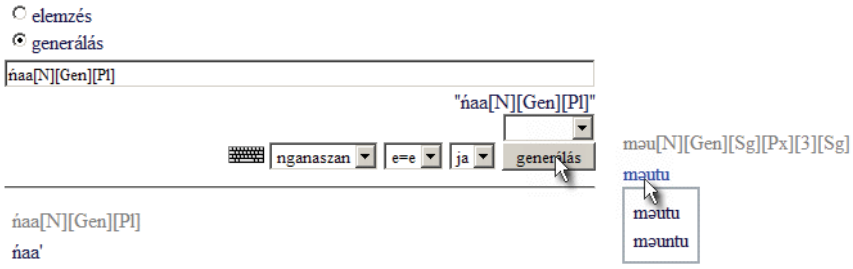


Az elkészült elemzések, illetve azok egyértelműsített változata elmenthető, az elmentett változatot a böngészőbe betöltve, az esetlegesen félbehagyott egyértelműsítő munka később folytatható.



Egyértelműsítés közben javíthatóak az elemzendő szövegben előforduló elgépelések is, ezután a szöveg újraelemeztethető. Ilyenkor nem vesznek el a korábban hozott egyértelműsítési döntések.

A webes felületen keresztül nemcsak morfológiai elemzők, hanem szóalakgenerátorok is elérhetők az egyes nyelvekhez. Ha egy adott morfémasorozat több formában is megjelenhet, akkor a generátor kimenete az elemző többértelmű kimenetének megjelenítéséhez hasonlóan jelenik meg a webes felületen, a lehetséges szóalakváltozatok itt is az egérmutatót a generált szóalak fölé mozgatva megjelenő pop-up ablakban láthatóak.



Terveink között szerepel, hogy a weblapot egyértelműsített korpuszokkal és korpuszkereső szolgáltatással egészítjük ki.

Bibliográfia

1. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications. Stanford University, Stanford (2003)
2. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, Szeged (2003) 138–145
3. Novák, A.: Language resources for Uralic minority languages. In: Proceedings of the SALTMIL Work-shop at LREC-2008: Collaboration: interoperability between people in the creation of language resources for less-resourced languages. Marrakech (2008) 27–32
4. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H., Yli-Jyrä, A. (szerk.): *Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford (2005) 116–125

MSD-KR harmonizáció a Szeged Treebank 2.5-ben

Farkas Richárd[1], Szeredi Dániel[2], Varga Dániel[2], Vincze Veronika[3]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
rfarkas@inf.u-szeged.hu
[2] BME Média Oktató és Kutató Központ
daniel@bme.mokk.hu, daniel@szeredi.hu
[3] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu

Kivonat: A magyar morfológiai erőforrások közül az egyik legelterjedtebben használt a morphdb.hu, amelynek morfológiai annotációs formalizmusa az úgynevezett KR-kódolás. A legnagyobb, kézzel egyértelműsített magyar nyelvi korpusz, a Szeged Treebank kódrendszere ezzel szemben az MSD-kódolást követi. A két kódolás nem kompatibilis egymással. Ez azt jelenti, hogy ha egy statisztikus módszerekkel tanított nyelvi elemző komponensben (POS-tagger, konstituenselemző, dependenciaelemző stb.) mindkét erőforrást ki kívánjuk aknázni, akkor nehézkes, információvesztéssel járó konverziós műveleteket kell végeznünk. Ebben a munkában beszámolunk a két kódrendszer (MSD és KR) közös nevezőre hozásáról, harmonizációjáról, amely megoldja a fenti problémát. A munka mindkét erőforrásban alapvető átalakításokkal járt. A konfliktusok nagyobb részében a harmonizációt közös finomítással igyekeztünk elvégezni, melynek hozadékaként jelentős mennyiségű manuális munka befektetésével a Szeged Treebank 2.5 által hordozott morfológiai információ részletgazdagabbá vált az előző verziókhoz képest.

1 Bevezetés

A magyar vonatkozású nyelvtechnológiai kutatásoknak és fejlesztéseknek alapfeltétele, hogy rendelkezésre álljon egy (lehetőleg egységes) nyelvi előfeldolgozó alapeszköztár. A rendelkezésre álló nyelvi elemzők egységesítésének legnagyobb akadálya a különböző morfológiai kódrendszerek használata. Cikkünkben beszámolunk két magyarra alkalmazott kódrendszer (MSD és KR) közös nevezőre hozásáról, harmonizációjáról. Ehhez tételesen ismertetjük a kódolások közötti elméleti különbségeket, majd az összehangolás során meghozott kompromisszumos döntésekről is beszámolunk. Az átalakított kódrendszernek megfelelően a morphdb.hu-ban [4] is változásokat eszközöltünk és a Szeged Treebank [2] szövegállományát is újrakódoltuk (a létrejött új verziót Szeged Treebank 2.5-nek kereszteltük). Célunk, hogy az egységes morfológiának köszönhetően létrejöhessen egy olyan morfológiai elemző, amely a Szeged Korpusszal is kompatibilis, annak érdekében, hogy a morfológiai elemzőre egy olyan POS-tagger legyen építhető, amely a magasabb szintű elemzé-

sekhez, illetve alkalmazásokhoz (dependenciaelemzés, információkinyerés) hasznos bemenetet szolgáltat.

2 Morfológiai kódrendszerek a magyar nyelvre

Az MSD morfológiai kódrendszer [3] több nyelvre, többek közt a magyarra lett kifejlesztve. A kódokon belül az első pozíció adja meg a fő szófaji kategóriát, míg a további pozíciók egyéb nyelvtani információkat tartalmaznak (pl. ige esetében az ige típusát, módját, idejét, számát, személyét, ragozását: a **Vmis2s---y** kód például egy kijelentő módú, múlt idejű, egyes szám második személyű tárgyas ragozású főigét jelöl).

A KR kódrendszer a magyar nyelv morfológiáját szem előtt tartva lett kidolgozva, bár alapvető szintaxisa nyelvfüggetlen, és a későbbiekben több más nyelvhez is készült a szintaxisra és a kódrendszer alapelveire épülő morfológiai erőforrás [4]. Magyar nyelvre történő implementációja, a morphdb.hu morfológiai elemző erőforrás létrehozásakor a legfontosabb célkitűzések a teljesség és az elméleti nyelvészeti szempontból való megalapozottság voltak, valamint hangsúlyos szempont volt a nyílt forráskódú szabad hozzáférhetőség. A kódrendszer hierarchikus jegy-érték struktúrában kódolja a nyelvészeti információkat: vannak alapértelmezett (default) jegyek (például egyes szám, harmadik személy), és csak az ettől eltérők jelennek meg a kódban. A fenti példa KR-kódolása a következő: **VERB<PAST><PERS<2>><DEF>**. A kódok inflexiós és derivációs információt is tartalmaznak.

A HUMor morfológiai kódrendszer az unifikációs nyelvleíráson alapul, azaz a tövek és morfémák más morfémákkal való együttes előfordulásra való képességük alapján jegyekkel vannak ellátva. E jegyek lehetnek egymást megengedők vagy egymásnak ellentmondók: egy szóalak csak olyan morfémákból épülhet fel, amelyek jegyei nem zárják ki egymást [5]. Az elemzés eredményeképpen a szó morfémákra bontott változatát kapjuk, minden morféma mögött szerepel a szófaji megjelölése, és ha eltér a szótári alakja, az is (*megy~me*), például: **mehetsz --megy[IGE]=me+het[HAT]+sz[e2]**.

Mivel a Szeged Korpusz építéséhez a szófaji előelemzést a HUMor morfológiai elemzőprogram végezte, melynek végeredményét automatikusan konvertálni kellett MSD-kódokra [1], az MSD és a HUMor kódrendszer harmonizációja már korábban megtörtént: a végeredmény a Szeged Treebank szófaji kódjaiban is tükröződik. Jelen cikkben a KR és MSD kódrendszerek összehangolására teszünk kísérletet.

3 A KR és MSD kódrendszerek harmonizációja

A kódrendszerek összehangolásában azt az alapelvet követtük, hogy a morfológiai kódoknak olyan (és csak olyan) információkat kell tartalmazniuk, amelyek a későbbi feldolgozás (szintaxis, különféle alkalmazások) szempontjából hasznosak. Ennek fényében mérlegeltük az egyes esetekben, hogy az MSD vagy pedig a KR rendszer megközelítését építsük-e be a harmonizált morfológiába.

Az egyik lényegi különbség a képzések kezelésében nyilvánul meg: míg a KR abszolút, addig az MSD relatív szótövekkel dolgozik. Ennek megfelelően a képzők nincsenek is kódolva MSD-ben, míg KR-ben igen, így adott esetben a szóalakok lemmája is eltér egymástól. A képzés hiányából adódóan az MSD kódrendszer nem tudja megkülönböztetni például ugyanannak az igének a műveltető vagy ható képzős alakjait a kód szintjén (természetesen a lemma eltérő) – ezzel szemben a KR-ben a lemma ugyanaz, de a kód különbözik.

Megoldásunk ebben az esetben az lett, hogy mindkét rendszerből átvesszük az indokolható megkülönböztetéseket. A relatív lemmák általában elég információt szolgáltatnak az alkalmazásoknak (pl. információ-visszakeresés), és a képzők annotálása a Szeged Korpuszban irreálisan nagy feladat lett volna, így a harmonizált kódrendszer is relatív lemmákkal dolgozik. Néhány esetben azonban indokolt volt kivételt tenni. A műveltető, gyakorító és ható[1] igék esetében fontos, hogy a képző csak aspektuális, illetve modális változást jelent, melyeket más nyelvek más – nem morfológiai, hanem például szintaktikai – eszközökkel fejeznek ki, aminek például a gépi fordításban lehet jelentősége. Ha pl. egy műveltető igealakot tartalmazó mondatot akarunk gépi úton angolra fordítani, akkor az MSD-kódolást használva abba a problémába ütközünk, hogy nagy valószínűséggel nem találunk a lemmának megfelelő szóalakot a szótárban. A KR-elemzést tekintve azonban a szótárban is megtalálható lemmából indulunk ki, és ha megfelelő fordítási szabályokat rendelünk a műveltetés (például használd a have + tárgy + ige 3 alakja szerkezetet) megfelelő kezeléséhez, akkor eljuthatunk a helyes fordításhoz.

Ezek alapján fontosnak tartottuk, hogy ezek az információk kódolva legyenek az MSD kódrendszerben is. Az igetípus pozíciójában azt is megjelöljük, hogy az ige műveltető (kódja: s), ható (kódja: o) vagy gyakorító (kódja: f) alakban szerepel-e.

Egy másik nagy elvi különbség a kódrendszerek között a névmások kezelése. Míg az MSD-ben külön szófaji kategóriának számítanak, addig a KR a helyettesített szófaj szerint kódolja őket. Az egységesítés eredményeképpen a KR rendszerbe is bevezettük a névmásokat PRONOUN jelöléssel.

A határozószavak kezelésében is mutatkoznak eltérések: az MSD-ben alosztályokba vannak sorolva, a KR-ben pedig egységesen <ADV> kóddal rendelkeznek. Az egységesítés folyamán az alosztályok megkülönböztetését választottuk, ugyanis ennek például a fokozásban van jelentősége. Az MSD kódrendszer képes jelölni a határozószavak fokozását, míg a KR-ből ez hiányzik: a *lejjebb*, *közelebb* alakok lemmája *lejjebb*, *közelebb*, kódolása pedig ADV. Az MSD-n belül mindez Rxc kódú (a c jelöli a középfokot), a lemmák pedig *lent* és *közel*. Viszont nem minden határozószó fokozható (a kérdő vagy általános határozószók például nem), ezért úgy szükséges módosítani a KR-kódolást, hogy csak bizonyos altípusok esetén legyen megengedve a fokozás lehetősége.

Az ún. személyes névmási határozószavak kérdése jelentette az egyik legjelentősebb elvi különbséget a két kódrendszer között. Míg MSD-ben a határozószavak egy altípusaként voltak kódolva (pusztán számot és személyt kódolva), addig a KR-ben

[1] Megjegyezzük, hogy az eredeti KR rendszerben a *-hAt* toldalék inflexióként jelenik meg, a harmonizált kódrendszerben azonban hasonlóképpen kezeljük a műveltető és gyakorító igeképzőkhöz, ezért itt tárgyaljuk.

főnévként: a határozórag alapúaknál (pl. *nekem*, *veled*) a személyes névmás szerepelt lemmaként, és a főnévi paradigmához hasonlóan kaptak esetet, a névutóból képzettek (*mögötted*, *szerintünk*) kódja pedig tartalmazta az eredeti névutót. Néhány példa: a *nekem* KR-elemzése én/NOUN<CAS<DAT>>, az MSD-elemzése Rl--s1 (*neki* lemmával), a *szerintem* szó esetében pedig én/NOUN<POSTP<SZERINT>>, illetve Rl--s1 (*szerinte*). A példákból ismét csak megmutatkozik az az eltérés a kódrendszerek között, hogy míg MSD-ben a kódolások megegyeznek, de a lemmák eltérnek, a KR rendszerén belül a lemmák megegyeznek, de a kódok különböznek.

Ennél a problémakörnél teljes egészében egyik rendszer megoldását sem vettük át. Mivel személyes névmásokból származtatjuk az alakokat, ezért a személyes névmási rendszerbe illesztjük be őket.

Szavak és szóalakok szófaji besorolását tekintve is találhatunk különbségeket a két kódrendszer között: jellemzően a kötőszavak és a határozószavak csoportjában fordul elő, hogy az egyik kódrendszerben kötőszó, a másikban határozószó az adott szóalak (pl. *majd*, *persze*). Ezek státuszáról egyenként hoztunk döntést, nyelvi disztribúciójukat mérlegelve.

Néhány kisebb horderejű különbség is megfigyelhető a két kódrendszer között. A főnevek kategóriáján belül ilyen például a köznév-tulajdonnév megkülönböztetés, mely az MSD sajátja. Mivel úgy gondoljuk, hogy nem a morfológiai elemző feladata eldönteni egy adott főnévről, hogy az tulajdonnév-e vagy sem (hanem egy NE-felismerőé), úgy döntöttünk, hogy az MSD-n belül sem érdemes ezt az elkülönítést alkalmazni. A familiáris többes számot a KR külön kódolja <FAM> jeggyel, az MSD-ben azonban ez nem szerepel. Mivel alkalmazási szempontból nem tűnt szignifikánsnak a többes szám kétféle jelölése, az egységes morfológiában csak egy "általános" többes számot használunk.

