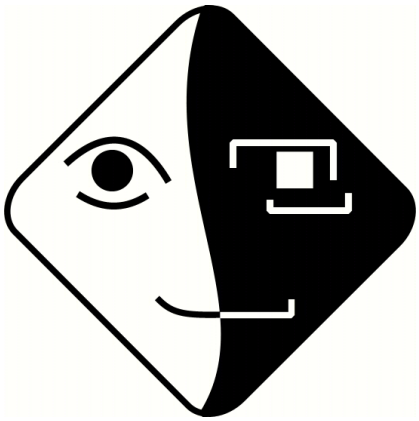
VI. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2009

Szeged, 2009. december 3-4.
http://www.inf.u-szeged.hu/mszny2009

Előszó

2009. december 3-4-én immár hatodik alkalommal lesz Magyar Számítógépes Nyelvészeti Konferencia. Nagy örömmel látom, hogy a rendezvény – kétévnyi szünet után – fokozott érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferencia fő célja az eddigi évekhez hasonlóan 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 és javaslatok közül a programbizottság 45-öt fogadott el, így 31 előadás és 14 poszter-, illetve laptopos bemutató alkotja a konferencia programját.

Külön kiemelendő, hogy a konferencián önálló szekciót szentelünk a Nyelv- és Beszédtechnológiai Platform által létrehozott Stratégiai Kutatási Terv részletes bemutatásának. A magyarországi nyelv- és beszédtechnológiai műhelyek közösen készítették el az ágazat magyarországi helyzetét tükröző Jelenképet, a lehetséges továbbfejlődési irányokat bemutató Jövőképet, illetve a Stratégiai Kutatási Tervet. Ez utóbbi a létrehozók céljai szerint az ágazati szereplők számára jövőbetekintő, kutatási sarokpontokat és módszereket meghatározó iránymutatásként szolgál.

Örömömre szolgál az is, hogy az idei konferenciára Kálmán László nyelvészkutató is elfogadta meghívásunkat, így az ő plenáris előadása is gazdagítja a szakmai programot. 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.

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, Szauter Dóra, Vincze Veronika) munkáját is.

Csirik János, a rendezőbizottság elnöke

Szeged, 2009. november

Tartalomjegyzék

I. Fordítás

II. Szövegbányászat

III. Korpusz, ontológia, lexikográfia

IV. Beszédtechnológia

V. Pszichológiai vonatkozású fejlesztések

VI. Gépi tanulás

VII. Poszter- és laptopos bemutatók

I. Fordítás

Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban

Recski Gábor, Varga Dániel, Zséder Attila, Kornai András

BME Média Oktató és Kutató Központ, e-mail:
{recski,daniel,zseder,kornai}@mokk.bme.hu

1. Bevezetés

Cikkünkben egy magyar-angol szövegfeldolgozó rendszert mutatunk be. Elsőként a maximális főnévi csoportok magyar, illetve angol nyelvre történő azonosítását végző `hunchunk` komponenst írjuk le. A 3. részben egy szótárépítő módszert ismertetünk, a 4. részben pedig leírjuk korpuszfeldolgozó rendszerünk néhány technikai részletét, melyek lehetővé teszik, hogy nagy mennyiségű nyers kétnyelvű szöveg birtokában hatékonyan – akár több szerveren párhuzamosan – végezzük el elemzett bikorpusz építését, és az adatok webes mondattárunkba integrálását. További terveinkről az 5. részben számolunk be.

A cikkben bemutatott kétnyelvű kísérletekhez a Hunglish Korpusz [1] mondatszinten párhuzamosított, magyar-angol nyelvű bikorpuszt használtuk. A korpuszban szépirodalom, jogszabályok szövegei, hírlapok és magazinok cikkei, filmszövegek, szoftverdokumentációk, valamint pénzügyi jelentések találhatók. A cikk további részében bemutatott elemzési eljárások elvégzését követően a Hunglish Korpuszról az 1. táblázatban látható statisztikát készíthetjük.

1. táblázat. A Hunglish korpusz számai

nyelv	token	típus	tő-típus	mondat	NP
magyar	31.4M	941k	342k	2.07M	7.6M
angol	37.1M	311k	248k	2.07M	5.2M

Magyar szövegre a morfológiai egyértelműsítést és tövezést a `hundisamb` eszköz végezi. Ez a `hunmorph` morfológiai elemző által felajánlott elemzések közül választ, a `hunpos` HMM-alapú morfológiai címkéző algoritmust alkalmazva. A hunmorph-ot ehhez tőkitaláló üzemmódban használja, amely heurisztikus elemzési javaslatokkal él, ha az elemzés a szótárában megtalálható szavakra nem vezethető vissza. A hunpos címkéző működéséhez szükséges modelleket magyar nyelvre a Szeged Treebank [2], angol nyelvre a Penn Treebank [3] segítségével tanítottuk.

Angol nyelvre a hundisamb eszközt nem alkalmazhattuk, mert a hunpos angol modellje Penn Treebank címkéket bocsát ki, amelyek közvetlenül nem feleltethetők meg a hunmorph angol morfológiai címkéinek. (Terveink között szerepel ennek a kellemetlen inkonzisztenciának az orvoslása.) Itt az angol tövezőnk

által javasolt alternatívák közül mindig a legrövidebbet választottuk. Szószintű párhuzamosításhoz és fordításhoz a legrövidebb lehetséges szótő és a Penn Treebank címke együttese jól használható mint a token normálformája, bár időnként nem teljesen helyes, pl. a *grind* és *ground* főnevek normálalakja e heurisztika szerint egybeesik.

2. Főnévi csoportok azonosítása

A morfológiai információk birtokában elvégezhetjük a szónál magasabb szintű egységek azonosítását. A főnévi csoportok azonosításához (NP-chunking) az itt bemutatásra kerülő `hunchunk` eszközünket [4] használjuk, mely a szegmentálási feladatot szószintű címkézési feladattá alakítva végzi. Elsősorban a szavak elemzésére támaszkodó jegyek segítségével maximum entrópia modellt tanít, majd címkézéskor a tanítókorpuszban megfigyelt átmenetvalószínűségeket figyelembe véve mondatonként azonosítja a legvalószínűbb címkesorokat.

2.1. A tanulóadatok előállítása

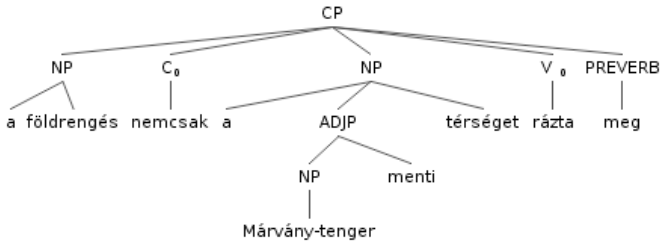
A magyar nyelvű tanulóadatokat a Szeged Treebankből nyerjük ki: a korpuszban található maximális NP-ket feleltetjük meg chunkoknak, tehát azokat a főnévi csoportokat, melyeket más NP nem dominál. Bár az NP-chunkok azonosítása a szakirodalomban leggyakrabban valamennyi minimális NP megkeresését jelenti, célszerűbbnek láttuk a fenti definíciót alkalmazni, mivel így lehetőségünk nyílik a mondatok közvetlen összetevőinek elkülönítésére és az igék argumentumszerkezetének feltérképezésére.

A tokenek címkézésekor a Start/End [5] konvenciót alkalmazzuk, mely az elterjedtebb IO és IOB konvencióknál [6] több címkét igényel, ugyanakkor lehetővé teszi többféle chunkbeli pozíció megkülönböztetését: míg az előbbi megoldások vagy egy címkével (I-NP) jelölik a chunkhoz tartozó szavakat, vagy ezen felül még a chunkot kezdő szót jelölik külön szimbólummal (B-NP), addig az általunk használt jelölés a chunkhoz nem tartozó szavakon (O) kívül négy címkét használ (B-NP, I-NP, E-NP, 1-NP), melyek rendre a chunk elején, közepén és végén álló, valamint az önmagában chunkot alkotó szavakat jelölik.

Az adatok kinyerésekor feljegyezzük azt is, hogy az adott NP-be milyen mélyen ágyazódnak további NP-k, így lehetőségünk nyílik egyfajta komplexitásfogalom alapján több chunktípust megkülönböztetni. Az effajta információk kinyerését nem tekintjük a címkéző feladatának, csupán a gépi tanulási feladatot könnyítjük meg vele: optimálisnak az a címkézés bizonyult, ahol csupán a legalacsonyabb – tehát további NP-t nem tartalmazó – chunkokat különböztettük meg (N_1) a komplexebbektől (N_2+). A fenti chunkdefiníció és címkézés eredményeképp a Szeged Treebank az 1. ábrán látható mondata a chunkkorpuszban a 2. ábra szerinti reprezentációt kapja.

A	földrengés	nemcsak	a	Márvány-tenger	menti	térséget	rázta	meg
B-N_1	E-N_1	O	B-N_2+	I-N_2+	I-N_2+	E-N_2+	O	O

1. ábra.



2. ábra.

Az angol nyelvű tanulóadatok kinyeréséhez a Penn Treebanket használjuk. Itt NP-chunknak tekintjük a maximális főnévi csoportok mellett azon prepozíciós frázisokat is, melyek tartalmaznak főnevet, nem tartalmaznak igét és nem képezik részét magasabb szintű NP-nek. Ezzel [7] definícióját követjük, melyet a szerző azzal motivál, hogy az NP és PP szerkezetek közti határt a különféle nyelvek nem ugyanott húzzák meg, illetve a két kategória számos nyelvben nem is különül el egymástól élesen. Fontosnak tartjuk megemlíteni, hogy az NP-chunk definíció mindkét nyelv esetében csupán a korpuszt előállító rendszer beállításaitól függ, így amennyiben eltérő egységeket tekintünk chunknak – így például a fent említett módon a minimális NP-ket szeretnénk azonosítani – úgy ahhoz egyszerűen állítható elő megfelelő tanítókorpusz.

2.2. A jegyek

A tanítás elsősorban szószintű jegyek alapján történik: egy szó jegyének tekintjük a szótövet és valamennyi morfológiai jegyet. A Szeged Treebank MSD-konvenció szerinti annotációját átalakítottuk a KR-formalizmusra, mivel az általunk használt morfológiai címkéző, elemző és egyértelműsítő egyaránt ezt a formátumot követi. Jegyként vettük fel az így előállított KR-kódok valamennyi elemi összetevőjét. A Penn Treebank esetében ezt nem tehettük meg, mivel az abban használatos morfológiai címkék nem kompozicionálisak, így ott a teljes címke mellett csupán annak első karakterét – mely a szófajt azonosítja – vesszük fel önálló jegyként. A szószintű jegyeket minden tokenre annak 5 szavas környezetében értékeljük ki.

Bevezettünk továbbá egy jegyet, mely egy szó adott hosszúságú környezetében az egymást követő szavak szófaji címkéinek sorozatait írja le a követ-

kező módon: ha a jegy sugarát r-rel, egy mondat i-edik pozíciójában álló szót w_i-vel, szófaji címkéjét pedig p_i-vel jelöljük, úgy, úgy bármely w_i szóra jegyként vesszük fel a $p_{i-r} \dots p_{i+r}$ sorozat összes összefüggő részintervallumát. A KR-mintákat kiválasztó jegy sugarát növelve a chunkolás F-pontszáma is nő, 3-nál magasabb sugár mellett azonban a jegyek magas száma nem teszi lehetővé a modell tanítását.

2.3. A statisztikus modell

A címkézési feladat modellezéséhez rejtett Markov Modellt (HMM, [8]) tanítottunk, melynek kibocsátási modelljét Maximum Entrópia modellből [9] nyertük. Az alábbiakban ismertetjük modellünket, és a mögötte álló statisztikai előfeltevéseket.

Jelölje $p(i, u)$ annak valószínűségét, hogy az i pozícióban álló szó az u címkét kapja. Feltételezzük, hogy $p(i, u)$ értéke kizárólag annak $w_{i-k} \dots w_{i+k}$ környezetétől függ. Ekkor $p(i, u)$ értékét $\hat{p}(i, u)$ kiszámításával becsüljük, melyet a korábban ismertetett jegyeken tanított ME modell szolgáltat. Jelölje $t(i, u, v)$ annak feltételes valószínűségét, hogy az i pozícióban álló szó u címkét kap, feltéve hogy az $i - 1$ pozícióban álló szó a v címkét kapta. Feltételezzük, hogy ez a valószínűség független i-től és a tanítókorpuszban megfigyelt feltételes relatív gyakorisággal ($\hat{t}(u, v)$) adunk rá becslést.

A címkézés során a rendszer egy adott mondatra adható legvalószínűbb címkesorozatot keresi. Ha $\hat{p}(i, u)$ csupán w_i-től függne (tehát nem számítana a környezet), akkor egy sorozat valószínűsége a feltételes függetlenségnek köszönhetően szorzatként állna elő és az alábbi képlettel lenne arányos:

$$\prod_i \frac{\hat{p}(i, u_i)\hat{t}(i, u_i, u_{i-1})}{P(u_i)}.$$

Ezen képlet maximuma, tehát a legjobb címkesorozat megtalálható a Viterbi algoritmus segítségével. Ez a modell valójában a ‘megfigyelések az állapotokban és nem az átmenetekben’ változata a Maximum Entrópia Markov Modellnek, ahogy [10] javasolja. Modellünket úgy írhatjuk le, mint ennek a modellnek az egyszerű általánosítását: megengedjük, hogy $\hat{p}(i, u)$ egy $w_{i-k} \dots w_{i+k}$ ($k > 0$) környezettől függjön és a fenti képlet segítségével becsüljük a tényleges valószínűséget. A tekintetbe vett környezet k sugara optimalizálandó paraméter, a cikkünkben említett összes feladaton az 5 sugár bizonyult optimális választásnak.