A Szeged Treebank 2.5 munkálatai nem csak elvi morfológiai átalakításokban öltöttek testet: a helyesírási hibát vagy elírást tartalmazó szóalakok mellé felvettük azok helyes alakját is annak lehetséges MSD-kódjaival együtt, majd a szövegkörnyezetnek megfelelően kiválasztottuk az aktuális kódot.

4 Konklúzió

Az előző fejezetben bemutatott harmonizációs lépéseket a morphdb.hu és a Szeged Korpusz manuális átalakításával valósítottuk meg. A két nyelvi erőforrás átalakításának statisztikai mutatóinak bemutatására hely hiányában nincs lehetőségünk, de részleteiben is elérhetőek a www.inf.u-szeged.hu/rgai/krmsd honlapon.

A cikkben bemutatott egységes morfológiának köszönhetően lehetővé vált olyan morfológiai elemző építése, amelynek kimenete a Szeged Treebankkel teljes összhangban van, és ezért a rá épülő, magasabb szintű nyelvi elemzést végző szövegfeldolgozó rendszerek (mint a magyarlanc[2] és hun* eszközláncok) a Szeged Treebank által hordozott minden morfológiai információt ki tudják használni statisztikus modelljeik tanításakor.

[2] www.inf.u-szeged.hu/rgai/magyarlanc

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Alexin, Z., Csirik, J., Gyimóthy, T., Bibok, K., Hatvani, Cs., Prószéky, G., Tihanyi, L.: Manually Annotated Hungarian Corpus. In: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03. Budapest, Hungary, 15-17 April (2003) 53-56
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123-131
3. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf
4. Kornai, A., Rebrus, P., Vajda, P., Halácsy, P., Rung, A., Trón, V.: Általános célú morfológiai elemző kimeneti formalizmusa. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2004) 172–176
5. Prószéky, G., Tihanyi, L.: Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. La tribune des industries de la langue 10, OFIL, Paris, France (1993) 28–29

Bizonytalanságot jelölő kifejezések és hatókörük azonosítása természetes nyelvi szövegekben: a CoNLL-2010 verseny tapasztalatai

Farkas Richárd[1,2], Vincze Veronika[2], Móra György[2],
Csirik János[1,2], Szarvas György[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{rfarkas, csirik, szarvas}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{vinczev, gymora}@inf.u-szeged.hu

Kivonat: A CoNLL-2010 konferenciához kapcsolódó nemzetközi versenyfeladat a bizonytalanságot jelölő kifejezések, és azok hatókörének azonosítását tűzte ki célul angol nyelvű szövegekben. Cikkünkben bemutatjuk a versenykiírást, a beérkezett rendszereket, a kiértékeléshez épített adatbázisokat, értékeljük az eredményeket, végül pedig beszámolunk egy – hasonló elvek alapján épített – magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról.

1 Bevezetés

A CoNLL-2010 konferenciához kapcsolódó nemzetközi versenyfeladat a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának szervezésében a bizonytalanságot jelölő kifejezések, és azok hatókörének azonosítását tűzte ki célul angol nyelvű szövegekben [1]. A feladat fontossága abban rejlik, hogy a különféle számítógépes nyelvészeti alkalmazásokban lényegi szerep jut a tényszerű és a bizonytalan, illetve tagadott információ megkülönböztetésének, hiszen például információkinyerés és szemantikus keresés esetében a felhasználónak többnyire tényszerű információra van szüksége, így alkalmazástól függően a rendszer vagy kiszűri a bizonytalan / tagadott szövegrészeket, vagy pedig a tényektől elkülönítve adja őket vissza a felhasználónak.

Cikkünkben összefoglaljuk a verseny tapasztalatait, valamint beszámolunk egy magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról.

2 A versenyfeladatok

A bizonytalanságot tartalmazó szövegrészek azonosítása történhet mondatszinten és hatókör szinten. Az első esetben elégséges a mondatról eldönteni, hogy az tartalmaz-e bizonytalan információt vagy sem, míg a második esetben a cél: megjelölni a mondaton belül a bizonytalanságot jelző nyelvi elemeket (kulcsszavakat) és azok mondatbeli

hatókörét. Noha az utóbbi feladat nagyobb kihívást jelent, a legtöbb alkalmazás számára mégis előnyt jelent ez a jelölési módszer, hiszen lehetnek olyan (általában összetett) mondatok egy szövegben, ahol a mondat egy része bizonytalan információt hordoz, más részében viszont hasznos tényszerű információ rejlik.

A versenykiírásban szereplő két feladat a fentieknek megfelelően mondat- és hatókör szintű címkézést tűzött ki célul. Az első feladat mondatszintű címkézést kívánt meg aszerint, hogy a mondat tartalmaz-e bizonytalan információt vagy sem. A rendszereknek biológiai témájú cikkek, illetve Wikipédia-szócikkek mondatait kellett osztályozniuk. A második feladatban biológiai cikkekben kellett bejelölni a kulcsszavakat és azok mondaton belüli hatókörét.

A versenyfeladatokhoz biztosítanunk kellett tanító és kiértékelő adatbázist is. Tanító adatbázisként a biológiai doménre a BioScope korpusznak [2] a tudományos cikkek absztraktjait és teljes cikkeket tartalmazó részét választottuk, a kiértékeléshez pedig újonnan annotáltunk 15 biológiai témájú cikket. A Wikipédia doménen pedig mind a tanító, mind a kiértékelő adatbázist az angol nyelvű Wikipedia *weasel* címkével ellátott (homályos, kétértelmű, túlzó vagy félrevezető információt tartalmazó) bekezdései közül válogattuk ki, melyekben kézzel megjelöltük a bizonytalanságot jelző szavakat. Néhány példamondat az adatbázisokból (<> jelöli a kulcsszavakat, míg () a hatóköröket):

The album, which was recorded in less than two weeks, contains <arguably> the band's two <most famous> songs, "Wonderwall" and "Don't Look Back in Anger" and their first UK #1 single, "Some Might Say". (Wikipédia)

Thus, misregulation of these genetic pathways (<may> confer unrestricted proliferative capacities to a range of glial cell types), but (how this occurs remains <unclear>). (biológiai publikáció)

Megjegyezzük, hogy a wikipédiás példamondat egyben azt is mutatja, hogy a kulcsszójelöltek nem minden előfordulásukban szerepeltek ténylegesen kulcsszóként: a *some* és a *might* gyakori kulcsszavak, de a fenti példában egy dal címének részeként – azaz metanyelvi használatban – nem utalnak bizonytalanságra.

A versenyzőknek lehetőségük nyílt az általunk rendelkezésekre bocsátott adatbázisok mellett további erőforrások használatára is a rendszerük fejlesztése során.

3 Versenyeredmények, értékelés

A versenyen 23 intézet kutatói vettek részt a világ minden tájáról. Az első feladat biológiai részére 22 csoport, a wikipédiás szövegek feldolgozására 16 csoport, míg a második feladatra összesen 13 csapat vállalkozott.

Az első feladat kiértékelése mondatszinten történt: a bizonytalan osztály F-mértékét alkalmaztuk mint fő kiértékelési metrikát. A második feladatban, ahol a kulcsszavakat és azok hatókörét is azonosítani kellett, egy szigorú, hatókör szintű kiértékelési metrikát használtunk: pontos találatnak csak azt fogadtuk el, ahol a kulcskifejezések és hatóköreik is pontosan lettek megállapítva.

A legjobb rendszerek az első feladatban 86% (biológiai domén), illetve 60% (Wikipédia) körüli F-mértéket értek el, a másodikban pedig 57% körülit. Utóbbi eredmény egyrészt a feladat nehézségét, másrészt pedig a kiértékelési metrika szigorúságát is jelzi: bizonyos esetekben a hatókörök rugalmasabb illeszkedése lenne kívánatos (például írásjelek, hivatkozások, zárójeles megjegyzések kezelése).

Az első feladatra a legjobb eredményt elért versenyzők biológiai szövegeken szekvenciajelöléses megközelítést alkalmaztak, míg a Wikipédia-szövegeken a szózsák típusú modellek bizonyultak sikeresnek.

4 Bizonytalanságot jelölő kifejezések a magyarban

A verseny résztvevőinek magas száma arra utal, hogy a bizonytalan szövegrészek azonosításának problémája élénken foglalkoztatja a számítógépes nyelvész kutatókat világszerte. Míg az eddigi kutatások nagy része az angol nyelvre irányult (azon belül is elsődlegesen az orvosi-biológiai szövegekre), szeretnénk a továbbiakban a magyar nyelvre is kiterjeszteni az ilyen témájú kutatásokat. E cél érdekében kísérleti jelleggel építettünk egy magyar nyelvű, Wikipédia-szócikkekből álló adatbázist[1], melyben kézzel annotáltuk a bizonytalanságot jelölő nyelvi elemeket, az ún. weasel szavakat[2]. A weasel szavak a véleményeket megfelelő forrás vagy alátámasztás nélkül tálalják: nem tükrözik egy enciklopédia szerkesztői (és olvasói) által elvárt semleges stílust. A következő példában az információ forrása nincs megadva, pusztán a *sokan* kifejezés utal a vélemény hordozójára:

*Ma már **sokan** úgy vélik, hogy ez a megítélés erősen szubjektív, hiszen Linné maga is rendszerint a svédországi fajt (vagy alfajt) látta a „legtipikusabbnak”.*

A korpusz létrehozásában követtük az angol nyelvű adatbázis építésekor alkalmazott alapelveket annak érdekében, hogy az eredményeket összevethessük a versenyfeladathoz épített korpusz adataival. Elsőként egy nyelvészeti szempontok alapján összeállított kulcsszólista segítségével gyűjtöttünk a magyar Wikipédia szócikkeiből bekezdéseket, majd ezekből – a kulcsszó-jelöltek gyakorisági adatait szem előtt tartva – véletlenszerűen válogattuk ki az annotálandó bekezdéseket. A végső annotált korpusz 1710 bekezdést tartalmaz. A munka során a nyelvészek bejelölték a weasel kifejezéseket a szövegekben, majd azokat a mondatokat minősítettük bizonytalannak, amelyek legalább egy kulcsszót tartalmaznak. A 11647 mondatból 953 volt ilyen (8,18%). Összehasonlításképpen: az angol tanító adatbázison 22,36%, a kiértékelő adatbázison pedig 23,19% volt a bizonytalan mondatok aránya.

A szövegekben összesen 1156 kulcsszó fordult elő, vagyis egy bizonytalan mondat átlagosan 1,21 kulcsszót tartalmazott. A leggyakoribb kulcsszavak, illetve kulcskifejezések a következők voltak: *számos N* (132 előfordulás), *valószínűleg* (128), *egyes N* (91). A kulcsszavak csoportjait tekintve elsődlegesen a határozatlan vagy általános

[1] A korpusz Creative Commons licenc alatt elérhető a www.inf.u-szeged.hu/rgai/uncertainty oldalon.

[2] http://hu.wikipedia.org/wiki/WP:WEASEL

kvantorokat (*egyes, néhány*) tartalmazó kifejezések és a bizonytalanságra vagy általánosságra utaló határozószók (*valószínűleg, feltehetőleg, általában*) domináltak a korpuszban, de a *más N* és *mások* kifejezések használata is jellemző volt.

A fenti számadatok azt mutatják, hogy a magyar Wikipédiában a bizonytalan mondatok aránya az angollal összevetve jelentősen kisebb. Ennek két fő oka lehet. Egyrészt, az angol Wikipedia szerkesztői közössége valószínűleg sokkal heterogénebb, mint a magyaré, ezért ott nagyobb az esély arra, hogy egy új szócikket egy kevesebb szerkesztői tapasztalattal rendelkező tag hozzon létre, növelve ezzel a bizonytalan szócikkek arányát. Másrészt, a Wikipédiák méretbeli különbségéből adódóan a bizonytalan szócikkek száma abszolút értékben véve jóval kevesebb a magyarban, azaz a szerkesztők könnyebben és gyorsabban ki tudják ezeket javítani.

A kulcsszavak gyakoriságát illetően a két nyelv között nincs számottevő eltérés. Az angol adatbázis leggyakoribb kulcsszavai a *some*, *may* és *others* voltak, míg a magyarban is gyakran fordultak elő a *számos*, *egyes*, *más*, *mások* kifejezések. Mivel a magyar nyelv morfológiailag fejezi ki a ható modalitást, az angol pedig a *may* segédigével, a *may* kulcsszó gyakorisága a *–hAt* morfémát tartalmazó elemek gyakoriságával vethető össze, ez pedig a két nyelv esetében hozzávetőlegesen megegyezik.

5 Összegzés

Tanulmányunkban beszámoltunk a CoNLL-2010 konferenciához kapcsolódó versenykiírásról, ahol a cél bizonytalanságot jelölő kifejezések azonosítása volt. Röviden bemutattuk a kiértékeléshez épített adatbázisokat, ismertettük a beérkezett rendszereket, végül pedig leírást adtunk egy – hasonló elvek alapján épített – magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról, mely a későbbiekben a magyar nyelvre fejlesztendő, bizonytalan szövegrészeket azonosító alkalmazások tanításában, illetve egységes kiértékelésében tölthet be fontos szerepet.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND, a BELAMI és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Uppsala (2010) 1–12
2. Vincze, V., Szarvas, Gy., Farkas, R., Móra, Gy., Csirik, J.: The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. Bioinformatics Vol. 9, No. 11 (2008)

Szemantikus annotációk létrehozása a weben nyelvtechnológiai eszközök támogatásával

Héder Mihály[1,2]

[1] MTA Számítástechnikai és Automatizálási Kutatóintézet
Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem
Filozófia és Tudománytörténet Tanszék

1. Absztrakt

A weben található hipertext minőségének egyik mutatója, hogy milyen mennyiségben található a szöveg mellett számítógép által is értelmezhető, azaz strukturált reprezentációja a szándékolt jelentésnek. Az így csatolt információkat szemantikus annotációknak is nevezhetjük, mivel úgy magyarázzák a számítógép számára az egyes szövegrészek értelmét, mint a széljegyzetek egy könyvben.

Az MTA Sztaki Internet Technológiák és Alkalmazások Központjában a szemantikus annotációkkal kapcsolatos kutatások és fejlesztések keretein belül létrehoztunk egy webes keretrendszert, amelynek segítségével az annotáció kérdései gyakorlati síkon is tárgyalhatókká váltak.

Az eszköz amellett, hogy megoldásokat kínál az annotációk granularitásával, szintaxisával és lekérdezhetőségével kapcsolatos néhány problémára, képes UIMA és egyéb interfésszel rendelkező nyelvfeldolgozó adapterek használatára is. Az angol nyelvű és nyelvfüggetlen adaptereken kívül az eszköz a Szegedi Tudományegyetem Nyelvtechnológiai csoportja által fejlesztett *magyarlanc* [1] UIMA-adaptereket is használja, a magyar nyelvű feldolgozás döntően ezen modulok segítségével történik.

Az annotáló szoftver leginkább végfelhasználóknak szóló alkalmazása egy Wikipédia-cikkszerkesztő. Ebben a konfigurációban a szoftver egy hivatkozásajánlásokat megfogalmazó névelem-felismerőt és a Hitec [2] keretrendszer magyar wikin betanított, webszervizen elérhető verzióját is használja. Ez az alkalmazás mutat rá a legmarkánsabban azokra a nehézségekre, amelyeket az annotációk helyes tárolása és karbantartása jelent egy olyan formátum (jelen esetben a wikitext) és létrehozási munkafolyamat esetében, amelyet ezen feladatokra nem készítettek fel.

Egy panaszlevél-kezelő alkalmazás is bemutatásra kerül. Ebben az alkalmazásban egyszerre próbáljuk segíteni a szemantikus keretekkel és szkriptekkel, illetve a szerkezeti egységekkel kapcsolatos kutatásainkat és a majdani felhasználót. Az Igazságügyi Minisztérium panaszleveleit tartalmazó korpusz, amellyel

ebben a projektben dolgozunk, megköveteli újfajta megközelítések és heurisztikák kikísérletezését.

Végezetül bemutatásra kerülnek azon kísérleteink, amelyek a nyelvfeldolgozással támogatott jogiszöveg-létrehozás és -annotálás gyakorlati kérdéseit vizsgálták.

Hivatkozások

1. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: Magyar Számítógépes Nyelvészeti Konferencia. (2009)
2. : Hitec. (*categorizer.tmit.bme.hu/trac/wiki*)

Melléknevek szűk szemantikai osztályainak detekciója a Magyar Nemzeti Szövegtárban jelentés-egyértelműsítés céljából

Héja Enikő[1], Takács Dávid[1]

[1] MTA Nyelvtudományi Intézet
{eheja, takdavid}@nytud.hu

A jelentés-egyértelműsítés a hazai és nemzetközi nyelvtechnológia egyik központi problémája. Számos alkalmazás (pl. információkinyerés, gépi fordítás) számára kiemelkedő jelentőségű. A jelentés-egyértelműsítés két részfeladatra bontható: (1) megfelelő jelentéstár kiválasztása, amelynek elemei hozzárendelhetőek a szövegekben szereplő tokenekhez, (2) megfelelő algoritmus kiválasztása, amely ezt a hozzárendelést elvégzi. Az annotátorok közötti egyetértést mérő vizsgálatok bizonyítják (l. [3, 4]), hogy a már létező, nem kifejezetten jelentés-egyértelműsítés céljából kialakított egynyelvű adatbázisok (pl. *Petit Larousse*, *Magyar Értelmező Kéziszótár*) alapján a jelentés-egyértelműsítés még az emberek számára is nehéz vagy megoldhatatlan feladat. Tehát az intuíción alapuló, nem jelentés-egyértelműsítés céljából létrehozott jelentéstárak alkalmatlanok a gépi jelentés-egyértelműsítésre.

Mivel kontextuális információ alapján tudjuk csak automatikusan lehorgonyozni a megfelelő jelentést, a jelentéstár kialakításánál is célszerű kizárólag a kontextuális információra támaszkodni. Ez a megközelítés egybecseng a jelentés disztribúciós felfogásával, amit Firth [2] így fogalmazott meg: "*You shall know a word by the company it keeps*". Az általunk javasolt módszer fontos tulajdonsága, hogy kizárólag disztribúciós információt vesz figyelembe, a jelentéstár kialakításánál az emberi intuíciót figyelmen kívül hagyja, és a jelentéstár szerkezetére vonatkozóan semmilyen előzetes megkötést nem teszünk.

Az általunk végzett kutatás célja, hogy felügyelet nélküli tanulással egy melléknevet tartalmazó jelentéstárat készítsünk a Magyar Nemzeti Szövegtár adatai alapján. [1] klikk-klaszterezési (*clique-based clustering*) eljárását alkalmazva az MNSZ-ben annotált főnév-melléknév kapcsolatokra azt várjuk, hogy a létrejövő klaszterek a melléknevek szűk szemantikai osztályaival (pl. színnevek) esnek egybe. Ezek a kontextus alapján létrejövő klaszterek képezhetik a jelentés-egyértelműsítő rendszer jelentéstár-komponensét.

Az eljárás az alábbi lépésekből áll: (1) az annotált korpusz alapján felépítjük a melléknevek disztribúciós mátrixát, ahol minden melléknevet a módosított főnevek halmazával jellemzünk. (2) Ebből a mátrixból egy távolsági mérték alkalmazásával meghatározzuk az egyes melléknevek közötti kontextuális távolságot. (3) Egy megfelelő vágási paraméter alkalmazása után a Bron-Kerbosch algoritmussal meghatározzuk a létrejövő gráf teljes részgráfjait, vagyis klikkjeit. (4) Az így létrejött, jellemzően

kis elemszámú klikkeket egy klaszterezési eljárással összeolvasztjuk, aminek eredményeképpen az egyértelműsítés számára megfelelő finomságú felosztást kapunk.

A klikk-klaszterezés járulékos előnye, hogy a lépések során az egyes melléknevek a különböző jelentéseik szerint egyszerre több klaszterben is szerepelni fognak. Emellett mindvégig megőrizzük a kontextuális információkat, így fölépíthetünk egy, a melléknév-kontextus párok halmazát a jelentések halmazára leképező függvényt, amelyet közvetlenül használhatunk a jelentés-egyértelműsítés során.

Jelen kutatás célja annak vizsgálata, hogy felügyelet nélküli tanulással a fent javasolt módszerrel létrehozható-e egy olyan jelentéstár, amely a magyar melléknevek jelentés-egyértelműsítésének alapjául szolgálhat.

Bibliográfia

1. Ah-Pine, J., Jacquet, G.: Clique-Based Clustering for improving Named Entity Recognition systems. In: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece (2009)
2. Firth, J. R.: Papers in Linguistics. Oxford University Press, London (1957) 1934–1951
3. Kuti, J., Héja, E., Sass, B.: Sense Disambiguation — "Ambiguous Sensation"? Evaluating Sense Inventories for verbal WSD in Hungarian. In: Proceedings of the LREC W22. Malta (2010)
4. Véronis, J.: Sense tagging: does it make sense? In: Wilson, A., Rayson, P. McEnery, T. (szerk.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Peter Lang, Frankfurt (2003)

Egy nyelvészeti UIMA-folyamat a kézi annotálástól az eredmények megjelenítéséig

Kiss Márton[1], Nagy Ágoston[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{mkiss, nagyagoston}@inf.u-szeged.hu

Kivonat: A MaSzeKer projekt indulásakor az UIMA nyelvészeti keretrendszert választottuk a fejlesztéshez. Az már a fejlesztés kezdetekor látszott, hogy a következő modulokra mindenképpen szükségünk lesz a projekt nyelvészeti részének megvalósításához: *kézi annotálás, gépi annotálás, a két annotáció összehasonlítása* és az *eredmények megjelenítése.* Ezen igények teljes körű kielégítésére nem találtunk implementált rendszert. Ezért kifejlesztettünk egy nyelvészeti UIMA-folyamatot támogató környezetet (UIMA-modulokat és hozzájuk kapcsolódó segédprogramokat), mely az előbb említett technikai elvárásokat megvalósítja. A cikkben bemutatjuk a létrejött rendszer mindazon részeit, melyek segítségével nyomon követhető, segíthető egy nyelvészeti kutatás a dokumentumok kézi annotálásától az eredmények megjelenítéséig.

1 Bevezetés

Cikkünkben bemutatjuk azokat a kifejlesztett UIMA-modulokat és segédprogramokat, melyek segítségével megvalósítottunk egy olyan rendszert, mely hatékony támogatást nyújt számítógépes nyelvészeti kutatásokhoz. A kifejlesztett rendszernek azon részeit mutatjuk be, melyek támogatják: kisebb méretű, pár száz dokumentumot tartalmazó *tanulókorpusz építését,* a korpuszon végzett gépi és kézi jelölések *összehasonlítását*, valamint az eredmények *vizualizált megjelenítését*. A kifejlesztett rendszer segítségével könnyen megtalálhatjuk a gépi rendszer hiányosságait és kijavíthatjuk az esetleges hiányosságokat.

2 Az UIMA-modulok és a segédprogramok bemutatása

A következő fejezetben végigvesszük a modulok és segédprogramokat, ismertetve a működésüket és technikai megvalósításukat. A fejezet végén pedig egy ábrán szemléltetjük a rendszer felépítését és a modulok kapcsolatát.

A kifejlesztett UIMA-modulok: *AnnotationComparator, HTMLViewer*. A segédprogramok: *Word <-> UIMA konverter, Word <-> TXT konverter*.

2.1 Kézzel annotált Word-dokumentum konvertálása UIMA XMI-be (Word-makró + Perl + UIMA-modul)

A nyelvész kollégák számára olyan annotálási módszert kellett kidolgoznunk, mely könnyen elsajátítható és kényelmesen végezhető vele a munka. Erre azt a megoldást találtuk a legalkalmasabbnak, hogy egy Word-dokumentumban jelöljék meg a releváns szövegrészeket valamilyen előre megállapított formázás segítségével, például változtassák meg a szöveg háttérszínét. Ezek után a Word-dokumentumból az UIMA számára is értelmezhető annotációkat kellett készíteni. Ezen technikai megoldással létrejöttek a tanulókorpuszok.

A modul első lépésben kiexportálja a Word-dokumentumokban formázással jelölt kézi annotációkat egy *egyszerű XML*-fájlba (Word -makró segítségével). Ezek után az XML-fájlból egy Perl-script segítségével olyan *konfigurációs fájlokat* készítünk, melyek tartalmazzák az annotációkat és a pontos karakterpozícióját a szövegben. Végül egy UIMA-modul a konfigurációs fájl segítségével létrehozza az *annotációkat.*

2.2 Word-dokumentum, TXT-konverter (Word-makró)

A Word-dokumentumot TXT formátumra is kellett hozni, hogy az UIMA rendszer moduljai bemenetként felhasználhassák. Ezt a problémát egy Word-makró segítségével oldottuk meg, mely egy könyvtár (egy korpusz) összes .doc kiterjesztésű fájlját átalakítja TXT formátumra.

2.3 Annotációk összehasonlítása (UIMA-modul)

Amikor előálltak a kézi és gépi annotációk is, szükségünk volt arra, hogy összehasonlítsuk a kettőt. A gépi algoritmus hatékonyságát a pontosság, a fedés és az F-mérték kiszámításával mértük. Az összehasonlítás során többféle illeszkedés is beállítható attól függően, hogy hogyan szeretnénk összehasonlítani a két annotációt. Választható illeszkedési típusok:

teljes: ekkor a két annotációnak teljesen meg kell egyeznie mind a kezdeti, mind a végső karakterpozícióban

tartalmaz: ebben az esetben a két annotáció akkor is egyezik, ha az egyik „csak” tartalmazza a másikat, vagyis annot1 és annot2 annotációk esetén: annot1.begin <= annot2.begin és annot2.end <= annot1.end

Az összehasonlítás során a további hatékonyság növelése érdekében az összehasonlító modul kigyűjti a rosszul bejelölt vagy nem megjelölt annotációkat.

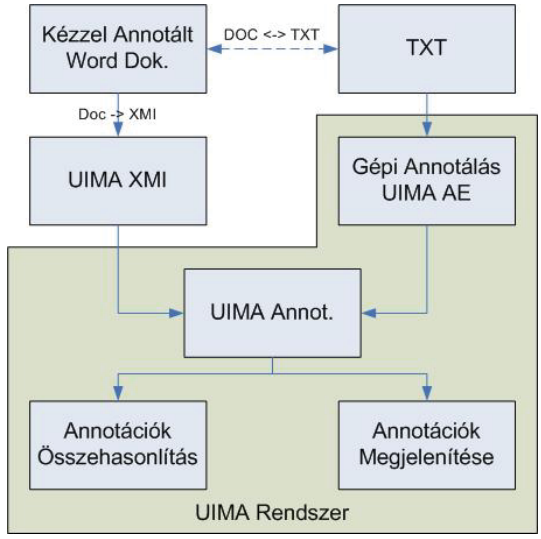
2.4 Megjelenítés (UIMA-modul + Perl + JavaScript)

Az eredmények vizualizációjára az ellenőrzés és az átláthatóság miatt volt szükség. Kétféle megjelenítőt készítettünk:

a) AZ UIMA InLine XML megjelenítése (XSL)

b) AZ UIMA XMI megjelenítése (Perl+UIMA+HTML)

Az a) esetben azon adatok, annotációk, kapcsolatok megjelenítésére van lehetőség, melyek *fastruktúrában* ábrázolhatóak: szülő és ős kapcsolat áll fenn két annotáció között. A b) esetben a feldolgozás során *bejelölt összes annotáció* megjeleníthető HTML formátumban.



1. ábra. A megvalósított UIMA-modulok és a segédprogramok kapcsolata.