Rendszerünknek még egy szabad paramétere a nyelvi modell súlyozása. Ez standard megoldás a HMM szakirodalomban. Esetünkben a fenti képletet ez úgy általánosítja, hogy egy pozitív λ kitevőt alkalmazunk a $\hat{p}(i, u_i)$-re és a $P(u_i)$-re. A λ paramétert kisméretű részkorpuszon optimalizáltuk egy-egy feladathoz.

2.4. A magyar és angol NP-chunking kiértékelése

A főnévi csoport azonosító kiértékeléséhez NP-korpuszainkat mindkét nyelven egy 1000000 token hosszúságú tanító- és egy 500000 token hosszúságú teszt-

korpuszra osztottuk véletlenszerűen. A tesztkorpuszon lefolytatott címkézések kimenetét a [11]-beli szabályokat követve értékeltük ki. Összehasonlítási alapmódszerként (baseline) magyar nyelvre minden szónak a szófaji címkéje alapján legvalószínűbb címkét osztottuk ki. A legegyszerűbb címkézési módszert követve, amely csupán az I-NP és O címkéket használja, a rendszer csupán 51.03%-os F-pontszámot ért el. Egy kevés bonyolítással - harmadikként bevezetve a B-NP címkét – az eredmény 60.37%-ra nőtt. Rendszerünk eredményei magyarra a 2. táblázatban láthatóak.

2. táblázat.

	Pontosság	Fedés	F-pontszám
baseline	60.24%	60.50%	60.37%
hunchunk	89.40%	89.97%	**89.68%**

Felhívjuk a figyelmet arra, hogy az NP chunk általunk adott, a szakirodalomban legelterjedtebbtől eltérő definíciója jelentősen hosszabb és szerkezetüket tekintve komplexebb NP-ket eredményezett, mint pl. a [11] szerinti, ún. alap NP („base NP"). Ez magyarázza a szakirodalomban szokásosan láthatónál alacsonyabb pontszámokat. Noha célunk a maximális NP-k azonosítása volt, algoritmusunkat egy minimális NP-feladaton is kipróbáltuk, hogy teljesítményét összevethessük a legkorszerűbbnek tartott statisztikus szegmentálóalgoritmusokéval. A CoNLL 2000 Shared Taskon, melynek tanuló- és tesztadata rögzített, és a szegmentálóalgoritmusok összehasonlításának standard terepeként szolgál, eszközünk 93.79%-os F-pontszámot ért el. Ez körülbelül fél százalékkal alacsonyabb, mint a modelltanításkor egy nagyságrenddel nagyobb számításigényű CRF algoritmusok eredménye: [12] 94.34%, [13] pedig 94.29% F-pontszámot publikált a feladaton. A `hunchunk` kétfajta feladaton elért eredményeit a 3. táblázat tartalmazza.

3. táblázat.

Feladat	Pontosság	Fedés	F-pontszám
max NP	79.33%	79.87%	**79.60%**
base NP	93.61%	93.85%	**93.73%**

3. Szótárépítés

Az alábbiakban egy egyszerű iteratív szótárépítő algoritmust mutatunk be, amely az együttes előfordulások aránya alapján rangsorolja a szótári tételeket. A miénkhez hasonló, úgynevezett Competitive Linkingen alapuló algoritmust

elsőként Melamed [14] publikált. Ezután bemutatjuk, hogy a Competitive Linking eljárás pontossága növelhető, ha kiaknázunk egy automatikus szószintű párhuzamosítást.

3.1. Az algoritmus

Szótárépítési eljárásunk alapja a Dice együttható néven ismert mérőszám egy magyar-angol szópár együtt-előfordulásának mértékére: ennek definícója $D = o_{he}/(o_h + o_e)$, ahol o_{he} a szópár együttes előfordulásainak száma, o_h és o_e az egynyelvi előfordulások száma. Ha egy bimondatban több előfordulás is van, akkor a két előfordulásszám mondatbeli minimumával (és nem szorzatával) járul hozzá a mondat az o_{he} mennyiséghez.

Az algoritmus első lépésként összegyűjti az összes olyan magyar-angol szópárt, amelyre D egy t küszöb felett van, ahol t az algoritmus paramétere. Ha egy szó egynél több így azonosított szótári tételben is szerepel, akkor csak a legnagyobb hasonlósági mértékűt tartjuk meg. Az iteráció kimenete az így összegyűjtött tételek halmaza. Ezután a korpusz bimondataiban összekötjük azokat a magyar-angol szópárokat, melyek megtalált szótári tételnek felel meg, és töröljük a korpuszból az összes összekötött szót. Ezen a ponton újrakezdhetjük az iterációt a Dice együtthatók kiszámításával, és joggal remélhetjük, hogy a korábbi tételek eliminálása után egyes új tételek hasonlósága a küszöb fölé lép. Az iterációt addig folytatjuk, amíg a szótár már nem bővül tovább – kísérleteinkben ehhez 10-15 iterációra volt szükség, egyre csökkenő hosszúságú iterációk mellett.

A most ismertetett eljárást ItCo-nak fogjuk nevezni az alábbiakban. Az eljárásnak megvizsgáljuk majd azt az ItCo+GIZA változatát is, amely (az alapváltozattal ellentétben) feltételezi egy szószintű párhuzamosítás meglétét. Ez a változat csak azokat az együttes előfordulásokat veszi számításba, amelyeknél a két szó között kapcsolat van a párhuzamosításban.

A szószintű párhuzamosítás építéséhez a szakirodalomban teljesen standardnak tekinthető GIZA++ és Moses eszközöket választottuk. Az IBM tanulóalgoritmusának [15] GIZA++ [16] által adott implementációja egy ún. IBM Model 5 fordítási modellt épít a tokenizált párhuzamos korpuszból, amiből egy asszimetrikus szószintű párhuzamosítást nyerhetünk ki. Ezt lépést a magyar-angol és angol-magyar fordítási irányokra egyaránt elvégezve két „félkész" párhuzamosítást kapunk. Ezeket a Moses [17] frázisalapú gépi fordítóhoz mellékelt heurisztikus algoritmus fésüli össze minél konzisztensebb szimmetrikus szószintű párhuzamosítássá.

3.2. Az eljárás kiértékelése

Méréseinkhez a Hunglish Korpusz tövezett, szószinten párhuzamosított változatát alkalmaztuk. A szavak halmazán szótárépítés előtt háromféle szűrést is végeztünk: elhagytuk a funkciószavakat, azokat a szavakat, amelyek nem szerepeltek legalább 10-szer a korpuszban, továbbá azokat a szavakat (szótöveket) is, amelyek nem szerepeltek magyar, illetve angol tövezett gyakorisági

szótárainkban. (Előbbinek a forrása a Szószablya Webkorpusz, utóbbié a Google 1T webkorpusz [18], mindkettőt a `hunmorph` eszközzel [19] töveztük.) A szótárépítés elvégzése után pedig elhagytuk a nagy kezdőbetűs tételeket is, hiszen ezek nagy pontossággal megfeleltek a tulajdonneveknek.

Az automatikusan létrehozott szótárakról automatikus eszközökkel igazán pontos minőségi információkat kinyerni nem lehet. De hogy már az előzetes kísérletek során fogalmat nyerhessünk a szótárak relatív minőségéről különböző paraméterbeállítások mellett, először mégis a Vonyó Attila szótárával való százalékos átfedésüket vizsgáltuk. Ezen mérések alapján úgy választottuk meg a szótárépítő algoritmus Dice paraméterét (0.095-nek), hogy egyensúlyt valósítsunk meg a szótár pontossága és mérete között. Az optimálisnak talált paraméterbeállítás mellett 21846 méretű szótárunk tételeinek 53.9%-a szerepelt a Vonyó-szótárban, amely arány 71.5%-ra nőtt, ha a szavaknak csak az 5 hosszú kezdőszeleteit illesztettük. Hangsúlyozzuk, hogy ez globális pontossági mértékként félrevezető, amennyiben a Vonyó-szótárban nem szereplő tételek többsége is legitim találat.

A paraméterhangolás után elvégeztük a manuális kiértékeléseinket. A hibafajták között nem meglepő módon a domináns az volt, amikor a szópár képzőtől eltekintve helyes volt. Ez előállhat akkor, amikor a magyar és angol szöveg különböző szófajú konstrukcióval fejez ki egy adott fogalmat, illetve ha a két tövező másként dönt egy képzett szó lexikalizált mivoltáról, pl. *vallásos-religion, szerencse-lucky, forgatás-rotate, szökött-escape, továbbfejleszt-development.* A manuális kiértékeléskor ezt a hibaosztályt külön jelöltük. Kétféle pontosságmértéket alkalmaztunk egy szótár minőségének manuális számszerűsítésére: a teljesen helyes tételek arányát, illetve a képzési hibától esetlegesen eltekintve helyes tételek arányát, amit nemhelytelen-nek neveztünk.

3.3. Eredmények

A kiértékelés alapjának [20]-gyel azonos módon a GIZA++ IBM Model 5 szótárépítőjét választottuk, amely a Model 5 fordítási modellből nyeri ki a szótárat. A rendszer minden szótári tételhez egy 0 és 1 közötti konfidenciaértéket ad. Ezek csökkenő sorrendje szerint rendezzük a szótári tételeket, így tetszőleges arányt megcélozhatunk méret és pontosság között. Egy általunk épített szótárral való pontossági összehasonlításkor a GIZA++ szótár akkora méretű kezdőszeletét választottuk, amekkora az összehasonlítandó szótár. A baseline szótár építésekor ugyanazokat a szűrőket alkalmaztuk, mint saját szótáraink esetében. Az eredmények szótáranként 200 véletlen minta vétele alapján kiértékelve a 4. táblázatban láthatók.

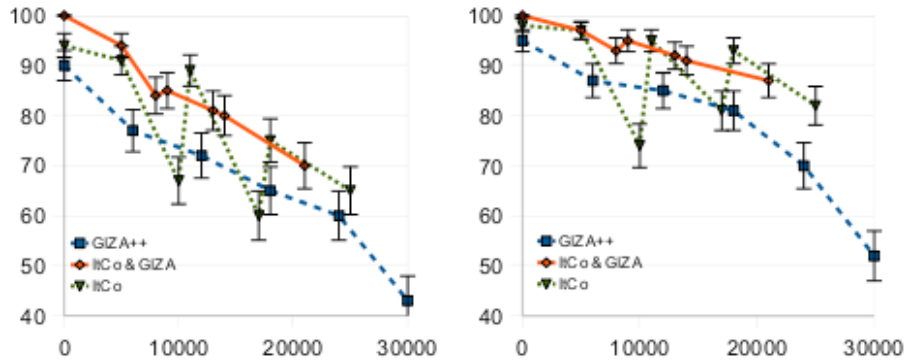
Az épülő szótár mérete az algoritmus paramétereitől függ. Az algoritmus futása jól elkülönülő iterációkból áll (10-15 egyre csökkenő méretű iteráció). Egy iteráción belül fokozatosan csökken a konfidencia és a tényleges pontosság is. Két iteráció között azonban felugrik a konfidencia és pontosság, hiszen a helyesen azonosított szótári tételeknek megfelelő szópárok elhagyása csökkenti a félrevezető kollokációk halmazát. Ez egy fűrésszerű pontossági grafikonhoz ve-

4. táblázat.

	Baseline	ItCo	Baseline	ItCo+GIZA
helyes %	68.5	77.0	69.2	81.5
nemhelytelen %	76.5	87.5	76.7	92.5
szótáméret	25422	25422	21846	21846

zet, ha az x-tengelyen azt ábrázoljuk, hogy a tétel hanyadikként lett azonosítva, az y-tengelyen pedig a simított pontosságot.

Ahhoz, hogy bemutathassuk ezeket a pontossági grafikonokat, mintavételezésre volt szükség, hiszen 25000 szótári tétel manuális ellenőrzése túlságosan időigényes feladat. A grafikon egy adatpontját ezért a következő módon hoztuk létre. Az x pozíció mintavételezéséhez véletlenszerűen kiválasztottunk a szótár (x,x+1000) intervallumából 100 tételt. Ezeket manuálisan klasszifikáltuk a már említett (helyes, képzéstől eltekintve helyes, helytelen) kategóriákba. Ez az adott x szótárpozícióhoz két különböző százalékos pontossági értéket rendelt: az egyik a helyes tételek aránya, a másik a nemhelytelen tételeké.



3. ábra.

A 3. ábrán látható, hogy a mintavételezés a GIZA++ által épített szótárak esetében szabályos lépésenként történt, a saját szótáraink esetében viszont nem. Ennek oka a grafikonok már említett fűrész-alakja. A lépésközt úgy választottuk, hogy az első két, 1000-nél még nagyobb méretű fűrészfog (azaz iteráció) belsejében két mintavételezési pont legyen: az iteráció elejéről illetve végéről. Az első, domináns méretű iterációnak a közepéről is mintavételezünk. Az ábrákról leolvasható, hogy lineárisan interpolálva a mintavételezési pontokat, a GIZA+ItCo módszer pontossága a GIZA módszeré felett van minden pontban, a ItCo módszeré pedig a legtöbb pontban.

4. Implementáció

Ebben a szakaszban szövegfeldolgozó rendszerünk néhány műszaki részletéről számolunk be.