Bibliográfia

1. Kano, Y., Nguyen, N., Sætre, R., Yoshida, K., Miyao, Y., Tsuruoka, Y., Matsubayashi, Y., Ananiadou, S., Tsujii, J.: Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. In: Proceedings of Pacific Symposium on Biocomputing (PSB), 13 (2008) 616–627
2. Ferrucci, D., Lally ,A.: Building an example application with the Unstructured Information Management Architecture. IBM Systems Journal Vol. 43 No. 3 (2004) 455–475
3. Kano, Y. et al.: U-Compare: share and compare text mining tools with UIMA. Bioinformatics, doi: 10.1093/bioinformatics/btp289 (2009)
4. D. Ferrucci, A. Lally: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Journal of Natural Language Engineering Vol. 10 No. 3-4 (2004) 327–348
5. Kunze, M., Rösner, D.: Tools for UIMA Teaching and Development. University of Magdeburg, Germany (2008)

A MASZEKER felhasználói felületének kialakítása

Minkó Mihály

[1] Szegedi Tudományegyetem, Könyvtár- és Humán Információtudományi Tanszék,
Szeged, Egyetem u. 2.
minko.mihaly@gmail.com

A MASZEKER felhasználói felületének tervezése és megvalósítása az SZTE Könyvtár- és Humán Információtudományi Tanszékének feladata volt. Az előadás a felület meg-tervezésének kiindulópontját, a tervezés folyamatát és a létrejött végeredményt mutatja be, illetve tárgyalja azokat a nehézségeket, amelyeket a felület tervezése során meg kellett oldani.

A felhasználói felület kialakítása kapcsán több olyan tényező is volt, amely nem tartozik szorosan a hagyományos értelemben vett keresőfelület (weben, dokumentumkorpuszokon vagy rdf-ekben kereső algoritmusok felülete) tervezéséhez. Az első feladat azoknak a sajátosságoknak a meghatározása és elkülönítése volt, amelyek a MASZEKER felhasználói felületét jellemzően megkülönböztetik ezektől a felületektől, illetve azoknak a hasonlóságoknak az áttekintése, amelyeket a hagyományos felülettervezésből átvéve sikeresen tudtunk alkalmazni. Ebben nagy segítséget nyújtott Marti A. Hearst: Search User Interfaces című munkája, amely az egyik - ha nem a - legalaposabb áttekintését adja a keresőinterfész tervezésekor számba veendő szempontoknak. Miután ez az összehasonlítás megtörtént, a rendelkezésre álló dokumentumok alapján elkezdődött az interfész tervezése. A konkrét kivitelezésben nélkülözhetetlen segítséget nyújtott a Törcsvári Attila által készített használatieset-leírás, amely biztos kiindulópontul szolgált egy UML-alapokon nyugvó fejlesztéshez és tervezéshez. A használati esetek azonban kevés vizuális és funkcionális támpontot adnak egy felület megtervezéséhez, a GUI (Graphical User Interface) elkészítéséhez. A tervezés folyamatának talán legfontosabb lépése volt egy olyan módszer felkutatása, amely segítségével a rendelkezésre álló használati esetekből a felhasználók által használatba vehető szoftverinterfész készülhet. A vonatkozó szakirodalom több megközelítést és transzformációs lehetőséget is bemutat. Ezek közül került kiválasztásra az a módszer, amelyet a felület tervezésekor alkalmaztunk és amelynek részletes bemutatása szükséges lesz. Ezen módszer segítségével elkészült egy olyan általános felületterv, amely már a drótvázak szintjén tartalmazta azokat az adattípusokat és interakciókat, amelyek segítségével a felhasználó operálhat a szoftverrel. A felületterv felhasználóknak történő bemutatása és felhasználók általi értékelése után megtörtént a terv véglegesítése és átadásra került a felület programozását végző kollégának, Danics Attilának, aki elkészítette a végleges szoftverinterfészt.

Bűnügyi névelem-felismerés

Molnár Gábor József[1], Kojedzinszky Tamás[1], Farkas Richárd[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
gjmolnar@inf.u-szeged.hu, Kojedzinszky.Tamas@stud.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: Ebben a munkában bemutatjuk a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében elkészült magyar nyelvű névelem-felismerő rendszert. A feladat bűnügyi dokumentumok szövegeiben található fontosabb szereplők felismerése, azaz előre definiált kategóriákba tartozó kifejezések azonosítása és azok megfelelő osztályba sorolása volt. A feladat és megoldásának érdekességei, hogy egyrészt egy sokosztályos klasszifikációt kellett megoldani, ahol az osztálycímkéken egy rendezés van definiálva, másrészt a szekvenciajelölés kétszintű: az első szinten jelölt tokenek további osztályokba sorolandók, valamint mivel a rendszer pontossága kiemelt szempont volt, ezért megvizsgáltuk kézi szabályok integrálási lehetőségeit is.

1 Bevezetés

Ebben a munkában egy komplex bűnügyi névelem-felismerő rendszert mutatunk be, amely a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében készült el. A rendszer alapvetően a korábban Szegeden kidolgozott névelem-felismerő keretrendszer [1] egy kiterjesztett változata.

A kiterjesztésre a feladat három specialitása miatt volt szükség.

- A feladat egyik érdekessége, hogy a szokványos négy névelemosztály (földrajzi hely, személynév, szervezetnév, egyéb tulajdonnév) helyett 13 szemantikai osztályt kellett megkülönböztetnünk.

- Mivel a felismert említésekben azonosítani kellett azok alkotóelemeit is (például személyneveken belül vezetéknevet és keresztnevet), kétszintű predikciós megközelítéseket implementáltunk és teszteltünk empirikusan. Ezen módszerek közül kettő egy-egy teljes modellt épít mindkét szintre, míg a harmadik módszer abban tér el a korábbiaktól, hogy a második szint minden egyes osztályára külön gépi tanuló modellt igényel.

- Habár a névelemosztályok többsége nem ismerhető fel jól reguláris kifejezések, listák illesztésével, ezek nagyban hozzá tudnak járulni egy statisztikai rendszer pontosságához. A szabályalapú módszereket többféleképpen kombináltuk statisztikai megközelítésekkel (feltételes valószínűségi mezőket alkalmaztunk), összesen három névelem-felismerő megközelítést vizsgáltunk.

Kiemeljük, hogy az egyes osztályok közti többértelműség igen magas volt, azaz a tulajdonnevek több osztályba is besorolhatóak. Ilyen esetek tipikusan a helységnevekkel kapcsolatosan fordulnak elő legtöbbször. A valós életben sok vezetéknévvel találkozhatunk, amelyek egyben településnevek is. Ilyen esetekben természetesen nem helyként szeretnénk klasszifikálni a kifejezést, hanem személynévként, de számos példa fordulhat elő helynevek és szervezetek (pl. *Szegedi Tudományegyetem*), vagy szervezetek és személynevek esetén is (pl. *Vörösmarty Mihály Általános Iskola*). Ezek a példák azért okoznak problémát, mert egy egyszerű listaillesztő rendszer képes felismerni helyként a *Szegedi* szót és szervezetként a *Szegedi Tudományegyetemet* is. Az általunk alkalmazott gépi tanulási módszer ezt kiküszöböli, a szövegkörnyezet alapján dönti el, hogy melyik jelölés alkalmasabb.

A dokumentumokban szereplő tokenek (szóalakok) képezték az osztályozás alapegységét, és rájuk épültek a gépi tanulási modellek. A tanító adatbázis kialakítása során a szövegek a whitespace és minden látható speciális karakter mentén lettek szavakra bontva.

Az egyes tokeneket bináris jellemzővektorok reprezentálták (részletes leírásukat [1] tartalmazza). A vektorban szereplő nullák és egyesek azt jelzik, hogy az adott tokenre teljesül-e az adott jellemző. A jellemzők megadására egy paraméterfájlon keresztül nyújt lehetőséget a rendszer[1]. Ebben a fájlban található minden olyan információ, amely a tanulási folyamat által használt jellemzőkhöz kapcsolódik.

A különböző rendszerek kombinációit empirikus módon értékeltük ki és vetettük össze. Az így kapott eredmények jól mutatják, hogy a probléma nehézsége ellenére a kapott jelölések jól közelítik az emberi annotációt.

2 Kétszintű címkézés

A kétszintű osztályozás miatt természetesen adódik az a lehetőség, hogy minden szintre egy külön modellt építsünk. A fő kizáró oka annak, hogy egy modellel hozzunk döntést az első és a másodrendű kifejezésekre is, az, hogy a két szint szoros kapcsolatban áll egymással.

Ahhoz, hogy képesek legyünk a szintezettségnek megfelelően címkézni, minden szinten különböző modellekre van szükség. Azonban a hierarchia miatt nem elegendő egyszerűen elfogadni az egyes modellek predikcióit, hanem szükséges még a szintekhez tartozó modellek kombinálása is.

[1] A paraméterezhető tulajdonnév-felismerő rendszer a Creative Commons licenc alatt elérhető: www.inf.u-szeged.hu/rgai/NER

A probléma megoldására egyetlen másodrendű modell helyett külön másodrendű modell készült az egy szülőhöz tartozó másodrendű jelölésekhez. Például egy-egy modell készült külön az előtag, vezetéknév, keresztnév osztályokhoz, amelyek a személynév első szintű jelöléshez tartoznak. Ez azt jelenti, hogy a tanító adatbázisból kigyűjtöttük a másodrendű címkével annotált kifejezéseket, úgy, hogy külön tanító adatbázis épült minden azonos elsőrendű szülővel rendelkező jelöléshez. Ezzel több új tanítóhalmaz alakult ki, amelyeknek száma megegyezett azon elsőrendű jelölések számával, amelyeknek léteznek leszármazottai (olyan első szintű osztályok is voltak, amelyekhez nem tartozott másodrendű leszármazott). Ezután az eredeti tanító adatbázist használtuk elsőrendű modell építésére, az újabb tanítóhalmazok segítségével pedig több kisebb modell készült. Ezek a kis modellek gyorsan felépültek, hiszen a tanító algoritmusnak nem kellett foglalkoznia azokkal a tokenekkel, amelyek egyik névelemosztályba sem estek (a névelem-kategóriákba eső kifejezések száma csak a töredéke azon tokeneknek, amelyek egyik tulajdonnévosztályba sem tartoznak), és képes volt kizárólag egy elsőrendű jelölés leszármazottaira fókuszálni.

3 Szabályalapú jelölés

A szabályalapú jelölések majdnem 20 címkére vonatkoznak (összesen több mint 50 első, illetve másodrendű osztály van), és egy-egy címkére több reguláris kifejezés is van definiálva. Az egyes felismerő kifejezések megírása különösen problémás volt, hiszen a nyelv változatossága miatt – akár az olyan szabványosnak vélt egyedek, mint a telefonszám is – több vagy összetett szabályok megírására volt szükség. Előnyük, hogy ha egy szövegrészre jelölést tesznek, az nagy valószínűséggel helyes is, azonban csak kevés egyedet fednek le.

Az ún. „Egymást követő" megközelítések a szabályalapú és a gépi tanulási módszereket külön-külön futtatják le, meghatározott sorrendben egymás jelöléseire adott megkötésekkel.

Az „RB + CRF" jelölés lényege, hogy elsőként a reguláris kifejezések illesztése történt a nyers szövegre, majd ezt követte a gépi jelölés (gépi tanuló modellként a Conditional Random Fields /CRF/-et használtuk [2]). Mivel a szabályok nagy pontossággal jó kategóriába sorolják a kifejezéseket, ezért abban az esetben, amikor a gépi tanuló modellnek is volt egy alternatív jelölése arra a kifejezésre, amelyre már a reguláris kifejezés illesztett, azt nem vettük figyelembe, és a szabály annotációját tekintettük érvényesnek.

A „CRF+RB" módszer esetén először a gépi tanuló modell jelölt, és csak utána következett a szabályalapú módszer. A modell nagy valószínűséggel jelöléseket végzett a szövegnek azon részein is, amelyekre egyértelmű szabályokat adtunk.

A „Bővített jellemzők" esetén (ez a módszer bizonyult a leghatékonyabbnak) a tanítás előtt a szabályalapú jelölő által készített jelölések bekerülnek a tanító adatbázisba a megfelelő tokenek mellé extra jellemzőként, így a statisztikai tanulóalgoritmus ezekkel a tulajdonságokkal egészítheti ki a jellemzőkészletét, majd a predikció során is lefutó szabályfelismerő növeli a felismerés pontosságát.

4 Eredmények

A tanuló adatbázis és a tesztfájlok kialakítására a rendelkezésre bocsátott 200 dokumentumból volt lehetőség. Mivel az összes dokumentum adatvédelmi okok miatt szigorú anonimizálási folyamaton esett át (pl. eredetileg egy nevet a „Vvvvv Kkkkk” karaktersorozatra cseréltek), így az anonimizált halmazon tanult modell és az ezen mért teszteredmények nem mutatnak pontos eredményeket.

A valós adatokon készített modell (a Szervezett Bűnözés Elleni Koordinációs Központban házon belül) és az azon végzett tesztelés összességében valamivel elmarad az anonimizált adatokon végzettétől, azonban vannak névelemosztályok, melyek esetében javultak az elért eredmények.

Az alábbi táblázatban látható az előző fejezetben bemutatott három különböző szabályalapú és gépi tanulási módszer kombinálására szolgáló módszer eredménye (néhány osztálycímkékre vonatkozóan és a végső rendszer pontosságára).

1. táblázat: Eredmények.

	RB + CRF	CRF + RB	Bővített jellemzők
Hely	78,12	78,5	79,72
Irányítószám	97,92	98,92	100
Város	85,25	84,43	94,71
Kerület	95,05	88,42	100
Utca	80,39	78,38	95,95
Házszám	81,25	77,73	98,29
Személynév	96,02	96,46	97,11
Előtag	89,06	89,6	88,72
Vezetéknév	95,19	95,19	98,2
Keresztnév	96,81	96,69	99,2
végső F-mérték:	**88,28**	**87,84**	**91,33**

Jól látható, hogy az „Egymást követő” módszerek közötti különbség elhanyagolható, viszont a „Bővített jellemzők” módszere szignifikánsan jobb eredményt produkált.