4.1. Keretrendszer

Elsősorban az a keretrendszer érdemel említést, amelyet az adatok feldolgozására kiépítettünk. Ennek feladata az egyes feldolgozó modulok (pl. tokenizálás, szófaji elemzés) hatékony futtatása nagy méretű adathalmazokon. A rendszer nagyon rugalmas keretet ad az általa futtatott moduloknak, nem kötelezi el magát például abban sem, hogy milyen programozási nyelven kell implementálnunk azokat.

A keretrendszer használatához az elvégzendő feladatok irányított aciklikus gráfját kell definiálnunk, megadva, hogy a csúcsokhoz tartozó feladatok milyen parancsnak felelnek meg. A keretrendszer feldolgozandó fájlok egy halmazára alkalmazza ezt a pipeline-t vagy valamely kijelölt részgráfját, egy standardizált struktúrájú könyvtárhierarchiát hozva létre. Két specializált szolgáltatást nyújt a rendszer, amelyek gyorsítják a feladat elvégzését, ezek akár egyszerre is kiaknázhatóak:

- Párhuzamosítás: A rendszer képes felhasználni a feladatok párhuzamos elvégzéséhez egy számítógép-klasztert, a klaszterben részt vevő számítógépek egyes processzorait párhuzamosan terhelve. Ehhez csupán arra van szükség, hogy a klaszter egyes tagjai hozzáférjenek az adatokat és modulokat tartalmazó fájlrendszerhez. Az ütemezés alapegysége a dokumentum, tehát egyetlen nagyméretű dokumentumot már nem tördel kisebbekre az ütemező.
- Daemon: A hunchunkhoz és hunnerhez hasonló gépi tanuló rendszerek statisztikus modelleket tartalmazó, sok megabájtos erőforrásfájlokat olvasnak be induláskor. Ezért ha sok kis dokumentumra futtatnánk ezeket, akkor a futásidő nagy részét inicializálással töltenék. A keretrendszer daemon üzemmódja ezt a problémát úgy orvosolja, hogy a munka kezdetén egyetlen alkalommal indítja csak el a címkéző/szegmentáló programot, majd az ütemezett fájlokat unix socketokon keresztül kommunikálva egymás után küldi el annak. A „becsomagoláskor” a keretrendszer a daemonként elindított programról nagyon kevés előfeltevéssel él. Ez a megoldás alkalmazandó akkor is, ha a címkéző/szegmentálókat például webszolgáltatás részeként kívánjuk alkalmazni.

A Hunglish Korpusz építését újraimplementáltuk a keretrendszerben, tehát az elemzési lépések kiindulópontja lehet nyers, formázatlan szöveg két nyelven. A megfelelő elemzési lépések elvégzése után a Hunglish Mondattár webes keresőrendszer indexelőjéhez vagy a Moses fordítórendszer modellépítőjéhez vagy dekóderéhez továbbíthatóak a feldolgozott adatok.

4.2. Huntag

A hunchunk a korábban publikált hunner rendszerhez [21] algoritmikusan nagyon hasonló – egyetlen különbségük, hogy a hunchunk a szegmentumok közti átmenet-valószínűségeket tanulja. A hunner rendszert ezért újraimplementáltuk, és a két szegmentálót egyetlen közös, huntag-nek nevezett eszközben valósítottuk meg, amelyet csak a jegy-számításért és paraméterezésért felelős leírófájlok adaptálnak egyik vagy másik feladathoz. A reimplementáció nem volt komoly hatással a hunner pontosságára, 96.35%/95.05%-ről 96.53%/94.81%-re változott a Szeged NER fejlesztő, illetve tesztelő adathalmazain.

5. További terveink

Elsődleges további tervünk olyan eljárás publikálása, és teljesítményének számszerűsítése, amely az azonosított maximális NP-ket párhuzamosítja a kétnyelvű szöveg bimondataiban. (Ilyen jellegű rendszert először Pohl [22] publikált magyar nyelvre, magyar-angol fordítómemória építése céljából.) Bár számszerűsíthető adataink a kézirat leadásakor még nincsenek, azt gondoljuk, hogy a maximális NP-k közt jóval nagyobb arányú az 1-1 párhuzamosság mint a szavak vagy alap NP-k közt, és hogy az NP-párhuzamosítási feladat hatékony megoldása nemcsak a gépi fordítást, hanem a mondatok argumentumszerkezetének megértését is segíteni fogja.

A keretrendszer és a huntag rendszer más technológiáinkhoz hasonlóan szabad forráskódúak. A cikk írásának időpontjában a :pserver:anonymous:anonymous@cvs.mokk.bme.hu:/local/cvs cvs-szerver tcg, illetve huntaggers moduljaiként már bárki számára elérhetőek, de célunk, hogy a rendszert a konferencia idejére megfelelő minőségű dokumentációval is ellássuk.

Hivatkozások

1. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Proceedings of the Recent Advances in Natural Language Processing 2005 Conference, Borovets. Bulgaria (2005) 590–596
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue. (2005) 123–131
3. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of english: The Penn Treebank. Computational Linguistics **19** (1994) 313–330
4. Recski, G., Varga, D.: Magyar főnévi csoportok azonosítása. Általános Nyelvészeti Tanulmányok (2010)
5. Uchimoto, K., Ma, Q., Murata, M., Ozaku, H., Isahara, H.: Named entity extraction based on a maximum entropy model and transformation rules. In: ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2000) 326–335
6. Sang, E.F.T.K., Veenstra, J.: Representing text chunks. In: EACL. (1999) 173–179

7. Koehn, P., Knight, K.: Feature-rich statistical translation of noun phrases. In: In Proc. of the 41st Annual Meeting of the ACL. (2003) 311–318
8. Rabiner, R.L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. IEEE. Volume 77. (1989) 257–286
9. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution. Technical report (1998)
10. Mccallum, A., Freitag, D., Pereira, F.: Maximum entropy markov models for information extraction and segmentation. In: Proc. 17th International Conf. on Machine Learning. (2000) 591–598
11. Sang, E.F.T.K., Buchholz, S., Sang, K.: Introduction to the CoNLL-2000 shared task: Chunking (2000)
12. Sun, X., Morency, L.P., Okanohara, D., Tsujii, J.: Modeling latent-dynamic in shallow parsing: a latent conditional model with improved inference. In: COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2008) 841–848
13. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Morristown, NJ, USA, Association for Computational Linguistics (2003) 134–141
14. Melamed, I.: (Empirical methods for exploiting parallel texts)
15. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics **19** (1993) 263–311
16. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
17. 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'07, The Association for Computer Linguistics (2007)
18. Brants, T., Franz, A.: Web 1t 5-gram corpus version 1. Technical report, Google Research (2006)
19. Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. (2005)
20. Karlgren, J., Sahlgren, M.: Automatic bilingual lexicon acquisition using random indexing of parallel corpora. Natural Language Engineering **11** (2005) 327–341
21. Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica **16** (2006) 293–301
22. Pohl, G.: English-hungarian np alignment in metamorpho tm. In: Proceedings of the EAMT. (2006)

Fordítások statisztikai alapú minőségvizsgálata tartalomelemzéssel

Puskás László

PTE BTK, Pszichológia Doktori Iskola
laszlopuskas@gmail.com

Kivonat: A tanulmány egy olyan eljárást mutat be, amely egy olasz szöveg magyar fordításának statisztikai jellemzőit vizsgálva igyekszik olyan általános statisztikai összefüggések bemutatására, amellyel meghatározott típusú szövegekben, a statisztikai jellemzők alapján, kiszűrhetők a hibás fordítást tartalmazó szövegrészek, illetve bizonyos típusú hibák. Az eljárás során az olasz szöveget meghatározott méretű szövegrészekre bontjuk, amelyeket azok magyar fordításával vetünk össze, a szövegrészek statisztikája alapján. Azt feltételezzük, hogy az eljárás az olaszon kívül más nyelvre is alkalmazható, azzal a megkötéssel, hogy a vizsgált statisztikai paraméterek eltérésének általános szabályait az adott nyelvre is ki kell dolgozni.

1 Bevezetés

A tanulmány egy az önéletrajzi emlékezet körébe tartozó olasz művet, és annak magyar fordítását megvizsgálva igyekszik meghatározni, hogyan szűrhetők ki bizonyos fordítási hibákat tartalmazó szövegrészek, statisztikai alapon. A cikket egyben vitaindítónak is szánom. Tanulmányomban a következő feltevések igazolására törekszem:

I. meghatározott típusú szövegek esetén, az olasz szövegrészekben szereplő szavak száma szinte mindig nagyobb a magyar szövegrészekben szereplő szavak számánál;

II. az olasz és a magyar szövegrészben szereplő szavak számának eltérése arányaiban általában jól behatárolható, de a mondat szintjén nem, csak a szövegrész szintjén alkalmazható;

III. az együttjárások elsősorban a történetek elbeszéléséhez, vagyis a narratív szemléletmódhoz köthetők;

IV. a szavak számának eltérése a szövegben szereplő szófajok arányainak eltérésével is együttjár;

V. az eljárás általános alkalmazása lehetővé teszi, hogy olasz és más idegen nyelvű szövegek hibás fordítását nagy valószínűséggel felismerjük, azaz a módszerrel nyelvfüggetlenül hasonlítsunk össze idegen nyelvű szövegeket magyar nyelvű fordításaikkal.

2 Nyelv és kultúra kapcsolata, nyelvi relativitás

A nyelvészetben már korábban is foglalkoztak azzal a gondolattal, miszerint a különböző nyelvek különböző gondolkodásmódokat takarnak, és meghatározzák használójuk világképét. Ezt a gondolatot először Wilhelm von Humboldt, német nyelvész és polihisztor, vetette fel a XIX. század elején. Később, a XX. század második felében, a kulturális antropológián belül indult meg a nyelvek kulturális összehasonlító vizsgálata, mely Edward Sapir és Benjamin Whorf nevéhez fűződik. A Sapir-Whorf hipotézis szerint a nyelv struktúrája és szemléletmódja meghatározza a valóságlátást és a külvilágból jövő ingerek érzékelését. Előadásomban ennek a gondolatnak egy sajátos megközelítésével kívánok foglalkozni: hogyan adható át egy gondolat két különböző szerkezetű nyelv között anélkül, hogy az átadott gondolat megváltozna, és statisztikai módszerekkel hogyan szűrhetők ki a fordítási hibák. A magyar nyelv a legtöbb európai nyelvtől különbözik. Az eltérő szerkezetű nyelvek fordítása során egy eltérő szerkezetű szöveg jön létre. Mivel az eltérések általában szisztematikusak, így statisztikai alapon vizsgálhatók. Feltételezésem szerint a nem megfelelő módon, szerkezetben átadott fordítás a megfelelő szerkezetű fordítástól eltérő statisztikai paraméterekkel rendelkezik, amely számszerűsíthető, ezzel kimutatva a hibás fordítást.

A Sapir-Whorf hipotézis szerint a világot a rendelkezésünkre álló fogalmakból tudjuk megérteni, és mivel ezeket a nyelv biztosítja számunkra, a más nyelven beszélők másképp látják a világot, más „megismerési univerzumban" élnek. Az elméletnek egy szélesebb körben elfogadott, enyhébb változata szerint a nyelvi különbségek hajlamossá tehetik az embereket, hogy másképp lássák a világot. [7] A hipotézist sok bírálat érte, miszerint a nyelv kultúra meghatározó szerepe nem igazolható, ugyanakkor az kevéssé vitatható, hogy a más nyelvet beszélők között kulturális különbségek vannak. A különböző nyelvek nemcsak eltérő szókinccsel és fogalmi repertoárral rendelkeznek, hanem az eltérő nyelvi szerkezettel a gondolatok szerveződésének egy egészen más módja valósul meg. Azt feltételezem, hogy az eltérő nyelvi szerveződések olyan rendszert alkotnak, amelyekben az eltérő nyelvi szerkezetek egyrészt megfeleltethetők egymásnak két különböző nyelv között, másrészt a szisztematikus eltérések részben számokkal mérhetővé tehetők.

A nyelvek közötti különbségnek azonban van egy másik vetülete is: nemcsak eltérő nyelvi szerkezetről van szó, hanem a gondolkozásnak egy olyan sajátos formájáról, amely ugyanarról a kérdésről, akár egészen más felfogásban fejeződik ki különböző nyelveken. A Sapir-Whorf hipotézis igazolására több kísérletet is végeztek. Ezek közül az egyikben az Egyesült Államokban élő kétnyelvű japán nőkkel készítettek interjúkat, akik mind a két nyelvet egyaránt jól beszélték. Két, egymástól elkülönült interjút készítettek, az első alkalom japán nyelven folyt, míg a második angol nyelven. Az interjú során azt a feladatot kapták az interjúalanyok, hogy egészítsék ki ugyanazokat a mondatokat, első alkalommal japánul, másodjára angolul. A gondolkodásmódban való eltérést, amikor különböző nyelven kellett megoldani a feladatot, a következő példák szemléltetik:

Amikor a vágyaim összeütközésbe kerülnek a családom vágyaival…
…ez nekem boldogtalanságot okoz. (japán)
…megpróbálom valóra váltani a vágyaimat. (angol)

Az igazi barátoknak…
…segíteniük kell egymást. (japán)
…őszintének kell lenniük egymáshoz. (angol) [6]

2 A vizsgálati módszer

A tanulmányban vizsgált szövegrészek Fabrizio Ciano Amikor nagyapa agyonlövette apát című könyvéből valók. [4], [5] Egészen pontosan annak első hat fejezete, amelyet meghatározott módszer szerint ötvenhárom szövegrészre osztottam, amely megfelelő esetszámnak bizonyul, ahhoz, hogy ezeken a szövegrészeken vizsgálatokat folytatva statisztikailag is értékelhető eredményeket kapjunk.