Bibliográfia

1. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In:, Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)

Igei igenevek problémája számítógépes nyelvészeti szempontból

Nádasdi Péter

Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
H-6722 Szeged, Egyetem u. 2.
nadasdi_peter@freemail.hu

Kivonat: Az igei igenevek nemcsak a leíró és elméleti nyelvészeti kutatások számára jelentenek problémát, hanem a gépi fordítás számára is, ezért fontos követelmény, hogy nem szavanként, hanem egységes szerkezetként kezeljük a vizsgált konstrukciót, így a gépi program csak azokat a szerkezeteket jelenítené meg, amelyek beletartoznak ebbe a típusba, de azok közül mindet visszaadná.

1 Bevezetés

A mai magyar nyelvben létezik egy olyan szerkezet (például *a Vágó István vezette vetélkedő*), amely egy főnév *(Vágó István)* és egy *-t/-tt* toldalékos verbális elem *(vezette)* szoros kapcsolatából áll, és ez a két elem együtt egy másik főnevet *(vetélkedő)* módosít. Feltűnő tulajdonsága, amely elkülöníti más jelzős és/vagy igeneves szerkezettől, hogy személyjelölő található a szerkezet igei tagján: *vezett-**e***. A szerkezet a gépi fordítás számára kihívást jelent, hiszen nincs egységes elképzelés e szerkezetről számítógépes nyelvészeti szempontból, ezért a szintaxis alkalmazása fontos tényező a vizsgált szerkezet számára.

2 A szerkezet jellemzői

Két fő elképzelés létezik erről a szerkezetről a generatív nyelvészeti szakirodalomban: igés (mai egyedüli képviselője [3]) és igeneves [1, 2], attól függően, hogy a szerkezet verbális elemét *(vezette)* igének vagy befejezett melléknévi igenévnek tartják. A szerkezet általános jellemzőit a következőkben foglalhatjuk össze:

1. A szerkezet egy nominális elem és egy verbális elem kapcsolata, amelyek mindig ebben a sorrendben követik egymást: *Vágó István* és *vezette.*
2. Egy ***-t*/*-tt*** toldalékos verbális elem található a szerkezetben: *vezet-**t**-e.*
3. Egy esetjelölés nélküli nominális elem van a szerkezetben: *Vágó István.*

4. A szerkezet több szóból is állhat, de tartalmaznia kell legalább egy nominális és egy verbális elemet, egyik vagy másik nélkül agrammatikus a szerkezet: *Vágó István vezette*[1].
5. Személyjelölő található a szerkezetben a verbális elemen: *vezett-**e***.
6. Jelzői szerepben áll egy módosított főnév (antecedens) előtt: *(Vágó István vezette) vetélkedő*.
7. A szerkezet két tagja között predikatív viszony van, azaz az esetjelölés nélküli nominális elem az alanyi argumentuma a verbális elemnek: a *Vágó István* a *vezette* elemnek.
8. A vizsgált szerkezetben tárgyas vagy alkalmilag tárgyas (egynél több argumentumú) igék fordulnak elő, például a nem tárgyas *lakik* is.
9. A szerkezetet általában névelő vezeti be, amely lehet határozott vagy határozatlan is, de alkalmilag el is maradhat: ***(a)** Vágó István vezette vetélkedő.*

A vizsgált konstrukció szintaktikai szerkezete tehát az alábbiak szerint ábrázolható:

Névelő	alárendelt (jelzői) tagmondat	módosított főnév (antecedens)
a	[*Vágó István* (tárgy $\varnothing_i$)*vezet-t-e*$_i$]	*vetélkedő*

A személyjelölő tárgyas személyragként értelmezhető, amely egy rejtett névmásra referál ($\varnothing$), és ez a rejtett névmás a módosított főnevet (*vetélkedő*) képviseli tárgyként a jelzői (alárendelt) tagmondatban.

3 Gépi fordítás

A szerkezet gépi fordítására rányomja a bélyegét, hogy nincs egységes elképzelés e szerkezetről, és nem kompakt egészként tekintenek erre a konstrukcióra. Az annotálása sem egyöntetű a szerkezet verbális tagjának, hiszen az MNSZ-ben igének és igenévnek is jelölik, vagy éppen ismeretlen kategóriának, ezért fontos lenne egységes jelölést alkalmazni. A félig kompozicionális főnév + ige szerkezetekkel mutat nagyfokú rokonságot ez a szerkezet [4], hiszen a kérdéses konstrukció is egy főnév *(Vágó István)* és egy ige(név) *(vezette)* kapcsolata: egy tagmondatot alkot *(Vágó István vezette),* és a szerkezet jelentését csak e tagmondat és a módosított főnév *(vetélkedő)* viszonyaként adhatjuk meg. Produktivitása és szintaktikai önállósága ellenére léteznek olyan szókapcsolatok, amelynek szavai gyakran fordulnak elő együtt, ezért is tekinthetők kollokációnak, mint például az *X.Y. vezette kormány, valamilyen természeti erő sújtotta terület.* A számítógépes nyelvészeti alkalmazásoknál is a legnagyobb problémát a kollokációk, illetve a kollokációszerű szerkezetek jelentik [4]. Nagyban megkönnyítené a gépi fordítását, ha nem szavanként, hanem egységes szerkezetként

[1] Legfeljebb igekötő (igemódosító) és tagadószó kerülhet a nominális és a verbális elem közé, mint például *a Lajos felásta föld* vagy *az ember nem lakta föld,* de adjunktum nem. A személyes névmásos alakoknál a személyes névmás ugyan elmaradhat, de ezek már nem produktívak, mint például *a szerette város.*

kezelnénk, így csak azokat a szerkezeteket jelenítené meg a gépi program, amelyek beletartoznak ebbe a konstrukcióba, de azok közül mindet visszaadná. A szerkezet ugyanis két részből áll: 1. egy beágyazott tagmondat *(Vágó István vezette)*, 2. egy módosított főnév *(vetélkedő)*. Természetesen fontos kiemelni, hogy ez a szerkezet nagyon közel áll a nem személyjelölt befejezett melléknévi igeneves szerkezetekhez, és ez a konstrukció könnyen átalakítható nem személyjelöltté: *a Vágó István által vezetett vetélkedő.* Éppen ezért idegen nyelvre, például angolra való fordításánál figyelembe kell vennünk, hogy kétféleképpen tudjuk elérni, hogy jó fordítás jöjjön létre: 1. posztnominális vonatkozó szerkezettel (angolban *wh*-konstrukció vagy *that),* 2. nem személyjelölt melléknévi igeneves szerkezettel (angolban *past participle+by*-os szerkezettel). A *(Vágó István vezette) vetélkedő* esetében a két változat az alábbiak szerint fordítható: 1. 'Az a vetélkedő, amelyet Vágó István vezet' = 'The quiz show (**that**/**which**) István Vágó present**s**.' 2. 'A Vágó István által vezetett vetélkedő' = 'The quiz show present**ed by** István Vágó.

A fenti konstrukció szintaktikai elemzése két fő részre osztja a szerkezetet: 1. *a vetélkedő,* 2. *Vágó István vezette,* ahol 1. egy DP, amelybe be van ágyazva egy többnyire két elemből álló tagmondat, amelynek az első eleme a beágyazott tagmondat alanya, a második eleme pedig az állítmánya. Fontos kiemelni, hogy vonatkozói szerkezetként való fordításánál az igei elemet a múlt idejű (befejezett melléknévi igenévi) forma ellenére is jelen idővel kell fordítani: *'amelyet rendszerint szokott vezetni'* (Simple Present). A gépi fordításnál elengedhetetlen, hogy a program felismerje, hogy az adott főnév és igei elem összetartozik [4]. Ennél a szerkezetnél pedig ezután azt is kódolnia kell, hogy a módosított főnév is a szerkezet része, de nem annak argumentuma, hiszen a módosított főnév bármilyen esetet felvehet, aszerint, hogy milyen mondatba foglaljuk bele: *Láttam a Vágó István vezette vetélkedő**t**.* Ugyanis ez az ige+főnév szerkezet egy beágyazott tagmondat, amely a főnevet (antecedens) módosítja. A Google fordító és a *webforditas.hu* a példamondatunk fordításakor rosszul formált szerkezetet hoz létre, viszont például *az Angela Merkel vezette kormány* esetében a két program közül a Google fordító jó fordítást hoz létre: *the government led by Angela Merkel*, de csak az igenevest alkalmazza, viszont ha kicseréljük a módosított főnevet az alábbi módon: az *Angela Merkel vezette vetélkedő,* akkor már rossz eredményre jut. Feltételezhető, hogy nem szerkezeti elemzéssel, hanem heurisztikákkal dolgozik, és statisztikai alapon jut jó megoldásra.

4 Összegzés

A cikkben azt a problémát vizsgáltam, hogy a kérdéses szerkezet kihívást jelent a számítógépes nyelvészeti alkalmazás szempontjából a gépi fordítás, illetőleg az annotálás számára, és a szintaxis alkalmazását tartom fontos lépésnek a konstrukció szempontjából a nehézséget jelentő megoldandó feladatok számára.

Hivatkozások

1. Kenesei, I.: On the Role of the Agreement Morpheme in Hungarian. Acta Linguistica Hungarica Vol. 36 (1986) 109–120
2. Laczkó T.: Néhány lexikai-funkcionális gondolat a személyragos -(*t*)*t* igenévről. In: Andor J. et al. (szerk.): Színes eszmék nem alszanak… Szépe György 70. születésnapjára. Lingua Franca Csoport, Pécs. (2001) 741−759
3. Nádasdi P.: Mozgatás nélkül? Egy prenominális vonatkozó szerkezet minimalista elemzése a magyarban. In: Gárgyán G., Sinkovics B. (szerk.): LingDok5. Nyelvész-doktoranduszok dolgozatai. JATEPress, Szeged (2006) 61–86
4. Vincze V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. In: Tanács A. et al. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009). JATEPress, Szeged (2009) 390–393

Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból különböző módszerek segítségével

Nagy Ágoston[1,2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
nagyagoston@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola,
Francia nyelvészet alprogram
nagyagoston@lit.u-szeged.hu

Kivonat: A cikk egy francia nyelvre készült saját, elsősorban szabályalapú, de statisztikai szűrőkkel is rendelkező terminológiakivonatoló leírását és eredményeit tartalmazza. Célunk annak feltárása, hogy a tisztán szabályalapú terminológiakivonatoló főmodulon kívül alkalmazott szabályalapú és statisztikai módszerek milyen mértékben járultak hozzá a fedés és a pontosság növeléséhez (vagy csökkenéséhez). A terminusok szabályalapú kinyerése véges állapotú automatával történik, a kimenet szűrése pedig először *stopword*-listával, majd tulajdonnév-felismerő modul alkalmazásával. A statisztikai módszereket szűrésre alkalmazzuk: a *unithood* érték mérésére a C és NC értékeket, a *termhood* mérése a *weirdness* arány segítségével valósul meg.

1 Bevezetés

A terminológiakivonatolás (a továbbiakban TE) során a TE-alkalmazás egy adott, írott nyers szövegből annak terminusjelöltjeivel tér vissza. A terminusjelöltek kinyerése és szűrése történhet szabályalapú és statisztikai módszerekkel. A leggyakrabban használt módszer ezek kombinációja, a hibrid módszer, ami [2] és [6] szerint először a terminusjelöltek kinyerésére a statisztikát alkalmazzák, majd azok szűrésére nyelvi filtereket.

2 Korpusz

Korpusznak négy, francia nyelvű informatikai témájú szabadalom leírását választottuk, amelyekben 854 különböző terminus található; a szövegek átlagosan 3500 tokennel rendelkeznek. A négy szabadalmat kézileg is annotáltuk: bejelöltük bennük az összes terminust.

3 Módszer

A vizsgálatunk során a terminusok kinyerésekor a megszokottól eltérően fordított sorrendet alkalmazunk: a terminusjelölt-listát szabályalapú módszerekkel nyerjük ki, majd ezt különböző szűrőkkel szűrjük a pontosság növelése érdekében. A szabályalapú kinyeréshez és szűréshez szükség van a szöveg előfeldolgozására, amelyhez a szöveget mondatokra, majd tokenekre bontó, illetve azokat szófaji címkékkel ellátó *Machinese*-t [3] használtuk.

3.1 Terminusjelölt-lista létrehozása és első szűrése szabályalapú módszerekkel

A terminusjelöltek listájának kinyeréséhez a leggyakoribb mintákból (pl. főnév+főnév, főnév+prepozíció+főnév) véges állapotú automatát hozunk létre. Ezt az automatát illesztjük a már szófaji címkékkel ellátott szövegre. Az *és/vagy* típusú koordinációkat visszaállítjuk az eredeti alakjukra. Az így kapott mintákat szűrjük egy *stopword*-listával, ami a leggyakoribb (főnevet is tartalmazó) kifejezéseket szűri ki a szövegből, hogy azok ne kerülhessenek be a terminusjelölt-listába. Ilyen típusú szerkezetek a *par exemple* 'például', *en effet* 'ugyanis' stb. A tulajdonneveket pedig az OpenCalais projekt keretében létrehozott *OpenCalais Web Service API* [8] nevű alkalmazással szűrjük.

3.2 Terminusjelölt-lista szűrése statisztikai módszerekkel

A terminusjelöltekre a C és NC [5], *weirdness* [1] értékek kiszámítására szolgáló algoritmust alkalmazzuk. Mindhárom értékre igaz az, hogy minél nagyobb egy adott terminusnál annak értéke, annál valószínűbb, hogy az adott jelölt ténylegesen terminus.

A C-érték egy *unithood* mérték, ami azt mutatja meg, hogy egy adott terminusjelölt gyakrabban fordul-e elő önmagában vagy egy nagyobb egység részeként. Így például kiszűrhetők azok a melléknévi utómódosítók, amelyek az adott terminusnak nem lehetnek részei, mert a terminus részét nem képező melléknévi utómódosító és a főnévi fej közötti kohéziós érték alacsony lesz, ha ritkán fordulnak elő együtt.

Az NC-érték azt vizsgálja meg, hogy az adott terminusjelölt környezetében lévő szavak milyen valószínűséggel jelzik azt, hogy előttük vagy mögöttük terminus áll.