Az eljárás módszere: a szöveget meghatározott méretű szövegrészekre bontjuk, majd az adott szövegrészt és annak fordítását statisztikai szempontból összehasonlítjuk. A szövegrészek hosszúságának megállapításánál két szempontot kellett figyelembe venni. A kiválasztott szövegrész ne legyen túl hosszú, mert ebben az esetben az esetleges hibák, a szavak nagy száma miatt, elveszhetnek, minthogy az apróbb hibák magas szószám esetén nem befolyásolják lényegesen a szövegrész statisztikáját. Másrészt túl rövid szövegrész esetén a két összehasonlított szövegrészben szereplő szavak aránya nagyobb ingadozást mutathat, és kevésbé közelít az átlaghoz. A vizsgált mű esetében általában oldalanként haladtam. Ettől két esetben tértem el. Ha egy oldalon kevesebb, mint húsz sor volt, akkor az előtte lévő oldallal közösen vizsgáltam. Az egyik oldalról a másikra átnyúló mondatokat minden esetben azzal az oldallal együtt vizsgáltam, ahol elkezdődtek. A fejezetcímeket az összehasonlító vizsgálatból kihagytam, mivel magára a szövegre voltam kíváncsi. Természetesen a könyvben szereplő oldalakat, mint a vizsgálat tárgyát képező szövegrészeket, nem lehet állandó statisztikai egységnek tekinteni, még akkor sem, ha ebben az esetben beváltak. Megvizsgálva a kiválasztott szövegrészeket, nagyrészt 150 és 400 szó közötti egységeket alkotnak, átlagos hosszúságuk 258,1 szó. Úgy tűnik, hogy nagyjából a megadott méretű szövegrészek mellett, az eljárás nagy pontossággal lefolytatható, és statisztikailag értékelhető adatokat eredményez.

A statisztikai adatokat Word program segítségével kérdeztem le. A szavak számán kívül rögzítettem a szövegrészekben szereplő karakterek számát is, mind az eredeti, mind pedig a lefordított szövegrészekben. Az eljárás lefolytatásához olyan segédprogram is előállítható, amely részekre bontja az eredeti szöveget, viszont ebben az esetben is szükséges a szövegrészek fordításának kijelölése, megfeleltetése, illetve ellenőrzése. Ilyenkor a meghatározott hosszúságú, vagy ahhoz közelítő szövegrészek kiválasztását célszerű bekezdéstől bekezdésig kijelölni.

A karakterek számának elérése az olasz és a magyar szövegrészekben rendszertelen ingadozást mutatott, így ezt az eljárás során nem használtam fel. Annyit mindenesetre érdemes megjegyezni, hogy bár az olasz szövegrészek szószáma minden esetben nagyobb volt a magyar szövegrészek szószámánál, a vizsgált teljes szövegben a karakterek száma szóköz nélkül a magyar szövegben 2,5 százalékkal magasabb volt, mint az eredeti olaszban. A karakterek száma szóközökkel együtt pedig egy elhanyagolható 0,2 százalékos eltérést mutatott a magyar szöveg javára, tehát az olasz nyelvű

szöveg ugyanazt a gondolatot gyakorlatilag ugyanolyan hosszan fejezte ki írásban, mint a magyar.

Az idegen nyelvű szövegrészek és magyar fordításuk összehasonlításánál meg kellett határoznom, hogy a vizsgálat szempontjából mit tekintek önálló szavaknak. Erre azért volt szükség, mert az olasz nyelv, meghatározott szabályok szerint, gyakran rövidít aposztróffal bizonyos szavakat, elsősorban a névelőket és bizonyos esetekben a tárgyas névmásokat. Mivel az aposztróf olyan írásjel, amit, a többitől eltérően, nem szóköz különít el az írásjelet követő szótól így ezekben az esetekben a Word program statisztikai szempontból a két szót egynek tekinti. Így létre kellett hoznom egy olyan állományt, amelyben a kiválasztott szövegrészekben az aposztróf és az azt követő betű közé szóközt tettem. Ezt követően megvizsgáltam az így módosított szövegrészek jellemzőit. A módosított olasz szövegrészekben és azok magyar fordításában megszámoltam a program segítségével a szavakat, és megvizsgáltam azt, hogy az olasz szavak száma milyen arányban tér el a magyar szövegészben szereplő szavak számától. Statisztikailag megvizsgáltam azonban azt is, hogy a javítatlan szövegrészek szempontjából, ahol az aposztróf előtti és utáni szót a program egyként kezeli, hogyan alakul ez az arány. (Míg a módosított szövegrészekben az átlagos szószám 258,1 szó volt, ez a módosítás előtt, átlagosan 251,2 szó volt.)

A megvizsgált szövegrészekben az olasz szavak száma minden esetben nagyobb volt az azok magyar fordításában szereplő szavak számánál. Az eltérés 8,4 és 26,7 százalék között mozgott, tehát egy jól behatárolható 18,3 százalékos intervallumban. Ha mind a hat kiválasztott fejezetet osztatlanul, egyben vizsgáljuk, akkor az olasz szöveg szavainak száma 17,2 százalékkal haladta meg a magyar fordítás szószámát. Ha a javítatlan szövegrészekkel hasonlítom össze magyar megfelelőjüket, akkor az olasz szövegben szereplő szavak száma legalább hat és legfeljebb 24,8 százalékkal haladta meg a magyar szövegrészek szószámát, hozzátéve azonban azt, hogy ebben az esetben nem nyelvtani ételemben használom a szó kifejezést, hanem statisztikai egységként, két különálló szöveg összehasonlítására. Ebben az esetben is egy a korábbihoz hasonló intervallumba esnek a szavak, amely 18,8 százalékos értéket mutat.

A fejezeteket és azok fordításait külön is megvizsgálva a következő eléréseket találtam a szavak számában az olasz szöveg javára:

- 1. fejezet: 15,9 százalék;
- 2. fejezet: 14,8 százalék;
- 3. fejezet: 15,7 százalék;
- 4. fejezet: 15,7 százalék;
- 5. fejezet: 19,2 százalék;
- 6. fejezet: 19,8 százalék.

A fejezetek és fordításuk szószámának elétérése egy jól behatárolt ötszázalékos intervallumban mozog. A fejezeteket hét-tizenhárom szövegrészre osztottam. Ha a szövegrészek statisztikai adatai alapján szeretnénk következtetéseket levonni a fordítás helyességére, arra van szükségünk, hogy a százalékos eltérések viszonylag szűk intervallumba essenek, hiszen minél szűkebbre szabott ez az intervallum annál érzékenyebben reagál az eljárás a szavak számában való eltérésre. A fejezetek megfelelnek ugyan ennek a kitételnek, a hosszúságuk miatt azonban nem válnánk be érzékeny indikátorként, viszont egy összetett vizsgálati eljárás egyik összetevőjeként, más feltételek együttes megléte estén jól használhatók. A szövegrészek 18,3 százalékos intervalluma is elég jól behatárolt, de kerestem annak a lehetőségét, hogyan tudnám

csökkenteni, így megvizsgáltam a szavak számaránybeli eltérésének szélsőértékeit, mégpedig az alsó és felső tizedbe tartozó értékeket. Ha az alsó és felső tizedbe tartozó értékek nélkül nézzük a szavak számbeli elérésének arányait, akkor egy tizenkettő és huszonöt százalék közé eső, tizenhárom százalékos intervallumot kapunk, ebbe esik bele a megvizsgált szövegrészek nyolcvan százaléka.

Ha tehát hosszabb szövegrészeket vizsgálunk, akkor a fordítások ellenőrzésére, minőségvizsgálatára a következő eljárást célszerű alkalmaznunk: a szöveget először nagyobb gondolati egységekre bontjuk, amelyeket azután a korábban ismertetett módon, meghatározott méretű szövegrészekre bontunk. A szövegrészek vizsgálatánál a tizenkettő és huszonöt százalék közötti elérést érdemes minden esetben elfogadnunk, a száznyolc és száztizenkettő százalék közötti eltérést, valamint a huszonöt és huszonhét százalék közötti eltérést pedig abban az esetben, ha a nagyobb gondolati egység és annak fordítása között a szavak számában való eltérés tizennégy és huszonegy százalék közé esik (ebben az esetben a vizsgálatnál kapott szélsőértékeket, fölfele és lefele egy-egy százalékkal kitoltam). Így egy statisztikailag jól körülhatárolt, könnyen kezelhető és érzékeny indikátort kapunk. Ha valamelyik szövegrész nem felel meg a meghatározott statisztikai paramétereknek, szükséges a fordítás ellenőrzése!

Amikor az eltérő hosszúságú fejezeteket kezdtem vizsgálni, akkor ezt abból a megfontolásból tettem, hogy ezek a részek önálló narratív egységeket alkotnak, így önálló narratív struktúrával rendelkeznek, és ezek a struktúrák, azt feltételeztem, hasonlóságuk révén jóval kisebb statisztikai eltéréseket fognak mutatni egymáshoz képest, mint az őket alkotó szövegrészek. A fordítások statisztikai alapú összehasonlíthatóságát pedig részben magának a narratív struktúrának tulajdonítottam.

Bruner az emberi értelem működésének két módját különbözteti meg: az egyik a logikai-tudományos vagy más néven paradigmatikus mód, a másik pedig a narratív mód. A paradigmatikus gondolkodásmódról az évezredek folyamán tekintélyes tudás halmozódott fel, és tudományos egzaktság uralja, míg a narratívum ezzel szemben „az emberi szándékok viszontagságaival foglalkozik". A történeteknek látszólag végtelen lehetséges elbeszélési módja van, azonban ez még sincs egészen így. „Egyes nézetek szerint az életszerű narratívum egyfajta kanonikus vagy „legitim" szilárd állapottal kezdődik, amely törést szenved és válságba kerül, hogy azután orvoslást nyerjen, s e ciklus megismétlésére nyitva a lehetőség." [3] Azt feltételezem, hogy nemcsak a narratívum, hanem az azt tartalmazó szöveg formai sajátosságai is hordoznak egy olyan struktúrát, amely a narratívumot hordozó nyelv szerkezetével és statisztikai jellemzőivel párhuzamosan vizsgálható, és felhasználható a más nyelvre fordított megnyilatkozás szerkezetével és statisztikai jellemzőivel való összevetésre.

Miért kötöm inkább a narratív gondolkodásmódhoz a szóstatisztikák összevethetőségét? Könnyű belátni ennek az elképzelésnek a valószínűségét, ha belegondolunk abba, hogy a képletekkel, anyagnevek felsorolásával, kísérletek logikai leírásával foglalkozó logikai-tudományos gondolkodásmód nem olyan szerkezettel rendelkezik, amely lehetővé tenné a vizsgált eljárás tökéletes alkalmazását (az anyagnevek és képletek a fordításkor nem eredményeznek mérhető változást a szavak számában, és a leírások igeidő használta kisebb változatosságot mutathat). Ha viszont egy önálló szöveg narratív struktúrával rendelkezik, annak szerkezete, más nyelvre való lefordításakor szisztematikusan változik, és ez a változás statisztikailag mérhető eredményt hoz magával.

Ha nem nagyobb gondolati egységeket vizsgálunk, vagy nem kívánjuk nagyobb gondolati egységekre bontani a vizsgálandó szöveget, akkor a szövegrészek vizsgálatakor egy árnyaltabb kategorizálást használhatunk, a tizenkettő és huszonöt százalék közötti különbségre azt mondhatjuk, hogy a megadott határértékek között van, míg a tanulmányban az alsó és felső tizedbe eső eredményeket a statisztikailag elfogadható kategóriába soroljuk. A nyolc százalék alatti és a huszonhét százalék feletti tartományba eső eltéréseknél minden esetben javasolt a fordítás ellenőrzése.

Kizárható-e a fordítás helyessége, ha a tanulmányban megadott határértékeken kívülre esik a szavak eltérési aránya? Természetesen a fordítás helyessége ebben az esetben sem zárható ki teljesen, még akkor sem, ha ez statisztikailag nem valószínűsíthető, ezért is fogalmaztam úgy a bevezető részben, hogy az olasz szöveg szószáma „szinte" minden esetben nagyobb a magyar szöveg szószámánál. A határértékeken kívül eső szövegrészek ellenőrzése azonban minden esetben célszerű. Az alkalmazott eljárás nem arra szolgál, hogy egy adott szövegrészről megállapítsuk, hogy annak fordítása helyes-e, sokkal inkább arra, hogy kiszűrjük vele az egyértelműen hibásnak vélelmezhető szövegrészeket, és az esetleges fordítási hibákat a szöveg ellenőrzését követően kijavítsuk.

Milyen típusú hibák kezelésére alkalmas az eljárás, és melyekre nem? A szavak félrefordításából eredő hibákat nem tudjuk kiszűrni ezzel az eljárással, hiszen a statisztikában nem jelenik meg semmilyen eltérés. Az összehasonlított szövegrészek szószámának eltérése részben a magyartól eltérő szerkezetekből, részben pedig a kifejezések, szófordulatok eltéréseiből adódik, amely sok esetben szintén az eltérő nyelvet használók szemléletbeli különbségeire vezethető vissza. A szövegrész nem megfelelő szerkezetben történő átadása nem adja vissza a szavak arányainak eltérését, így a félrefordított szöveg könnyen kiszűrhető. Egy másik hibalehetőség, amikor a fordításból kimarad valami. Egy mondat, vagy akár egy fél mondat kimaradása is olyan elérést eredményezhet az eredeti arányokhoz képest, amely könnyen kiszűrhetővé teszi a hibát.

3 A nyelvhasználat és az eltérő nyelvi szerkezetek hatása az eredeti és a lefordított szöveg szóstatisztikája közötti különbségre

Nem csak két különböző nyelven megnyilatkozó ember beszéde, írása között találunk különbséget, hanem gyakran az azonos nyelvet beszélők megnyilatkozásaink formája is eltérhet egymástól. Az eltérő nyelvhasználatot nem csupán az eltérő gondolkodásmód okozza, hanem az adott társadalmon belüli rétegződésbeli, tanultságbeli és szocializációs különbségek. Ezen kívül még számolnunk kell a különböző nyelvjárásokban beszélők nyelvi normájának különbözőségével is, amely bizonyos nyelvek esetében markáns különbségeket mutathat. A felsorolt különbségek részben kulturális jellegűek, de bizonyos esetben az adott nyelv tökéletlen elsajátítása is eredményezheti. Meg kell tehát vizsgálnunk, hogy a felvázolt elmélet alkalmazható-e ezekben az esetekben is, és ha igen, akkor milyen megkötésekkel, valamint azt is, hogy az adott nyelven belüli különbségek miben térnek el az adott nyelv sztenderdjétől.

A különböző társadalmi rétegek eltérő nyelvhasználatának társadalmi hátrányokat továbbörökítő hatásával már több szerző is foglalkozott. Basil Bernstein ennek első-

sorban szociolingvisztikai hátterével foglalkozott, míg Bourdieu inkább szociológiai szempontból vizsgálta a kérdést.

Bernstein szegényebb és gazdagabb gyerekek beszédhasználatát vizsgálta, nem a szókincs vagy a verbális képességek különbségei érdekelték, hanem a nyelvhasználat szisztematikus különbségei. Azt tapasztalta, hogy az alsóbb osztályok nyelvhasználatára a korlátozott kód jellemző, ami az jelenti, hogy a nyelvet sok olyan előfeltevéssel használják, amelyről azt felételezik, hogy a hallgató számára is ismertek, azaz olyan nyelvhasználatról van szó, ahol a mondanivaló nem választható le a helyzetről, amelyben létrejött. A felsőbb osztályok gyermekeire a kidolgozott kód használata jellemző, ami azt jelenti, hogy a mondanivaló leválasztható arról a helyzetről, amelyben létrejött, kevésbé kontextusfüggő, így ezek a tanulók könnyebben fejeznek ki általánosításokat és elvont fogalmakat is. Bernstein szerint azok a gyerekek, akik a kidolgozott kódot sajátították el, sikeresebben küzdenek meg az iskolai próbatételekkel, hiszen az oktatás kidolgozott kódban folyik, így ők egy ismerős nyelvi közeggel találkoznak, míg az alsóbb osztályok gyermekei könnyen kudarcként élik meg az iskolai nyelvhasználattal való találkozást. [8]

Bourdieu az iskola szerepéről írva egészen odáig megy, hogy az iskolának komoly része van a társadalmi szelekció fenntartásában – amelyben nyilvánvalóan szerepe van a Bernstein-féle nyelvhasználati különbségeknek is:

„A kiváltságos osztályok egyre teljesebb mértékben az iskolára ruházzák át szelekciós hatalmukat. Úgy tűnik, mintha ezzel egy teljesen semleges hatalom javára mondanának le a nemzedékek közötti hatalomátadás hatalmáról, s mintha feladnák a kiváltságok átörökítésének kiváltságát. Az iskola eljárása azonban a következő: formailag kifogástalan ítéleteket hoz. Ezek objektíve mindig az uralkodó osztályt szolgálják, hiszen még technikai érdekeit sem sértik soha – hacsak nem társadalmi érdekeik védelmében. Ilyen módon az iskola minden eddiginél jobban – s egy demokratikus ideológiára hivatkozó társadalomban az egyetlen elképzelhető módon működik közre a fennálló rend e reprodukciójában, mert minden eddiginél jobban leplezi el azt a funkciót, amelyet betölt…" [2]

Ha ilyen komoly különbségek lehetnek egy adott nyelvet használók között, akkor vajon alkalmazhatjuk-e ilyen esetekben a szóstatisztikákról vázolt összefüggéseket, eredményeket fordítások összehasonlításánál? Ha igen, milyen megkötéseket kell tennünk az elmélet felhasználásával kapcsolatban, szükséges-e ilyen megkötések megtétele? Hogyan illeszthető be ez az eddig vázolt elméletbe? Ahhoz, hogy erre a kérdésre kielégítő választ kapjunk, magában a szövegben kell számba vennünk, hogy milyen tényezők befolyásolják a szóstatisztikákban mutatkozó különbségeket. Ezek a tényezők a következők:

I. Eltérő nyelvi szerkezet. Az olasz nyelvet vizsgálva azt tapasztalhatjuk, hogy amikor egy bizonyos gondolatot megfogalmazunk, akkor a magyarhoz képest bizonyos szavaknak, szófajoknak az előfordulási gyakorisága minden esetben magasabb, ugyanannak a gondolatnak a magyar megfogalmazásához képest. Vannak azonban olyan szavak, amelyeknek a fordítást követően, azok magyarra fordításánál nem következik be lényeges változás. Vegyük sorjában a változásokat előidéző szerkezetbeli különbségeket! A felsorolás természetesen nem lehet teljes körű és részletekbe menő, hiszen az kimerítené ennek a tanulmánynak a kereteit, viszont fontos áttekintenünk azokat a főbb szerkezei eltéréseket a két nyelv között, amelyek meghatározzák a szóstatisztikában való elérést, hogy lássuk, hogy olyan szerkezeti sajátosságokról van

szó, amelyek szisztematikus eltérést mutatnak, és amelyek bizonyos keretek között kiszámíthatóvá teszi a magyar és az olasz szövegek statisztikai összehasonlítását.

Az igék és az összetett igeidők használata során a szavak száma az olaszul megfogalmazott szövegrészekben gyakorlatilag szinte mindig magasabb lesz, mint azok magyar megfelelőiben. Ennek egyik oka az, hogy az olasz nyelv gyakran olyan esetekben is használja a létige ragozott alakját, amikor a magyarban ezt nem használjuk. Másrészt, míg a magyar nyelvben egyféle múlt időt használunk kijelentő módban, az olaszban ötféleképpen fejezhetjük ki ugyanebben a módban egy bizonyos cselekvés múltidejét (passato prossimóval, imperfettóval, trapassato prossimóval, passato remotóval és trapassato remotóval). Ezek közül az igeidők közül kettőnél az ige ragozott alakját használjuk (imperfetto, passato remoto), míg a másik három esetben összetett múlt időt. A passato prossimónál egy az avere vagy essere segédige jelen idejű ragozott alakjából és egy múlt idejű melléknévi igenévből álló szerkezettel fejezzük ki a múlt időt. Hasonló a helyzet a trapassato prossimónál és a trapassato remotónál is, azzal a különbséggel, hogy előbbinél az avere vagy essere segédigét imperfettóban, míg utóbbinál passato remotóban ragozzuk. A szenvedő és a műveltető szerkezet is az olasz szöveg szószámát növeli a magyar megfelelőhöz képest.

További eltérést eredményez a magyar és olasz szöveg szóstatisztikájában, hogy az olasz nyelv használ elöljárószavakat, míg a magyar nyelvben ezek funkcióját a ragok és névutók tölti be. Amennyiben névelőt is használunk, az az olasz nyelvben összeolvad az elöljárószóval. A részelő névelő használata is ismeretlen a magyar nyelvben.

A névmások használatában is jelenős elérés mutatkozik a magyar és az olasz nyelv között. Az olaszban ismeretlen a tárgyas igeragozás, így az olaszban a tárgyesetű személyes névmást minden esetben ki kell tennünk, amikor a magyarban tárgyas ragozást használnánk. A magyar nyelvben azonban a tárgyas ragozás már önmagában is kifejezi a tárgyat, így egyes szám 1. személyű alany és 2. személyű tárgy esetén nem szükséges kitennünk, mint ahogy az egyes szám 3. személyű tárgyat sem szükséges kitenni. A többes szám 2. és 3. személyű tárgyat viszont a magyarban is mindig kitesszük.

A birtokos jelző kifejezésére az olaszban olyan szerkezeteket használunk, amelyek szintén a szavak számának eltérését eredményezik a magyar nyelvű szerkezethez képest. Míg a magyarban a birtoklás kifejezésére a birtokhoz mindig birokos személyragot teszünk, az olaszban ezt kifejezhetjük egy birtok + di elöljárószó + birokos szerkezettel, jelzői birtokos névmással (birokos determinánssal), valamint használhatunk személyes névmást a birtokos determináns helyett. „A birtokosra igen gyakran nem a birokos determináns, hanem a személyes vagy visszaható névmás *hangsúlytalan részes esete* utal. Ez jellegzetesen olaszos, a *magyartól teljesen eltérő* szerkezet.” [1]

Az olasz szövegben szereplő főnevek, melléknevek, számnevek általában a magyar szövegbe is ekként kerülnek lefordításra, de itt is lehetnek kivételek, például ha az olasz szövegben szereplő jelzős főnév a magyarban egy olyan kifejezést alkot, amely egy összetett szó, vagy ha az olasz főnév megfelelőjét a magyarban két szóban írjuk.

Az olasz és a magyar szövegrészek szószámbeli elérését leginkább a nyelvtani szerkezeti különbségeknek tulajdonítom, amelyek természetesen a gondolkozásbeli, szemléletmódbeli különbséggel függenek össze. Természetesen a továbbiakban leírt okok is közrejátszanak a szavak számának elérésében, de önmagukban nem lennének

elegendőek ahhoz, hogy egy jól behatárolt keretek között mozgó, szisztematikus eltérést vizsgáljunk.

II. Eltérő kifejezés- és gondolkodásmód. Amikor arról írtam, hogy a kétnyelvű japán nők gondolkodásmódja eltér, amikor japánul, illetve angolul kellett megválaszolniuk egy kérdést, eszembe jutott néhány olasz közmondás, amelyek magyarul egészen másképp hangzanának, ha szó szerint próbálnánk lefordítani őket, de néhány gyakori szófordulattal is ez a helyzet. A szófordulatok részben követik a nyelv eltérő szemléletmódjából adódó struktúrákat, részben pedig gyakoriságuk alapján vagy eleve szerepelnek a már meghatározott arányszámokban, vagy ha ritkán fordulnak elő, eleve nem befolyásolják lényegesen a kialakított eltérési arányszámokat.

III. Nyelvjárás, nyelvi rétegződés. A Wikipédia honlapján a következőket olvashatjuk az Olaszországban használt dialektusokról: „Az olasz dialektológia az újlatin nyelvészet egyik leggazdagabb területe. Olaszország területén számos (egyes becslések szerint 200 körüli) újlatin dialektust és aldialektust (nyelvjárást) használnak. Ezek az olasz dialektusok (*dialetti italiani*) északról dél felé haladva erősen különböznek egymástól, oly mértékben, hogy két távolabbi beszélő meg sem érti egymást: így a kölcsönös érthetőség végett mindenkinek beszélnie kell a sztenderd olasz nyelvet (olasz köznyelv).

Az erős dialektális tagolódás oka az egységes Olaszország, illetve az olasz irodalmi nyelv késői kialakulása volt. Sok olasz dialektust ma már teljesen önálló újlatin nyelvként tartanak számon, amelyek már saját helyesírással is rendelkeznek. Ilyenek a szicíliai, a nápolyi, az emilián-romanyol, a velencei, a lombard, a ligur, a piemonti, a szárd és a korzikai.” [9]

„Az olasz irodalmi nyelv alapjául a középolasz dialektuscsoport, ezen belül elsősorban a középkori toszkán dialektus szolgált. Az ebből kialakuló mai sztenderd olasz nyelvre azonban a többi közeli középolasz dialektus, így a római dialektus is hatást gyakorolt. Érdekes módon a mai toszkán dialektus a sztenderd olasz nyelvtől a kiejtésében észrevehetően elkülönül, például egész Olaszországban egyedül a toszkán dialektus használja a 'h'-mássalhangzót a 'k' helyett: például a 'come' szó *hóme* ejtése a köznyelvi *kóme* helyett. A köznyelvi 'cs'-mássalhangzót - a rómaihoz hasonlóan - a toszkán dialektus is "s"-nek ejti, szemben a sztenderd olasz ejtéssel: például a 'cinquecento' szó *sinkvesento* ejtése a köznyelvi *csinkvecsento* helyett.” [9]

Ahogy látjuk, egyrészt megkülönböztethetünk olyan dialektusokat, amelyek az idők folyamán önálló nyelvvé váltak, olyanokat, amelyek erősen eltérnek az irodalmi olasz nyelvtől, és olyat is, amely hangzóiban, egyes tájszavaiban tér el a mai olasz nyelvi sztenderdtől. Amikor egy dialektus annyira eltávolodik a nyelv általánosan bevett normáitól, hogy önálló nyelvvé válik, akkor erre a nyelvre külön meg kell határoznunk a szöveg fordításából adódó elérési arányt a szavakra. Ugyanez a helyzet az irodalmi, illetve a köznyelvtől való jelentős eltérés esetén is. Fordítási szempontból ezt az eltérést akkor tekinthetjük jelentősnek, ha a két nyelv szóstatisztikáinak eltérései rendre a meghatározott intervallumon kívülre esnek.