A *weirdness* pedig egy olyan *termhood* mérték, amely azt mutatja meg, hogy az adott terminusjelölt az adott szakszövegben vagy egy általános nyelvű korpuszban fordul elő gyakrabban. Ehhez egy általános keresőmotort használunk, az Exalead vállalat online keresőjét [4]: a saját alkalmazásunk minden egyes terminusjelöltnél lekérdezi annak gyakoriságát egy köznyelvi újság, a Le Figaro weboldaláról [7]. A Le Figaro keresési feltételként történő megadásának célja, hogy a keresőmotor ne kereshessen bárhol, hiszen így szakmai szövegekben is keresne, amit el kell kerülni.

A fent említett mértékeket először külön-külön alkalmazzuk az adott korpuszra, és megnézzük, hogy milyen hatékonyság érhető el ezeknél. Megkeressük minden változónál azt a határértéket, amely felett a legjobb a pontosság, fedés, illetve F-érték. Ezt követően egy összevont értéket is alkalmazunk, ami minden érték együttes eredményét veszi alapul.

4 Eredmények

A terminológiakivonatolás esetén a fedés a helyesen kinyert terminusok számának és az adott szövegben lévő terminusok számának a hányadosa, a pontosság a helyesen kinyert terminusok és az összes kinyert terminusok számának hányadosa, az F-érték pedig a fedés és pontosság szorzatának duplája osztva a fedés és a pontosság összegével [2].

A tisztán szabályalapú algoritmussal, tehát a mintákkal, körülbelül 0,78-as fedést és 0,59 értékű pontosságot (F-érték: 0,67) érhetünk el. A fedés és a pontosság értékei már akkor jelentősen nőnek, ha a mintaillesztés után a terminusjelölteket szűrjük az előre megadott, főnevet is tartalmazó fordulatokkal, valamint a benne szereplő tulajdonnevekkel: ekkor a pontosság 0,66 a fedés 0,83 (F-érték: 0,74). A statisztikai módszerek a várt eredményeket hozták: a pontosságot tudták növelni, de ezáltal a fedés csökkent. A legjobb pontosságot az általunk létrehozott kombinált érték biztosította, mely által ez az érték 0,89 lett. Az 1. táblázat foglalja össze az eredményeket a különböző algoritmusok esetén, ahol a legjobb értéket vastaggal emeltünk ki.

1. táblázat: Fedés, pontosság és F-érték a különböző módszerek esetén.

alkalmazott módszer	határérték	fedés	pontosság	F-érték
kinyerés mintákkal	-	0,7834	0,5895	0,6728
kinyerés mintákkal + szabályalapú szűrés	-	0,8285	0,6609	0,7353
weirdness				
	-	**0,8285**	0,6609	0,7353
	> 0,2595	0,7109	**0,6901**	0,7003
	> 0,0011	0,8285	0,6626	**0,7363**
C-érték:				
	-	**0,8285**	0,6609	0,7353
	> 2,8074	0,4574	**0,6917**	0,5506
	> -6,3399	0,8274	0,6618	**0,7354**
NC-érték:				
	-	**0,8285**	0,6609	**0,7353**
	> 1,5388	0,5620	**0,7098**	0,6273
C-NC érték				
	-	**0,8285**	0,6609	0,7353
	> 2,3807	0,4682	**0,6922**	0,5586
	> -4,8251	0,8274	0,6618	**0,7354**
kombinált érték				
	-	**0,8285**	0,6609	0,7353
	> 0,8468	0,0701	**0,8904**	0,13
	> 0,0867	0,8123	0,6759	**0,7379**

Bibliográfia

1. Ahmad, K., Gillam, L., Tostevin, L. Weirdness indexing for logical document extrapolation and retrieval (wilder). In: The Eighth Text REtrieval Conference (TREC-8). (1999) 717–724
2. Cabré, M. T., Bagot, R.E., Vivaldi Palatresi, J.: Automatic term detection. A review of current systems. In: Bourrigault, D., Jacquemin, Ch., L'Homme, M-C. (szerk.): Recent advantages in Computational Terminology. John Benjamins Publishing Co., Amsterdam/Philadelphia (2001) 53–87
3. Connexor – Technology – Machinese – Demo, http://www.connexor.eu/technology/machinese/demo/
4. Exalead search, http://www.exalead.com/search
5. Frantzi, K. T., Ananiadou, S.: The c/nc value domain independent method for multi-word term extraction. Journal of Natural Language Processing Vol. 6, No. 3 (1999) 145–179
6. Ha, L.A., Fernandez, G., Mitkov, R., Corpas, G.: Mutual bilingual terminology extraction. In: Calzolari, N. et al. (szerk.): Proceedings of LREC 2008 (CD-ROM). ELRA, Marrakech (2008) 1818–1824
7. Le Figaro online, http://www.lefigaro.fr/
8. OpenCalais Web Service API, http://www.opencalais.com/documentation/calais-web-service-api

Szótáralapú kémiai NE-felismerő rendszer

Nyilas Sándor, Németh Gábor, Almási Attila

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{nyilasster, nemeth.gabor3, vizipal}@gmail.com

Kivonat: A MASZEKER projekt szabadalmakon futó szemantikus keresőrendszer kifejlesztését célozta meg, melynek az orvostudományi és kémiai szabadalmak esetében egyik lényegi lépése a kémiai névelemek felismerése. Ehhez szükség volt egy szótárfájl létrehozására, mivel a névelemeket jelölő program nem elemzi szemantikusan a mondatok szavait, és nem az alapján dönt, hogy melyik szó kémiai névelem és melyik nem. A szótárfájl, amely soronként elkülönített szavakból áll, tartalmazza a kémiai névelemeket. Ennek a szótárfájlnak a rendszeres frissítése és karbantartása szükséges ahhoz, hogy a program minden kémiai névelemet fel tudjon ismerni.

1 A szótárfájl előállításáról

A kémiai vegyületneveket tartalmazó szótárfájlt az Environmental Chemistry oldalról[1] gyűjtöttük ki, mely soronként egy vegyületnevet és egy tabulátorral elválasztott egy- vagy kétjegyű előfordulási számot tartalmaz, amelyet eddig figyelmen kívül hagytunk.

1.1 Az eredeti szótárfájl

A szótárfájlból több névelem (NE) is hiányzott (pl. *sodium*), ezért szükség volt a szólista bővítésére. Mivel más kémiai névelem-adatbázis nem állt rendelkezésünkre, de feltételeztük, hogy az összetett kémiai NE-k tartalmazzák a hiányzó, elemibb NE-ket (pl. a (*9-Octadecenoic acid (Z)-, iron(3+) salt*) tartalmazza az *iron*-t), ezért az összetett NE-k felbontása mellett döntöttünk. Ehhez szükség volt egy bővítmény létrehozására, melyet az alábbiak szerint hajtottunk végre: az eredeti szótárfájlról két másolatot készítettünk, melyeken a következő változtatásokat hajtottuk végre:

- kis- és nagybetűket nem megkülönböztetve ábécérendbe rendeztük az NE-ket,
- minden sor végéről a tabulátort és az utána szereplő számot töröltük,
- az 1. másolatban minden nem szám- és betűkaraktert kicseréltünk szóköz karakterre, az összes szóközt sortörésre cseréltük, s az így kapott szavakat a következő algoritmussal szűrtük:
  - a tördelésnél keletkezett felesleges szavak közül kivettük a duplikátumokat;
  - a könnyebb kezelhetőség kedvéért a dupla sortöréseket is töröltük;

[1] http://environmentalchemistry.com/yogi/chemicals/

• az 1. másolatot hozzáadtuk a 2.-hoz, majd újból szűrtük a dupla sortöréseket és a duplikátumokat;

• a 2. másolatból a felesleges szavakat kézileg eltávolítottuk. E 2. másolatot a beillesztést követően az első verziójú szótárfájlnak tekintjük.

Mivel a program nem tett különbséget a rövidítéseknél a kis- és nagybetűk között, helytelenül került feljelölésre az *at* mint prepozíció és helyesen az *At* mint *Astatine*. A hiba kiküszöbölése érdekében a szótárfájlt kettévágtuk a legfeljebb három, illetve az annál több karakterből álló szavakra. Ezentúl két lista létezik, melyeket együttesen nevezünk szótárfájlnak. Az első listafájlt (három és annál kevesebb karakterből álló szavak), kis- és nagybetűt megkülönböztetve, a másikat (háromnál több karakterekből álló NE-k) továbbra is kis- és nagybetűt nem megkülönböztetve vizsgáljuk. A két listafájl jelenleg 81959 vegyületnevet tartalmaz.

2 A program működése

A szótárfájlt és a vizsgált szabadalmat betöltjük a programba, majd megvizsgáljuk, hogy a szótárfájlból soronként betöltött NE-k megtalálhatók-e a vizsgált szabadalomban. Az eredmények javításához feszítő-szűrő szabályrendszert alkalmazunk.

2.1 Feszítés

Mikor a szó átadódik a feszítő algoritmusnak, már biztosak lehetünk abban, hogy vegyületnevet találtunk. Először a kezdő-, majd a záróindex értékét változtatja a kód, értelemszerűen a kezdőindexet csökkentve, a záróindexet pedig növelve azért, hogy a vegyületnév-töredék indexeit ráfeszítse az egész vegyületnévre. A folyamat akkor áll meg, ha a program megfelelő karakterpárt talál. Ezek a karakterpárok a vegyületnév elejét vagy végét jelzik. Ha egy szó után szünet van, és a következő karakter valamilyen betű vagy szám, akkor ott a vége a vegyületnévnek. Ugyanez igaz a szó elejével kapcsolatban is. Ezen kívül a következő kritériumok esetében áll meg a feszítés, és dönt úgy, hogy a vegyületnévnek az adott helyen eleje vagy vége van:

eleje:

- <bármilyen karakter és/vagy szám> és szóköz
- <írásjelek (, . ! ? (stb...)> és szóköz
- <sorvége, kocsivissza, \0 jelek> és szóköz
- pontosvessző

vége:

- szóköz és <bármilyen karakter és/vagy szám>
- szóköz és <írásjelek (, . ! ?) stb...)>
- <szóköz, írásjelek> és <sorvége, kocsivissza, \0 jelek>

2.2 Szűrés

Mikor a feszítés befejeződik, átadja a találatok kezdő- és végindexeinek listáját a szűrő algoritmusnak, mely azért felel, hogy a találatok közül mindig csak a „legbővebb" legyen feljelölve (pl. csak a (*Threitol 1,4-bis (methanesulfonate*)) és ne a *Threitol* és *methanesulfonate* külön-külön), ezért egy algoritmussal a kezdő és az ahhoz tartozó végindexeket rendezi. Ezután a program addig hasonlítja össze a találatokat, ameddig a szűrő már nem változtat a találati listán.

A szűrő egyik funkciója az, hogy megvizsgálja, hogy az n-edik találatnak a k kezdőpontjához és v végpontjához képest hol helyezkedik el az n+1-edik találatnak a k2 kezdőpontja és a v2 végpontja. A sorba rendezés miatt alapfeltevés, hogy k<k2:

- ha v<k2, akkor nem változik
- ha v>=k2, de v<=v2, akkor v-t egyenlővé tesszük v2-vel és az n+1 találatot töröljük
- ha v>=k2, de v>v2, akkor v megtartja az értékét, és az n+1 találatot töröljük.

Ha két találat közt csak egy karakter távolság van, akkor a kettőt egynek veszi, és megkapja a két találat legkisebb kezdőértéket és a legnagyobb végértéket.

3 A névelemek annotációja során felmerült problémákról

A főigénypontokban szereplő NE-ket három csoportba rendeztük: 1) kémiai elemek (nitrogén), elemcsoportok (halogének), vegyületek (Na_2O) stb.; 2) általános anyagnevek (só), vegyületfajták (szénhidrát) stb.; 3) konkrét betegségek (Alzheimer-kór), betegségcsoportok (immunhiányos betegségek) és tünetek (másnaposság).

Néhány gyakoribb hibatípus [1]:

- A program nem különíti el a NE-k főnévi és jelzői használatát: pl. *antibiotic* – az angolban főnév és melléknév is lehet; főnévi használatban jelöltük csak.
- Az előforduló helyesírási hibák miatt a program nem megfelelően szegmentál bizonyos elemeket: pl. *…alkarylamino, fluoro, chloro, bromo iodo and trifluoromethyl…* – két, egyébként külön jelölendő NE-t egynek vett; az annotáció a szándékolt tartalomnak megfelelően történt.
- Szófaji problémák:
    1. *water-soluble*, *wax-like* kifejezések a magyarban nem NE-k – nem jelöltük;
    2. *carboxylic*, *enantiomeric* jelzők – nem jelöltük;
    3. *O-glycosidically* határozószó – nem jelöltük.

4 Eredmények

A névelem-felismerő program találati pontossága a fejlesztésekkel rohamosan nőtt. Ezt egy segédrendszerrel teszteltük, amely összehasonlította a kézzel annotált 313 dokumentumot az automatikusan feljelöltekkel. A két feljelölés között jelentős számbeli különbség mutatkozott: pl. a *salt* a kézi feljelölés alapján 17, a gépi feljelölés szerint pedig 43 alkalommal fordul elő NE-ként.

Eddig négy különböző programverzió készült, a negyedik még tesztfázisban van.