A nyelvi rétegződés szóstatisztikára gyakorolt hatásával foglalkozva szintén azt kell szem előtt tartanunk, hogy a nyelvi normáktól való eltávolodás milyen mértékben zajlik le, és mekkora hatással van az olasz és a magyar nyelv szóstatisztikája közötti különbségre. A korlátozott és a kidolgozott kód kérdéskörére visszaérve, azt feltételezem, hogy a korlátozott kódban elhangzó megnyilatkozások az adott nyelv nyelvtani szerkezetét követik, egy leegyszerűsített, hiányos szerkezetben, amely a magyar nyel-

vű fordításnál valószínűleg olyan szisztematikus eléréseket eredményez, amely a megadott statisztikai határértékek közé esik. Megjegyzem ugyanakkor, hogy az írásban elhangzó közlések sajátossága, hogy kidolgozott kódban fogalmazódnak meg, és ha esetleg irodalmi környezetben meg is jelennek egy bizonyos társadalmi közeg bemutatására, hatásuk még ebben az esetben is elhanyagolható, hiszen egyrészt a szövegkörnyezet, amelyben szerepelnek kidolgozott kódban fogalmazódik meg, másrészt a cselekmény megértése szükségessé teszi, hogy a korlátozott kódot használók előfeltevéseit, a mondandót, amely nem választható le a helyzetről, amelyben létrejött, a szerző egyértelművé tegye az olvasó számára. Nem tartom azonban kizártnak, hogy a vizsgált szövegek között olyan jól körülhatárolható kategóriákat találjunk, amelyek statisztikai tulajdonságaikban eltérhetnek egymástól. Ezek azonban nem a felvázolt eljárás cáfolatai, hanem annak árnyalásai az eljárás szabályszerűségeinek felhasználásával.

III. Idegen nyelvű szövegek beékelődése a szövegbe. Amikor olasz nyelvű szöveget fordítunk magyarra, figyelembe kell vennünk a szövegben esetlegesen hosszabb terjedelemben szereplő idegen nyelvű idézeteket. Ha egy olasz szövegben, illetve annak részekre bontásánál egy szövegrészben hosszabb idegen nyelvű idézet van (például angol, német vagy francia), amit lábjegyzetben magyaráz meg a fordító, és ami így változatlan formában és szószámban kerül be a magyar szövegbe, a két szöveg statisztikai összehasonlítását nyilvánvalóan befolyásolhatja. Hosszabb idézet vagy beékelés esetén érdemes a vizsgált nyelvtől eltérő, idegen nyelvű szöveg nélkül összehasonlítani a szóstatisztikákat.

IV. A helyesírás aktuális szabályai. Amikor egy adott szöveget és annak fordítását vizsgáljuk, figyelembe kell vennünk az eltérési arányok meghatározásánál az egybe és különírás időszerű szabályait, amelyek időről időre változhatnak, valamint meg kell vizsgálnunk a szóban forgó nyelv és a magyar nyelv, adott korra jellemző nyelvhasználatát.

4 Összegzés

Olasz nyelvű szövegek magyar fordításának vizsgálata alapján igazolódni látszik az a feltevés, hogy az olasz szöveg szószáma szinte minden esetben magasabb a magyar szöveg szószámánál, és ez az eltérés jól behatárolható értékek között mozog. Ha bizonyos szövegrészek statisztikája a megadott határértékeken kívülre esik, akkor minden esetben szükséges az adott szövegész fordításának ellenőrzése. Amennyiben a megadott szövegrészek rendre eltérnek a megadott határértékektől, és a fordítás is helyes, akkor vagy olyan szöveggel van dolgunk, amelyre nem alkalmazható az eljárás (például képleteket, tudományos leírásokat, és anyagneveket felsoroló szöveg), vagy egy olyan szöveggel, amelyre eltérő statisztikai határértékek a mérvadók, amely az adott szövegtípusra is meghatározható (például nyelvjárási szöveg, rétegnyelv vagy egy jól körülhatárolható szövegtípus).

Feltételezhető, hogy az eljárás elsősorban a narratív struktúrákhoz kötött, az elbeszéléshez, és az elbeszélésben szereplő igeidők és szerkezetek váltakozásához, továbbá az is, hogy nemcsak a narratívum, hanem az azt tartalmazó szöveg formai sajátosságai is hordoznak egyfajta struktúrát, amely a narratívumot hordozó nyelv szerkeze-

tével és statisztikai jellemzőivel párhuzamosan vizsgálható, és felhasználható a más nyelvre fordított megnyilatkozás szerkezetével és statisztikai jellemzőivel való összevetésre. Mivel szisztematikus eltéréseket vizsgálunk, ezek más nyelvre is kiterjeszthetőek, amelyre szintén külön meg kell határoznunk a szóstatisztikák közötti eltérés intervallumát.

A kutatás további lehetséges irányai: az eljárás alkalmazhatóságának vizsgálata más nyelveken és más szövegekre, valamint egy adott mű több különböző magyar nyelvű fordításának összevetése. Az így kapott eredmények tovább árnyalhatják a tanulmányban ismertetett módszert, és lehetőséget adhatnak az eredmények szélesebb körű felhasználására.

Hivatkozások

1. Angelini, M. T., Móritz Gy.: Gyakorlati olasz nyelvtan, Nemzeti Tankönyvkiadó, Budapest, (2006)
2. Bourdieu, P.: A társadalmi egyenlőtlenségek újratermelődése. Tanulmányok. Fordította: Ádám P., Ferge Zs., Léderer P. Gondolt, Budapest, (1978)
3. Bruner, J.: A gondolkodás két formája. Forrás: László János, Thomka B. (szerkesztette): Narratívák 5. Narratív pszichológia. Kijárat Kiadó, Budapest, (2001) 27-58
4. Ciano, F.: Amikor nagyapa agyonlövette apát. Fordította: Puskás L. Kézirat.
5. Ciano, F.: Quando il nonno fece fucilare papà. A cura di Cimagalli, D.. Arnoldo Mondadori Editore, Milano, (1991)
6. Farb, P.: Word Play: What Happens When People Talk. Vintage Books, New York, (1993)(http://cyberartsweb.org/cpace/theory/luco/Hypersign/Language.html)
7. Forgács J.: A társas érintkezés szociálpszichológiája. Fordította: László J. Gondolat Könyvkiadó, Budapest, 2. kiadás, évszám nélkül.
8. Giddens, A.: Szociológia. Osiris Kiadó, Budapest. Fordította: Babarczy E., Nagy M., Nagy Zs., Tóth L. (1995)
9. Olasz nyelv. Forrás: http://hu.wikipedia.org/wiki/Olasz_nyelv

Kísérletek statisztikai és hibrid magyar–angol és angol–magyar fordítórendszerek megvalósítására[1]

Novák Attila, Prószéky Gábor

MorphoLogic
1116 Budapest, Kardhegy u. 5.
{novak,proszeky}@morphologic.hu

Kivonat: Cikkünkben két olyan kísérletről számolunk be, amelyek arra irányultak, hogy a tisztán szabály alapú MetaMorpho rendszerünknél jobb minőségű fordításokat hozzunk létre. Két ilyen rendszer készült: az egyik rendszerben a Moses statisztikai dekódert használtuk a MetaMorpho által előállított fordítások rangsorolására, illetve a részleges fordításokból teljes fordítások összeállítására; a másik kísérleti rendszer egy tisztán statisztikai morfémaalapú magyar–angol fordítórendszer volt. Előbbi rendszerünkkel a tisztán szabály alapú rendszernél kicsit jobb minőségű fordítást kaptunk, az utóbbi azonban gyengébb eredményeket produkált.

1 Bevezetés

A MorphoLogic MetaMorpho fordítórendszere (Novák, Tihanyi & Prószéky, 2008) egy sok emberévtizednyi munkával létrehozott szabályalapú fordítóprogram, amely a magyar és az angol nyelv között mindkét irányban képes fordítani. Időközben létrejöttek ezen nyelvpár tagjai közötti automatikus fordítást kínáló más kísérleti és üzleti alkalmazások, illetve online szolgáltatások is. Ezek között megjelentek a statisztikai gépi fordítási paradigma keretében készült rendszerek is, ám ha az anonimizált gépi fordítások szubjektív emberi minőségi rangsorolását tekintjük mércének, mind a mai napig a MetaMorpho kínálja a legjobb minőségű fordítást. Ebben a cikkben két olyan kísérletről számolunk be, amelyekben a MetaMorphóénál jobb minőségű fordítást produkáló fordítórendszereket próbáltunk létrehozni.

Az eredeti MetaMorpho rendszerben a lehetséges fordítási opciók közötti választás sok esetben nem feltétlenül optimális. Ha a rendszerbe épített mondatelemzőnek sikerült teljes elemzést előállítania a lefordítandó mondathoz, akkor egyszerűen a legelsőként előálló elemzésnek megfelelő fordítást adja vissza, ahelyett hogy esetleg több lehetséges fordítást előállítana, és azok közül választaná ki a legjobbat. Abban az esetben pedig, amikor nem áll elő a fordítandó mondathoz teljes elemzés, és a program részfordításokból próbál a teljes mondatot lefedő fordítást összeállítani, a részfordítások kiválasztásánál nem ellenőrzi, hogy az egyes fordításrészletek a célnyelven

[1] Ehhez a kutatáshoz az Európai Bizottság részleges támogatást nyújtott az EuroMatrix (FP6-IST-5-034291-STP) projektum keretében. Szeretnénk köszönetet mondani Laki Lászlónak és Siklósi Borbálának statisztikai fordítórendszerünk létrehozásában való közreműködésükért.

mennyire jól illeszkednek egymáshoz. Ezért úgy döntöttünk, hogy létrehozunk egy olyan kísérleti hibrid fordítórendszert, amelyben mind a teljes fordítások rangsorolására, mind a részfordítások kiválasztására és azokból a teljes fordítás összeállítására a MetaMorpho eredeti algoritmusa helyett a Moses statisztikai dekódert használjuk (Koehn és munkatársai, 2007).

Létrehoztunk emellett egy teljesen statisztikai alapon működő alternatív fordítórendszert is (szintén a Moses felhasználásával), amelyben a hagyományos szóalapú megoldás helyett morfématokeneket használtunk. Ezt a megoldást a magyar és az angol nyelv közötti alapvető strukturális különbségek és az ezek által okozott szómegfeleltetési (alignment) problémák motiválták, amelyek a jelenleg elterjedt frázis alapú statisztikai gépi fordítási paradigmában alapvetően behatárolják az angol–magyar viszonylatban elérhető fordítási minőséget. Sajnos azonban utóbbi rendszerünk nem bizonyult sikeresnek: az általa generált fordítások minősége mind a BLEU-pontszám, mind a szubjektív emberi megítélés szempontjából messze elmaradt a szabályalapú rendszer (és a szóalapú statisztikai rendszerek) fordításainak minőségétől.

2 A MetaMorpho fordítórendszer

A MorphoLogic MetaMorpho szabályalapú fordítórendszere strukturálisan különbözik a legelterjedtebb szabályalapú fordítórendszerektől: nem tartalmaz külön transzfer komponenst. Nyelvtana (beleértve a lexikont is) olyan mintapárokból áll, amelyeknek egyik tagját a forrásmondat (alulról felfelé történő) elemzésekor használja a fordítórendszer mondatelemzője, és az ehhez tartozó célnyelvi mintát (vagy több célnyelvi minta valamelyikét) felhasználva generálja az adott forrásnyelvi mondatrészlet célnyelvi megfelelőjét a fordítás (felülről lefelé történő) generálásakor. A mintapárok tagjai jegyekkel kibővített kontextusfüggő szabályok. A nyelvtan architektúrája teljesen homogén: az általános szerkezeti szabályoktól a többé-kevésbé idiomatikus frázisokon keresztül a teljesen lexikalizált szótári tételekig minden nyelvi elemet és azok fordítását azonos módon ábrázolja, ezek csak az egyes elemek alulspecifikáltságának mértékében különböznek egymástól.

A célnyelvi szerkezetek létrehozása és a lexikai elemek beillesztése nem igényel utólagos transzfer műveletet: a forrásnyelvi elemzési fa részstruktúráinak az alkalmazott szabálypárok szerint megfelelő célnyelvi struktúrákat csak ki kell olvasni, és azokat a célnyelvi szóalak-generátor közvetlenül fordítássá alakítja.

A MetaMorphóban a forrásnyelvi szöveg elemzése az alábbi lépésekből áll. Az első lépés a szöveg mondatokra bontása. Ezt a szavakra bontás, azaz tokenizálás és a tokenek morfológiai elemzése követi, amely morfoszintaktikai jegyvektorokat rendel a tokenekhez. Ezután következik a többértelmű tokensorozatok által alkotott háló elemzése a nyelvtan forrásoldali szabályainak felhasználásával. A nyelvtanban jegyeket használunk egyrészt az elemzett szövegre vonatkozó lexikai, morfoszintaktikai és vonzatkeret-információk tárolására, másrészt arra, hogy az elemzési, illetve generáló szabályok alkalmazhatóságára vonatkozó megszorításokat tegyünk (pl. másként fordítunk egy igét, ha az alanya élő, mint ha nem az).

Amikor az elemzés kész, és nem marad több illeszthető elemzési szabály, a fordítás a forrásnyelvi mondat elemzési fáját felülről (a mondatszimbólumtól kezdve) bejárva az egyes forrásnyelvi részstruktúráknak megfelelő célnyelvi struktúrák kombinálásával, a bennük szereplő lexikai és morfoszintaktikai jegyegyüttesek interpretációjával áll elő. A forrásnyelvi szabályok bármelyikéhez egynél több célnyelvi szabály is tartozhat. Az adott esetben alkalmazandó célnyelvi megfelelő kiválasztásakor a rendszer az adott forrásnyelvi szabály alkalmazásakor kitöltött jegyekre tett megszorításokra támaszkodik.