- ***verzió 0.1***: szűrés és feszítés nélkül a program a tesztelés során csak ~ 6,5%-ot ért el. A programban nem volt szűrési rendszer: a gépi megjelölés jóval több, mint a kézi, mert összetettebb vegyületnevek esetén az elemibb NE is feljelölésre kerültek, és az is annotációnak számított.
- ***verzió 0.2***: a szűrőrendszer beiktatása után már 70%-os teljesítményt ért el a program. A gépi annotációk száma számottevően kisebb volt, mint a kézzel feljelölteké. Szükség volt a szótárfájl bővítésére és egyben szűrésére is, mert még ~ 2000 feljelölés fölösleges, illetve helytelen volt.
- ***verzió 0.3***: a szótárfájl bővítése és szűrése után az F-mérték 90.13%-ra javult.
- ***verzió 0.4***: finomítottunk az annotálási elveken és három kategóriát vettünk fel: ***speciális NE-k***, ***általános NE-k***, ***betegségek*** (l. 3. fejezet). A tesztanyagban a kézzel jelölt annotációkat a fentieknek megfelelően módosítottuk. A szabályrendszer átalakítására nem volt szükség, csupán a program beolvasási és osztályba sorolási rendszerén kellett változtatni. Az NE-k és betegségek megkülönböztetése érdekében a szótárfájlt – a korábbi kettő helyett – négyfelé vágtuk, majd beillesztettük a programba a négy fájl beolvasását. Amikor a program egy NE-t felismer, feljelöli, és besorolja a megfelelő osztályba. Az, hogy melyik osztályba kerül egy NE, kizárólag attól függ, hogy a kifejezés melyik szótárfájlból származik. A betegségek szótárfájlt bővíteni kell egy másik adatbázis[2] segítségével. A kézi jelölés módosítása után az F-mérték 95.25%-ra javult.

1. táblázat

	verzió 0.1	verzió 0.2	verzió 0.3	verzió 0.4
Gépi NE-k száma	17 779	9 799	11 407	10 373
Kézi NE-k száma	10 874	10 874	10 874	11 355
Helyes NE-k száma:	932	7 306	10 041	10 348
Precision / Recall:	8.57 / 5.24	67.18 / 74.56	92.33 / 88.02	91.13 / 99.75
F:	6.50	70.68	90.13	95.25

[2] http://www.who.int/classifications/icd/en/

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Vincze V., Nagy Á., Klausz Á., Almási A., Kiss M.: Nyelvészeti problémák a szabadalmak feldolgozásában. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2010) 168–179

Lényegkiemelő módszerek összehasonlítása közlekedési zajban történő beszédfelismerés céljából

Sárosi Gellért[1], Tobler Zoltán[2], Mihajlik Péter[1,2], Fegyó Tibor[1,3]

[1] Távközlési és Médiainformatikai Tanszék,
Budapesti Műszaki és Gazdaságtudományi Egyetem
{sarosi, tobler, mihajlik, fegyo}@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.
[3] Aitia International Inc.

Kivonat: A gépi beszédfelismerés egyik döntő fontosságú eleme a beszéd akusztikai lényegének kiemelése, különösen a zajos környezetben történő alkalmazásoknál, amely jelen esetben közlekedési zajjal terhelt akusztikai környezetet jelentett. Emiatt helyeztük vizsgálatunk középpontjába a zajtűrő és hagyományos beszédfelismerési lényegkiemelési eljárásokat. A tanítást és tesztelést hat nyelven végeztük el: angol, francia, magyar, német, olasz, spanyol. Teszteléshez a telefonos hálózaton keresztül az utcáról vagy járműből rögzített adatbázist használtunk. Alaprendszerként teszteltük a HTK és a SPHINX eszközkészletben, vagy általunk is implementált Mel Frequency Cepstral Coefficients (MFCC) és Perceptual Linear Prediction (PLP) módszereket. Az újabb módszerek között a Power-Normalized Cepstral Coefficients (PNCC) és a Perceptual Minimum Variance Distortionless Response (PMVDR) szerepel.

1 Bevezetés

Feladatunk közlekedési zajban üzemelő folyamatos beszédfelismerő rendszer összeállítása. A rendszernek hat nyelven kell működnie: angol, francia, magyar, német, olasz és spanyol. A cél: felismerni a nyilvános mobiltelefon-hálózaton érkező hívásokban, hogy a hívók milyen célobjektumot (POI – Point of Interest) szeretnének megtalálni, mint például egy múzeumot, éttermet vagy konkrét címet. A rendszernek legalább a POI-k többségét megbízhatóan fel kell ismernie annak ellenére, hogy az utcán sétálva, vagy valamilyen járműben utazva a beszédkörnyezet legtöbbször zajjal terhelt.

2 A lényegkiemelők

Az automatikus beszédfelismerés kritikus lépése a lényegkiemelés, hiszen ekkor alakítjuk át a beszédet a gép számára feldolgozható lényegvektorok sorozatává. Emiatt helyeztük kísérleteink középpontjába különféle lényegkiemelő eljárások vizsgálatát.

A *Mel Frequency Cepstral Coefficients* (MFCC) egy elterjedten alkalmazott módszer, sokféle implementációja létezik, ezek közül hárommal foglalkoztunk. Az egyik a HTK (Hidden Markov-Model Toolkit) [10] nevű, rejtett Markov-modellek építésére és manipulációjára alkalmas eszközkészlet. A munkánk során részint a beépített lényegkiemelő eszközöket, részint az akusztikus modelltanító és -kiértékelő eszközöket használtuk fel. A másik a SPHINX [1] nevezetű, kifejezetten beszédfelismerésre készült rendszer, ennek csupán a lényegkiemelő részét használtuk fel. A harmadik a saját implementációnk, mely a Voxerver[1] nevezetű felismerő szoftver része. Mindhárom módszer magja a Mel-szűrőbank és a logaritmikusamplitúdó-kompresszió.

Zaj szempontjából robosztusabb megoldást kínálhat a *Perceptual Linear Prediction* (PLP) módszer [2], mely lineáris predikciót (LP) használ a beszéd spektrális burkolójának előállításához. A perceptualitást a Bark-szűrés és – a hallás frekvenciával változó érzékenységét követő – azonos hangosságú előkiemelés adja.

Az újabb módszerek között szerepel a *Perceptual Minimum Variance Distortionless Response* (PMVDR) [9], mely szűrés helyett egy paraméterezhető ún. frekvenciahajlítást (frequency bending) alkalmaz, a LP-együtthatókból pedig MVDR-spektrumot, egy felső spektrális burkolót számít.

Szintén új módszer a *Power-Normalized Cepstral Coefficients* (PNCC) [3], amely ún. Gammatone-szűréssel [7], teljesítményeltolással és exponenciális amplitúdóösszenyomással reprezentál egy zajrobosztus lényegkiemelést.

3 A kísérleti környezet

Tanítási célokra a SpeechDat [8] adatbázist használtuk, mely az általunk vizsgált nyelveken, egyenként 500-5000 beszélőtől származó, a vezetékes és a mobilhálózaton keresztül is rögzített felvételeket tartalmaz. Ez alól kivétel a magyar nyelv, amelynél az akusztikai modellek tanításához az MTBA-t (Magyar Telefonos Beszéd Adatbázis) [4] használtuk. Ez 500 beszélőtől tartalmaz felolvasott szöveget szintén a telefonhálózaton keresztül rögzítve, tehát teljességgel SpeechDat-szerű, és felhasználható a kísérleteinkben. Mind a tanító-, mind a tesztadatokra közösen jellemző a 8kHz mintasűrűség, az egy csatorna és a 8 bit A-law kódolás. Az angol nyelv esetén fellépett adatelégtelenségi problémák miatt ezt a nyelvet kivettük a zajtűrési vizsgálatainkból. A teljes adatbázisból 10 órányi adatot használtunk fel a tanításra, mert kísérletünk célja az egymáshoz viszonyított javulások vagy romlások feltérképezése, nem pedig az abszolút legjobb felismerés, ez esetben pedig az adatbázis mérete nem kritikus.

A felismerési teszteket két szakaszban hajtottuk végre. Elsőre verifikációs teszteket végeztünk magyar nyelven, hogy beállítsuk az optimális tanító és tesztelő környezetet. E célból felhasználtunk mintegy 15 percnyi hanganyagot egy magyar nyelvű műsorszóró adó híradójából, és ugyanennyi telefonos hanganyagot. A tesztek lényegi részét a többnyelvű felismerések adták, melyekhez a telefonos hálózaton keresztül az utcáról vagy járműből rögzített, tájékozódási célú kérdésekből és kijelentésekből álló, alacsonyabb jel-zaj viszonyú adatbázist használtunk. Tartalmuk egy-egy POI megta-

[1] Aitia International Inc.

lálásához kapcsolódó kérdés vagy jellemzés, de vannak POI-t nem tartalmazó bemondások is a szófelismerés pontosabb méréséhez.

Miután a tanító adatokon lefuttattuk a lényegkiemelést, fonémaszintű címkézés és flat-start módszer [10] alkalmazásával is Maximum Likelihood módszerrel tanítottunk GMM (Gauss Mixture Model) alapú, szóhatárokon átívelő trifón akusztikai modelleket. A felismerési teszteket a már korábban említett Voxerverrel végeztük, mely egy WFST (Weighted Finite State Transducer) [5] alapú dekóder szoftver. Az akusztikai modell mellett WFST alapú nyelvtanokat használtunk, melyeket a lehetséges kérdező, kérő mondatszerkezetekből és POI-kifejezésekből generáltunk a [6] módszerei szerint.

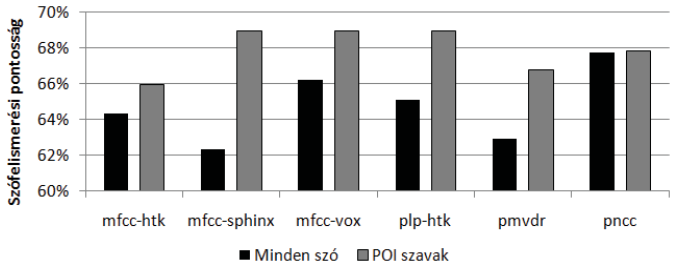
4 Eredmények

Minden kísérletben 39 dimenziós, az energia jellemzőt is magában foglaló lényegvektorokat állítottunk elő. Az akusztikai modelleket alkotó Gauss-függvények számát 10-ben maximalizáltuk. A nyelvenként és módszerenként lefutott legjobb felismerési eredményeket a 1. táblázat tartalmazza, ahol WAcc a szófelismerési pontosságot jelenti, az *All* és *POI* pedig hogy az eredmény minden szóra vagy csak a POI-kra vonatkozik. Az első három sorban a három MFCC változat, a másik háromban a zajrobosztus frontendek pontosságai szerepelnek, kiemelve a nyelvenkénti legjobbat.

1. táblázat: Felismerési eredmények.

WAcc	Francia		Magyar		Német		Olasz		Spanyol	
(%)	All	POI	All	POI	All	POI	All	POI	All	POI
mfcc-htk	56.8	70.3	70.1	57.4	62.6	56.5	62.7	**81.2**	75.4	77.6
mfcc-sphinx	**65.9**	74.4	68.3	60.6	59.5	69.6	57.7	73.9	71.6	65.7
mfcc-vox	**65.9**	70.3	67.5	56.9	**69.3**	63.0	62.1	76.1	77.1	76.1
plp-htk	61.4	**75.2**	71.1	**62.3**	66.3	69.6	**63.7**	70.3	79.8	**85.1**
pmvdr	65.5	74.4	69.7	59.7	62.0	**71.7**	59.3	75.4	68.3	56.7
pncc	64.8	70.3	**71.3**	61.7	67.5	67.4	59.5	68.1	**83.6**	80.6

A legkiemelkedőbb a magyar 71.3%-os szófelismerési arány, melyet viszonylag nagyobb tesztadatbázis mellett sikerült elérni. A spanyol 83.6% is figyelemre méltó, de a kevesebb tesztadat miatt kevésbé megbízható. Az olasz, francia és német adatokon is magas POI-pontosságot értünk el, de általában az összes szó felismerése ettől nem marad el jelentősen.



1. ábra. Felismerési eredmények módszerenkénti átlagai.

A kapott értékekből módszerenkénti átlagot képeztünk, hogy megkeressük a globálisan optimális lényegkiemelő módszert. Ez látható az 1. ábrán. Az átlagosan legjobban teljesítő módszer a PNCC, legmagasabb átlagos szófelismerési pontossággal. Még a PNCC-nél is jobb POI-felismerést adott a HTK PLP és saját MFCC-implementációnk. Az összes szót tekintve már rosszabban teljesítettek, viszont a saját MFCC-implementációnk jobb teljesítményt mutatott a másik két változatnál. A SPHINX MFCC módszere szintén kiemelkedő POI-pontosságot ért el, de az összes szót tekintve a leggyengébben szerepelt. A PMVDR és HTK MFCC átlagosan gyengébben teljesített, bár az eredmények közti eltérések nem jelentősek.

5 Összefoglalás

Öt nyelven készült el egy olyan beszédfelismerő rendszer, amellyel nyelvenként hatféle lényegkiemelési módszer szerint végeztünk kísérleteket. Eredményeink alapján a PNCC teljesített a legjobban, mert sok nyelv esetén a legjobb, vagy ahhoz közeli felismerést adott. Szintén kiemelkedő a Voxerver MFCC és HTK PLP teljesítménye, de átlagban kissé elmaradnak a PNCC-től. Ráadásul a Voxerver MFCC jobban teljesít a másik két implementációnál. Az angol rendszer gondjai kevés fejlesztéssel megoldhatóak, és az abszolút felismerési eredmények is tovább javíthatóak, ha a teljes adatbázist felhasználjuk a tanításhoz, ezt tekintjük a lehetséges folytatás fő irányának.

Köszönetnyilvánítás

Kutatásainkat részben támogatták: OM-00102/2007, OMFB-00736/2005, TAMOP-4.2.2-08/1/KMR-2008-0007.