A klasszikus transzfer alapú fordítóktól eltérően, a MetaMorphóban a fordításkor alkalmazandó szórendi átrendezéseket is a forrásnyelvi szöveg elemzése során alkalmazott szabályok és az elemzési fában kitöltött jegyek tulajdonképpen már elemzési időben meghatározzák. A kimenet generálásakor csak a már meghatározott és átrendezett struktúrák szöveggé alakítása történik. A generált célnyelvi fa terminális pontjain levő morfoszintaktikai és lexikai jegyegyüttesek interpretálását a célnyelvi szóalak-generátor végzi, amely a megfelelő célnyelvi szóalakokat előállítja.

A többértelműségek kezelése a tisztán szabályalapú rendszerekben mindig nehéz. A MetaMorpho két módszert alkalmaz a nem kívánt többértelműségek kiszűrésére: vagy magas szintű heurisztikákat használ az alternatívák közötti választásra (pl. egy összetevőnek vonzatként való elemzését preferálja a szabad határozóként való elemzés helyett), vagy a specifikusabb szabályok explicit módon blokkolják az adott esetben nem alkalmazandó általánosabb szabályok alkalmazását.

Általában a MetaMorpho csak az első sikeres elemzéshez tartozó első lehetséges fordítást állítja elő. Kellően hosszú, és megfelelő számú lehetséges strukturális többértelműséget tartalmazó fordítandó mondatok esetében azonban így is előfordulhat, hogy elemzés közben túl sok hipotézis áll elő. Eredetileg erre a problémára az volt a megoldás, hogy az elemző egyszerűen leállt azon a ponton, amikor egy beállított időkorlátot túllépve túl sok időt töltött egy mondat elemzésével. Ez a megoldás ugyan biztosítja azt, hogy a fordítórendszer válaszideje minden bemenetre korlátos maradjon, azonban ennek a megoldásnak az eredményeképpen az ilyen, túl hosszú mondatokra olyan fordítás jött létre, amely a mondat végén lefordítatlanul maradt szavakat tartalmazott. Erre a problémára jobb megoldást sikerült találni azzal, hogy a túl hosszúnak tűnő mondatokat már a mondatokra bontás során rövidebb egységekre bontjuk (a korábbinál agresszívebb módon), és így már szinte egyáltalán nem fordul elő, hogy szükség lenne az elemzés idő előtti megszakítására, és ennek megfelelően sokkal ritkábban maradnak lefordítatlan részek a fordításban.

3 A hibrid fordítórendszer

Elemzés közben a MetaMorpho mondatelemzője hierarchikusan egymásba épülő részleges szintaktikai struktúrákat állít elő. Ha nem sikerül teljes elemzést találni az adott lefordítandó mondathoz, akkor a MetaMorpho jobb híján egy olyan heurisztikát alkalmaz, amely ezekből a részleges struktúrákból egy a teljes bemenő mondatot mintegy mozaikszerűen lefedő sorozatot kiválasztva állítja elő a fordítást. Az így előálló fordítások általában nem optimálisak, mert a teljes elemzés hiányában bizonyos strukturális (pl. az egyeztetéssel kapcsolatos) információ elvész.

3.1 A névmástörlés

A magyar–angol fordítási irányban a magyar névmások kiesése (az ún. pro-drop) további problémát jelent, amikor részfordításokból próbáljuk a teljes fordítást összerakni. Mivel az alany számát és személyét, vagy tárgyas igék esetében a tárgy határozottságát az igeragok általában önmagukban pontosan jelzik. Az explicit alanyi és tárgyi névmások tehát a magyarban általában elhagyhatók, és gyakran el is hagyjuk őket (hacsak nem állnak fókuszban, vagy egyéb módon kiemelten hangsúlyosak). A probléma az, hogy pontosan ugyanazokat az igealakokat használjuk kitett teljes alany és tárgy mellett, mint amiket az elhagyott névmások esetében. Ebben az esetben azonban ugyanezek az igei végződések nem tartalmaznak inkorporált névmást, és hiba, ha a fordításban névmás jelenik meg.

Hallja.	*He/she/it hears him/her/it.*
Fred hallja a doktort.	*Fred hears the doctor.*

Puszta (egyszavas) magyar igealakok fordításaként a MetaMorpho kizárólag olyan angol frázisokat generál, amelyek explicit alanyi névmást tartalmaznak (illetve határozott tárgyas igealakok, pl. a *hallja* esetében tárgyi névmást is: *he hears it*), mert az igéket a nyelvtanban kizárólag a vonzataikat is tartalmazó lexikai minták reprezentálják. Ennek következtében fölösleges beszúrt névmások jelenhetnek meg azokban a mozaikszerűen összerakott fordításokban, ahol testes alany, illetve tárgy is szerepel a mondatban, abban az esetben, ha az algoritmus olyan forrásnyelvi részmondat fordítását is felhasználja, amelyben explicit alany vagy tárgy nem szerepelt.

Hasonló jelenség figyelhető meg a harmadik személyű birtokos szerkezetek esetében (itt birtokos névmások jelenhetnek meg birtokos szerkezetek helyett):

háza	*his house*
Fred háza.	*Fred's house.*

Egy példa a fentiekre a következő fordítás:

Bemenet: *A repülő objektumok + nem viselkednek teljes mértékben úgy, mint ahogy az az ősi gravitációs törvény + alapján + elvárható + lenne.*

MMO: *The flying objects + they do not behave in a full measure the way that ancient gravitational law + his basis + can be expected + he would be.*

3.2 A Moses dekóder bevetése

Az eredeti részfordítás-kombináló algoritmus nem használ célnyelvi nyelvmodellt arra, hogy a lehetséges részekből összerakott fordításokat rangsorolja. Kísérleteinkben az eredeti algoritmust statisztikai modellel helyettesítettük. A hibrid rendszerben a MetaMorphót a nyílt forráskódú Moses statisztikai dekóderrel kombináltuk (Koehn és munkatársai, 2007): a szabályalapú komponens által előállított részfordításokat, illetve teljes fordításokat tartalmazó frázistáblából a Moses dekóder állít össze és

választ célnyelvi nyelvmodell felhasználásával optimalizált fordítást. Azt reméltük, hogy így az eredetinél jobb minőségű fordítást kapunk ezekben az esetekben. A MetaMorpho elemzőjét kiegészítettük egy olyan felülettel, amely az elemzés közben létrejött összes részstruktúrát a lehetséges fordításaival együtt ki tudja írni a Moses dekóder frázistáblájának megfelelő formátumban.

Ennek felhasználásával aztán a Moses dekóder segítségével generáltunk célnyelvi nyelvmodell felhasználásával optimalizált fordítást az eredeti fordítandó mondatokra. Mivel jobb becslésünk nem volt a fordítási valószínűségekre, egyenletes eloszlást feltételeztünk a frázistáblában az egyes frázisok lehetséges alternatív fordításai felett, és a Moses konfigurációjában zérus súlyt rendeltünk a fordítási modellhez. Lexikalizált torzítási modellt sem használtunk (a statisztikai fordítási zsargonban a szórendi átrendezést nevezik torzításnak). Így a dekóder a célnyelvi nyelvmodell által a fordításhoz rendelt pontszám alapján rangsorolja a fordításokat. Kísérleteinkben 5-gram (5 szavas) nyelvmodellt használtunk, amelyet a Hunglish korpusz (Halácsy és munkatársai, 2005) jogi és irodalmi részéből generáltunk. Sajnos nagyobb egynyelvű korpuszokból generált nyelvmodellek előállítását és betöltését a használt tesztgépben levő operatív memória menyisége nem tette lehetővé. [2]

Számos paraméterbeállítással és frázistábla-építési módszerrel kísérleteztünk. A teljes elemzéssel rendelkező mondatok esetében a részfordítások felvétele a frázistáblába a fordítási minőség egyértelmű romlásához vezetett. Ugyanakkor – nem meglepő módon – az összes lehetséges teljes fordítás felvétele a frázistáblába (ha volt a mondatnak sikeres teljes elemzése) és a legjobb fordítás nyelvmodell segítségével való kiválasztása a MetaMorpho-alapértelmezéssel, azaz az első sikeres elemzésnek megfelelő fordítást kiíró megoldással szemben egyértelműen javította a fordítás minőségét. A dekóder konfigurációs fájljában meg kellett növelnünk a maximális megengedett frázisméret értékét az alapbeállításról ahhoz, hogy a dekóder a teljes mondatfordításokat is használja (ellenkező esetben nagyon drasztikusan romlott a fordítások minősége).

Szintén kedvező hatása volt, ha azokhoz a frázisokhoz, amelyeknek a fordítása esetleg felesleges névmást tartalmazott, olyan alternatív fordításokat is generáltunk a frázistáblába, amelyekben a névmások nem szerepeltek, mert ez tényleg csökkentette a fordító által generált felesleges névmások számát.

Míg a MetaMorpho eredeti részfordítás-összerakó algoritmusa soha nem próbálja meg átrendezni a generált darabokat, a hibrid rendszerben kísérleteztünk különböző torzítási (pontosabban: szórend-átrendezési) beállításokkal, hiszen ez a lehetőség benne van a Mosesben. (Azért nem egészen „ingyenes” ez a szolgáltatás: az átrendezés megengedése drasztikusan növeli a dekódoláshoz – az optimális fordítás kiválasztásához – szükséges időt.) Azt találtuk, hogy ha nem adtunk büntetőpontokat a szórendi átrendezésekért a dekódernek, akkor határozottabban rosszabb minőségű fordításokat kaptunk. Az alapbeállításban szereplő torzítási büntetés (a torzítási büntetést és a nyelvmodell által adott pontszámot azonos súllyal vettük figyelembe), és megengedett maximális mozgatás (d=6, azaz 6 szón átívelő mozgatás megengedése) gyak-

[2] Lehetséges megoldások erre a problémára (amellett, hogy több memóriát teszünk a gépbe): alacsonyabb rendű nyelvmodell használata (ezzel persze a távolabbi függőségek ellenőrzését csökkentjük), az egyszeri előfordulások elhagyása és/vagy a nyelvmodell szótárának a leggyakoribb frázisokra korlátozása.

ran olyan fordításokat eredményezett, amelyekben a fordításként generált mondat végén teljesen elkeveredett fordításdarabok sorakoztak. A legjobb eredményt – a BLEU-pontszám tekintetében is – abban az összeállításban kaptuk, amelyikben az átrendezést teljesen megtiltottuk (d=0), annak ellenére, hogy ez sokszor szórendileg szerencsétlenebb fordítást eredményezett, különösen a magyar–angol fordítási irányban, ha a fordítandó magyar mondatnak a végén állt az ige. Az átrendezés letiltása a dekódolási időt is tizedére csökkentette.

Az alábbi mondat esetében látható egyrészt a feleslegesen generált névmások elhagyásának kedvező hatása, másrészt itt a hibrid fordító egyébként is sokkal érthetőbb fordítást generált, annak ellenére, hogy az egyik ige nem a megfelelő helyre került a fordításban:

Bemenet: *„Az anomáliáért a sötét anyag lehet felelős, amely talán akár egészen a Föld közelében is megtalálható”, írja Adler.*

MMO: *The dark substance, which the Earth is entirely in his neighbourhood even possibly, may be responsible for the anomaly can be found, Adler writes it.*

MMO+Moses: *The dark substance may be responsible for the anomaly, that possibly even all near the Earth can be found, Adler writes.*

3.3 Eredmények

A kísérleti összeállításokat a 2009-es athéni EACL konferencia mellett rendezett *Fourth Workshop on Statistical Machine Translation*-re kiadott angol–magyar tesztkészleten teszteltük (Callison-Burch és munkatársai, 2009).

Legeredményesebbnek a következő kísérleti összeállítás bizonyult:

- a frázistáblát kiegészítettük olyan alternatív részfordításokkal is, amelyekből töröltük a beszúrt névmásokat,
- a Moses dekódert úgy paramétereztük, hogy ne rendezze át az összetevők sorrendjét,
- azokra a mondatokra, amelyekre a MetaMorpho teljes fordítást adott, nem használtuk a részfordításokat, hanem pusztán a teljes fordítások rangsorolására használtuk a dekódert.

Az utóbbi összeállítással mindkét fordítási irányban a puszta MetaMorphónál kissé jobb minőségű fordításokat sikerült elérni mind a BLEU-pontszám, mind a szubjektív emberi megítélés szempontjából, azonban a javulás mértéke elmaradt a várakozásainktól (BLEU: magyar–angol irányban 9,96 → 10,10; angol–magyar irányban 8,13 → 8,44). Az alábbi táblázatban összefoglaltuk az eredeti MetaMorpho rendszer és néhány hibrid összeállítás által generált fordítások BLEU-pontszámait:

1. táblázat: A fordítások és azok BLEU-pontszámai.

magyar–angol	
MetaMorpho	9.96
d=6, nincs átrendezési büntetés, teljes elemzésnél is lehet összerakás	9.62
d=6, van átrendezési büntetés, teljes elemzésnél nincs összerakás	9.70
d=0, nincs átrendezés, teljes elemzésnél nincs összerakás, névmástörlés	10.10
angol–magyar	
MetaMorpho	8.13
d=6, van átrendezési büntetés, teljes elemzésnél nincs összerakás	8.22
d=0, nincs átrendezés, teljes elemzésnél nincs összerakás	8.44

4 Morfémaalapú statisztikai fordítórendszer

A magyar–angol fordítási irányban kísérleteztünk egy további fordítórendszerrel is, amelyben a szabályalapú komponenst mellőzve, a statisztikai nyelvmodelleket algoritmikus morfológiai elemzővel és szófaji egyértelműsítővel előállított morfémaalapú reprezentáció felhasználásával állítottuk elő. Ebben a rendszerben szintén a Moses dekódert használtuk.