Bibliográfia

1. CMU Speech Recognition Engine (SphinxTrain 1.0): http://www.speech.cs.cmu.edu/

2. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America Vol. 87 No. 4 (1990) 1738–1752
3. Kim, C., Stern, R. M.: Feature Extraction for Robust Speech Recognition using a Power-Law Nonlinearity and Power-Bias Subtraction. In: INTERSPEECH (2009) 28–31
4. Magyar Telefonos Beszéd Adatbázis: http://alpha.tmit.bme.hu/speech/hdbMTBA.php
5. Mohri, M., Pereira, F., Riley, M.: Weighted Finite-State Transducers in speech Recognition. Computer Speech and Language Vol. 16 No. 1 (2002) 69–88
6. Mozsolics T., Tarján B., Mihajlik P., Fegyó T.: Környezetfüggetlen és sztochasztikus nyelvtanok összehasonlítása többnyelvű gépi beszédfelismerési feladatban. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 203–215
7. Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M. H.: Complex sounds and auditory images; Auditory and Perception (1992) 429–446
8. SpeechDat(II) telephone network database. http://www.speechdat.org/SpeechDat.html
9. Yapanel U. H., Hansen, J. H.L.: A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition In: EUROSPEECH (2003) 1281–1284
10. Young, S., Ollason, D., Valtchev, V. and Woodland, P.: The HTK book. (for HTK version 3.4), March 2009. http://htk.eng.cam.ac.uk

Valós idejű szövegosztályozás a Wikipédia szolgálatában

Solt Illés[1], Héder Mihály[2], Tikk Domonkos[1,3]

[1] Budapesti Műszaki és Gazd.tud. Egyetem, Távközl. és Médiainf. Tansz. (TMIT), H-1117 Budapest, Magyar Tud. krt. 2, e-mail: {solt,tikk}@tmit.bme.hu
[2] MTA SZTAKI, Internet Technológiák és Alkalmazások Központ (ITAK), H-1132 Budapest, XIII. Victor Hugo u. 18–22, e-mail: {mihaly.heder}@sztaki.hu
[3] Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics (WBI), D-10099 Berlin, Unter den Linden 6, e-mail: {tikk}@informatik.hu-berlin.de

A szöveges, azaz humán felhasználásra szánt információk hozzáférhetőségén javíthat, ha a szövegek több aspektus mentén is lekérdezhetőek. Ilyen rendszerre példa a Wikipédia közösségi enciklopédia, melyben az információ egysége a szócikk, melyek nem csak kereszthivatkozások mentén, hanem egy kategóriarendszer mentén is böngészhetőek. A szócikkek kategóriákba sorolásának koherenciája és a kategóriarendszer megválasztása határozza meg, hogy mennyivel könnyebben férhetők hozzá az összetartozó információk. Mind a kategóriarendszer kialakítását, mind a kategóriákba sorolást önkéntes szerkesztők végzik, akik nem feltétlenül ismerik a kategóriarendszert, melynek akár részleges feltárása is időigényes lehet, így nem várható el a szerkesztőtől. A szerkesztők munkája támogatásának kézenfekvő módja egy, a szócikk tartalma alapján kategóriaajánlatokat tevő rendszer. Itt bemutatunk egy elosztott, gyors válaszidejű, széleskörűen integrálható szövegosztályozó rendszert. Rendszerünk alkalmazhatóságát a magyar nyelvű Wikipédiába integrálható „okos” szerkesztővel demonstráljuk.

A szövegosztályozás, dokumentumok kategóriákba sorolása a természetes nyelvek feldolgozásának talán legkiforrottabb területe [1]. A szövegosztályozást a legtöbb módszer az alábbi lépésekben valósítja meg:

0. Nyers szöveggé alakítás (dokumentum → szöveg)
1. Nyelvi feldolgozás (szöveg → szófolyam): szavakra bontás, szótövezés, zajszavak eltávolítása;
2. Indexelés (szófolyam → egész vektor): egyedi szavak előfordulásainak összeszámlálása, a korpuszban túl gyakori vagy túl ritka szavak eltávolítása;
3. Súlyozás (egész vektor → valós vektor): a szavak dokumentumra vonatkozó fontosságának meghatározása,
4. Predikció (valós vektor → súlyozott kategóriák): betanított/felépített osztályozómodell alkalmazása.

Munkánk újdonsága nem a terület előremozdításában áll, hanem a szokásos offline, csővezeték jellegű feldolgozástól eltérő, közel valós idejű működésben. Tehát a bemutatásra kerülő rendszer a fenti lépéseket nem egy egész dokumentumgyűjteményre, hanem az egyes dokumentumokra külön végzi el, a válaszidőt előtérbe helyezve az átlagos feldolgozási idővel szemben. A kategóriajavaslatok

mellett a rendszer evidenciát is szolgáltat a döntésre a dokumentum adott kategóriára releváns szavainak kiemelésével.

A legtöbb fent vázolt technológiai lépés elvégzésére számos szabad szoftver (pl. NLTK, Snowball, Weka) és üzleti programcsomag található (pl. SPSS Textmining). Az itt bemutatásra kerülő megvalósításban[1] a nyelvi előfeldolgozást és indexelést az Apache Lucene[2], a súlyozást az Apache Mahout[3], az osztályozást pedig a HITEC osztályozó[4] [2] végzi. Az osztályozó választásakor a döntő szempont a HITEC mellett az volt, hogy támogatja a hierarchikus kategóriarendszereket, mint amilyen a Wikipédiáé is. A szócikkek nyers szöveggé alakítását a Devijver-féle elemző[5] módosított változata végzi.

Az osztályozó szolgáltatás a könnyű integrálhatóság érdekében HTTP REST felületen keresztül érhető el, a kimeneti formátumok között szerepel olvasható HTML és gépi feldolgozásra szánt XML. Az osztályozó példányok számának növelésével érhető el a rendszer horizontális skálázása, amely várhatóan elengedhetetlen Wikipédia méretű alkalmazás esetén.

A rendszer válaszideje 10 kB méretű (hosszú) szócikkekre 150 ms körüli, mely közel egyenlő arányban oszlik meg az előfeldolgozás (1–3) és a predikció (4) között. Ez a válaszidő ergonómiai szempontból nyilvánvalóan megfelel egy hálózati szolgáltatással szemben támasztott követelményeknek.

Az automatikus és azonnali kategóriajavaslatok felkínálása csak egy módja a Wikipédia-szerkesztők támogatásának. Folyamatban van az itt vázolt rendszer kiegészítése, mely a hasonló szócikkek, azok kategóriái, valamint a kategóriák jellemző szócikkeinek felkínálásával segíti a szerkesztőket a kategóriarendszerben való eligazodásban és így a jobb minőségű kategóriarendszer kialakításában.

Hivatkozások

1. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 2002; 34(1):1–47.
2. Tikk D., Biró Gy., Törcsvári A.: A hierarchical online classifier for patent categorization. *Emerging Technologies of Text Mining: Techniques and Applications* 2007; 244–67.

[1] `http://categorizer.tmit.bme.hu/trac/wiki/HITEC-java`
[2] `http://lucene.apache.org/`
[3] `http://mahout.apache.org/`
[4] `http://categorizer.tmit.bme.hu/trac/`
[5] `http://code.google.com/p/java-wikipedia-parser/`

A HG-1 treebank: a nyelvtanírástól az online konkordanciáig

Tóth Ágoston[1]

[1] Debreceni Egyetem, Angol-Amerikai Intézet
4010 Debrecen, Egyetem tér 1.
tagoston@delfin.unideb.hu

Kivonat: Kutatócsoportunk a magyar nyelv LFG nyelvtanának kifejlesztését és egy treebank elkészítését tűzte ki célul. Az implementált nyelvtani jelenségek vonatkozásában a korpuszt mondattani annotációval látjuk el, majd egy kiválasztott alkorpuszt manuálisan egyértelműsítünk. Ehhez kapcsolódóan olyan eszközt is fejlesztünk, amellyel az alternatív elemzéseket kapó mondatok elemzését grafikus felületen szerkeszthetjük. A korpusz hozzáférhetőségének és kereshetőségének biztosítására online konkordanciaprogram-alkalmazást fogunk megvalósítani, így lehetővé válik a weben keresztüli keresés tetszőleges szóra, szófajra, alaktani jegyre, mindezt igény esetén meghatározott mondattani címkével ellátott összetevőre szűrve, a kimeneten megfelelően vizualizálva. Jelenleg a korpuszfejlesztési munka tervezési fázisának lezárásáról számolunk be, és áttekintjük a folyamatban lévő fejlesztéseket és az előttünk álló feladatokat.

1 Bevezetés

A Debreceni Egyetem Angol-Amerikai Intézetének Laczkó Tibor által vezetett Lexikai-Funkcionális Grammatikai Kutatócsoportja megkezdte egy 1,5 millió szavas, magyar írott nyelvi szövegeket tartalmazó treebank összeállítását és annotálását a saját készítésű LFG nyelvtanának felhasználásával. A készülő korpuszunkat, melynek a HunGram-1 (HG-1) nevet adtuk, mondattani annotációval látjuk el automatizált módon, majd egy kiválasztott alkorpuszt manuálisan ellenőrzünk és egyértelműsítünk. A kézi annotálás tapasztalatait a nyelvtanírásban felhasználjuk. A korpusz hozzáférhetőségének és kereshetőségének biztosítására online konkordanciaprogram-alkalmazást fejlesztünk.

2 Az elméleti keret és az implementációs környezet

A treebank projekt lényegi hátterét adó nyelvtanírási munka a **lexikai-funkcionális grammatikát** (LFG) használja, amely egy erős lexikalizmusra építő nem transzformációs generatív keret.

Kutatócsoportunk munkája közvetlenül kapcsolódik a nemzetközi **ParGram** együttműködéshez, melyben több nyelvhez (angol, német, norvég, francia, japán,

urdu, arab, török stb.) készül LFG nyelvtan, folyamatos egyeztetések mellett. A grammatika minden elemét – a többi ParGram projekthez hasonlóan – a Xerox Linguistic Environmentben (XLE) implementáljuk.

3 A treebank projekt feladatai

A tervezési fázisban meghatároztuk az alkalmazandó adatbázis-szerkezetet és a kifejlesztendő adatbázis-kezelő rendszer funkcióit. A mondattani fák tárolását a TigerXML [1] segítségével oldjuk meg, amely kiváló eszköz fák reprezentálására, és felhasználták – többek között – a Penn Discourse Treebank XML-re konvertálására is [2]. A projekt szoftver-infrastruktúráját – a parser kivételével – házon belül hozzuk létre.

Az induláshoz két kész korpuszt használunk nyersanyagként: a *Hunglish* korpuszt, mely egy nem annotált magyar-angol párhuzamos korpusz, és a *Szeged Treebank 2.0* korpuszt, ami egy 1,2 millió szavas magyar treebank (melynek annotációját projektünkben nem használjuk fel). Mindezt kiegészítjük egy saját gyűjtésű „nyers" korpusszal, ami főleg szépirodalmat, technikai dokumentációkat és híreket tartalmaz.

A programozási feladataink:

1) Mondatok elemeztetése a készülő nyelvtannal feltöltött XLE elemzővel, és a kimenet rögzítése (alternatív elemzésekkel). A korpuszt ettől a ponttól XML dokumentumban tároljuk.
2) Az összes lehetséges elemzés c-struktúrájának kibontása és tárolása.
3) Alkorpuszok kezelése:
   - korpuszfájlok darabolása és egyesítése,
   - indexelés, statisztikák készítése (faszélesség, -mélység, szavak és mondatok száma).
4) Kiválasztott mondat kézi egyértelműsítése, illetve annotációja saját fejlesztésű, grafikus felületű szerkesztőprogrammal, melynek a tervezett főbb funkciói a következők: ábrázolás, ágrajz kézi szerkesztése (melyhez bármelyik automatikusan generált elemzés kiindulópontként választható; a többszavas kifejezések lexikai egységként megjelölhetők, a morfológiai címkék megváltoztathatók; az ágrajzon élek és csomópontok létrehozhatók és törölhetők), a felhasználó által helyesnek vagy rossznak ítélt elemzések megfelelő feljelölése, megjegyzések elhelyezésének lehetősége.
5) Online lekérdezési felület a következő főbb funkciókkal:
   - keresés szóra vagy lemmára reguláris kifejezések használatával,
   - keresés szűrése morfológiai jegyekre és a keresett szót tartalmazó összetevőre (szűrés beállítása űrlap segítségével),
   - a találatok KWIC konkordanciaként való megjelenítése,
   - a konkordanciából kiválasztott mondat ágrajzainak megjelenítése.

A korpuszt tartalmazó XML dokumentumot több lépésben hozzuk létre, a fent említett eszközök segítségével. A nyelvtanunkkal feltöltött XLE parser PROLOG kódot ad vissza elemzésként, mely tartalmazza a „csomagolt" (a többértelműségeket részfákra lokalizáltan tároló) LFG c-struktúrát és f-struktúrát. Az XML fájl első változata

az eredeti mondaton (és az ahhoz kapcsolódó alapadatokon, valamint az esetlegesen meglévő angol fordításon kívül) ezt, az XLE-ből közvetlenül kapott elemzést tárolja. A következő lépésben a PROLOG kódból automatikusan létrehozzuk az összes lehetséges elemzést, majd ezzel egységes szerkezetben tároljuk a kézi annotáció eredményét.

A mondattani fák reprezentálását egy Tiger-XML alapú leírónyelv segítségével oldjuk meg. Egy ágrajz kódolása a gyökérelem kijelölésével indul, utána a terminális szimbólumok felsorolása következik, melynek során a lexikai egységekhez kapcsolódóan a szófajt, a lemmatizált alakot és a morfológia által visszaadott összes jegyet tároljuk. Ezt követi az összes többi csomópont leírása legalább 1-1 kapcsolódó él meghatározásával.

4 Felhasználási lehetőségek

A készülő korpusz (a tervezett lekérdezési lehetőséggel) felhasználható a nyelvoktatás, nyelvtanulás területén, a konkordanciaalapú megoldások összes előnyével: autentikus élőnyelvi szövegekkel dolgozhatunk olyan módon, hogy a tanulás nyelvi felfedezéssé válik. Ugyancsak fontosak számunkra a lehetséges lexikográfiai alkalmazások, valamint a korpusz felhasználása elméleti nyelvészeti kutatásokban: ez utóbbira példa a kutatócsoportunk nyelvtanírási projektje is, amelyhez a korpuszfejlesztési alprogram folyamatos tesztelési lehetőséget és visszajelzést biztosít.

Köszönetnyilvánítás

A munkát részben az OTKA (K 72983), részben a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatja. A projekt az Új Magyarország Fejlesztési Terven keresztül az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósul meg.

Bibliográfia

1. Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G.: The TIGER Treebank. In: Proceedings of the Workshop on Treebanks and Linguistic Theories (2002) 24–41
2. Yao, X., Borisova, I., Alam, M.: PDTB XML: the XMLization of the Penn Discourse TreeBank 2.0. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) (2010) 2022–2027

Szerzői index, névmutató