4.1 A rendszer felépítése

A tanítókorpusz magyar oldalát a *Humor* morfológiai elemzővel (Prószéky & Novák, 2005) elemeztük és tövesítettük, és a *Hunpos* szófaji egyértelműsítővel (Halácsy, Kornai & Oravecz, 2007) egyértelműsítettük. Az angol oldal egyértelműsítésére a *CRFTagger*-t (Phan, 2006) használtuk, és a *morpha* elemzővel tövesítettünk (Minnen, Carroll & Pearce, 2001). Az utóbbinak megfelelő *morphg* morfológiai generátorral állítottuk elő célnyelvi fordítások felszíni alakjait. Sajnos a *morpha* elemző nem különbözteti meg a létige nem harmadik személyű alakjait a harmadik személyűektől, ezért ezt a hibát javítanunk kellett ahhoz, hogy a kimeneten a létige helyes alakja generálódjon.

Rendszerünkben a tokenek nem szavak, hanem morfémák voltak. Az alábbi példa a tanítókorpusz egy mondatát mutatja a rendszerben használt morfémaalapú reprezentációban.

Magyar: *a[det] 137[szn] apró[mn] csillag[fn] [ela] álló[mn] spirál[fn] meg+[ik] duplázódik[ige] [me3] .[punct]*

Angol: *the_dt spiral_nn of_in 137_cd tiny_jj star_nn s_nns double_vb ed_vbd itself_prp ._.*

Megközelítésünket több tényező motiválta. Egyrészt a magyarban a szavaknak több ezer lehetséges toldalékolt alakja van, és nincs az a korpusz, amelyben példaként

szerepelne minden szó minden lehetséges alakja (vagy akár csak a leggyakoribbak). Ezért az adatorientált megközelítés lépten-nyomon abba a problémába ütközik, hogy hiányzik az éppen szükséges adat, ha a tokenek szóalakok. Másrészt rendszeresen kötött morfémák felelnek meg a magyarban angol funkciószavaknak (pl. elöljárószóknak, birtokos és egyéb névmásoknak). Emellett rendszeres morfémasorrendi különbségek is vannak: az angol prepozícióknak a magyarban megfelelő toldalékok, illetve névutók követik, és nem megelőzik a névszói csoportokat, ugyanez igaz a birtokos névmásokra (és a nekik megfelelő birtokos ragokra), illetve az alanyi névmásokra (amelyeknek a magyarban leggyakrabban csak az igei személyragok felelnek meg).

Ezek a tényezők elég súlyos problémákat okoznak már a statisztikai fordító betanításához használt tanítókorpuszban az egymásnak megfeleltethető szóalakok összepárosítását végző Giza++ számára is, illetve jelentősen csökkentik a szóalapú fordítórendszer általánosítóképességét. Azt reméltük, hogy morfémaalapú rendszerünk frappánsan megoldja ezeket a problémákat.

A frázistáblát az alapértelmezett *grow-diag-final* heurisztikával állítottuk elő a Giza++ szóösszerendelésekből, amelyet a tanítókorpusz morfémaalapú reprezentációjából állítottunk elő. Ebben a rendszerben használtunk lexikalizált átrendezési táblát, a torzítási paramétert az alapbeállításon hagytuk. A rendszerben 5-gramos célnyelvi nyelvmodellt használtunk (ebben az esetben ez öt morfémát, nem öt szót jelent). Sajnos ebben az esetben is csak korlátozott méretű korpuszból tudtunk nyelvmodellt építeni a tesztrendszer korlátozott memóriakapacitása miatt. A rendszer betanításához a *Hunglish* korpusz irodalmi és jogi részét használtuk, a tesztkorpusz azonos volt a hibrid rendszer esetében használttal.

A MERT paraméteroptimalizációs eljárást (Och, 2003) úgy futtattuk, hogy az a korpuszból kiválasztott hangolókészleten kapott morfémaalapú BLEU-pontszámot próbálta optimalizálni. Az optimalizálás több napig futott.

4.2 Eredmények

A rendszer tesztelésekor először a morfémaalapú BLEU-pontszámot optimalizáló MERT eljárás által javasolt paraméterbeállításokat használtuk. A célnyelvi angol szóalakokat a *morphg*-vel állítottuk elő a dekóder által előállított morfémaalapú fordításokból. Számítottunk rá, hogy a morfémaalapú rendszer új problémával szembesít majd minket: olyan helyekre fognak keveredni morfémák, ahol normális esetben nem fordulhatnának elő, és így nem tudunk majd értelmes szóalakot generálni az adott morfémasorozatból. Így is lett. Ezekben az esetekben egyszerűen kihagytuk a rossz helyre került morfémát, bár ez nyilván nem optimális megoldás.

Sajnos ez az összeállítás várakozásainkkal ellentétben nem produkált nagyon biztató eredményeket. A fenti összeállítás a detokenizált kimenetre 7,82-es BLEU-pontszámot adott. Mikor a dekódert újrafuttattuk egy korábbi félbeszakadt MERT folyamat során kapott paraméterekkel, kicsit jobb BLEU-pontszámot kaptunk: 7,95-öt. Ez is sokkal gyengébb volt, mint a MetaMorphóé, de a fordítás emberi megítélése szempontjából még ennél is jelentősebb mértékben elmaradt a minősége a szabályalapú fordítóétól. Nagyjából ugyanez mondható el a rendszer kimenetét szóalapú statisztikai rendszerek által magyar–angol irányban produkált fordításokkal összevet-

ve is: a BLEU-pontszámok különbsége ebben az esetben még nagyobb, és a szubjektív minőség is jelentősen rosszabb a szóalapú rendszerekhez viszonyítva is.

A Giza++ szóösszerendeléseket átnézve azt tapasztaltuk, hogy várakozásainkkal ellentétben a tanítókorpusz morfémákra bontása önmagában nem oldotta meg a szóösszerendelések minőségével kapcsolatos problémákat: az összerendelések még roszszabbak voltak, mint amiket a korpusz minden morfológiai feldolgozás nélküli változatára kaptunk. Ugyanakkor a morfémaalapú megközelítés mindazon hátrányai, amikre előre számítottunk: a nyelvmodellekben és a frázistáblában megragadott lokális függőségek csökkent távolsága annak következtében, hogy a bemenet ugyanakkora szakaszát több token fedi le, mint a szóalapú változatban; a rossz helyre keveredett morfémák stb. valóban bekövetkeztek.

5 Összefoglalás

Cikkünkben a magyar és angol nyelvpár tagjai közt fordító hibrid és morfémaalapú statisztikai kísérleti fordítórendszereinket mutattuk be. Sajnos átütő eredményekről nem számolhattunk be. Ugyan hibrid rendszerünk egyértelműen jobb fordításokat hozott létre, mint a tisztán szabályalapú MetaMorpho rendszer, a javulás mértéke elmaradt várakozásainktól. Morfémaalapú statisztikai fordítórendszerünk pedig egyértelműen nem váltotta be a hozzá fűzött reményeket.

Hivatkozások

1. Callison-Burch, Chris; Philipp Koehn, Christof Monz, Josh Schroeder: Findings of the 2009 Workshop on Statistical Machine Translation In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece (2009) 1–28
2. Halácsy, Péter; András Kornai, Csaba Oravecz: 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
3. Halácsy Péter, Kornai András, Németh László, Sass Bálint, Varga Dániel, Váradi Tamás, Vonyó Attila: A Hunglish korpusz és szótár. In: Csendes D., Alexin Z. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia 2005, Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (2005) 134–142
4. Koehn, Philipp; Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst: Moses: Open Source Toolkit for Statistical Machine Translation 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) 177–180
5. Minnen, Guido; John Carroll, Darren Pearce: Applied Morphological Processing of English, Natural Language Engineering, 7(3). (2001) 207–223

6. Novák, Attila; László Tihanyi, Gábor Prószéky: The MetaMorpho translation system. In: Proceedings of the Third Workshop on Statistical Machine Translation at ACL 2008, Columbus, Ohio, (2008) 111–114
7. Och, Franz Josef: Minimum Error Rate Training for Statistical Machine Translation. In: Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, (2003) 160-167
8. Phan, Xuan-Hieu: CRFTagger: CRF English POS Tagger. (2006) http://crftagger.sourceforge.net/
9. Prószéky, Gábor and Attila Novák: Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.): Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday, Gummerus Printing, Saarijärvi/CSLI Publications, Stanford. (2005) 116-125

webforditas.hu: egy internetes nyelvtechnológiai szolgáltatás tanulságai

Prószéky Gábor, Tihanyi László

MorphoLogic
1116 Budapest, Kardhegy u. 5.
{proszeky,tihanyi}@morphologic.hu

Kivonat: Ez az előadás az immár kilenc éve indult MetaMorpho gépi fordítási projekt ingyenes internetes fordítórendszerként való működtetésével foglalkozik, illetve egy szótári, helyesírási és több más szolgáltatással kiegészített nyelvtechnológiai portál, a *webforditas.hu* működtetésének gyakorlati tapasztalatait és az ebből levonható elméleti következtetéseket mutatja be.

1 A *webforditas.hu* felépítése

1.1 Az alapszolgáltatások: weblapfordító, szövegfordító és szótár

A MetaMorpho gépi fordító rendszer (Tihanyi 2003, 2004, 2005, 2006, 2007) kifejlesztését követően döntés született arról, hogy a fordítóprogram legfontosabb funkcióit nemcsak a Windows alatti személyiszámítógép-felhasználók, hanem a teljes internetes közösség számára is elérhetővé tesszük. Ehhez a kifejlesztett alaprendszert internetes használatra is alkalmassá kellett átalakítani: ennek a szolgáltatásnak az alapváltozata *webforditas.hu* néven 2006 végén indult el, az angol-magyar fordítóprogramra építve (Tihanyi, 2007).

Az ötszáz karakterben limitált tetszőleges, formázatlan felhasználói szöveg fordítására kialakított modul mellett a MetaMorpho egy weblapfordító szolgáltatásnak is a háttérmotorját adja. Ez utóbbi arra van hivatva, hogy tetszőleges angol vagy magyar nyelvű weboldal teljes formázását megtartó formában forduljon le a másik nyelvre, azaz a képek elhelyezésétől kezdve a betűtípusok kiválasztásáig minden hűen tükrözi az eredeti weboldalt, csak a szöveg tartalma jelenik meg a másik nyelven. Ilyen szolgáltatás más nyelvpárok esetében létezett, ám üzemszerűen működő formában a magyar és bármilyen más nyelv között ez a megoldás volt az első.

Ahhoz, hogy a *webforditas.hu* valóban nyelvtechnológiai alapeszközök portáljaként működhessen, a korábban éveken át *www.mobidictionary.com* alatt működő internetes szótárszolgáltatást is ideemeltük, és egy másik „fül” alatt elérhetővé tettük a szótári rendszert is a fordítóprogram felhasználói számára. 2006 októberében tehát ezzel a három alapszolgáltatással – az angol és magyar nyelvekre működő weblapfordítóval, a szövegfordítóval és a szótárral – indult el a *webforditas.hu* portál.

1.2 További szolgáltatások: kereső, elemző, helyesírás-ellenőrző, felolvasó

2007 márciusában egy újabb szolgáltatással jelentkezett a *webforditas.hu*, és ez a keresés volt. Ez egy korábbi ITEM-pályázat részleges támogatásával megvalósított nyelvileg kiegészített internetes keresőmodul integrálásával történt. A rendszer nem a beírt karakterfüzért, hanem a keresőkifejezés tövére (vagy adott esetben: töveire) adott találatokat mutatja meg, és ezt akár szinonimák vagy idegen nyelvi alakok felajánlásával – és nem „vak" automatizmussal, hanem a felhasználó aktív közreműködésének igénybe vételével – a hagyományos keresésnél sokkal hatékonyabban képes megtenni. Mindezeket a funkciókat a Google által közzétett hívási felületen keresőmotoruk egyfajta kiegészítéseként jelentettük meg saját weboldalunkon, a *Kereső* fül alatt. Ezt a megoldást később kiegészítettük a kapott idegen nyelvű találatok keresés nyelvére való visszafordításának felajánlásával (1. ábra). Sajnálatos módon azonban, a Google keresőprogramban 2009 februárjában megjelent magyar fordítómodul hatással volt a *webforditas.hu* keresőfülének látogatószámára is. A továbbiakban mégis építünk erre a szolgáltatásra, ugyanis a keresőprogramot használóknak az a része, akik nem beszélnek idegen nyelvet, maguk még a keresőkérdést sem tudják megfogalmazni, nem hogy a találatot elolvasni. Ezért fontos, hogy a találatok egyszerű lefordítását célzó Google-megoldással szemben a *webforditas.hu* lehetővé tudja tenni a világ weblapjain megbúvó esetleges találatok magyar nyelvi elérését, illetve a külföldiek érdeklődésének felkeltését a magyar weblapokon található magyar nyelvű tartalomban való idegen nyelvű keresés irányába.

2007 májusában a MetaMorpho rendszer mondatelemzési technikáját illusztrálandó bevezettük az *Elemző* fület. Ezzel az igényesebb felhasználók a magyar és angol mondatok gép által „látott" nyelvi szerkezetének mibenlétéről is tudomás szerezhetnek. A belső struktúra egy némiképp leegyszerűsített formában, grafikusan megjelentett faszerkezeteken keresztül jelenik meg (2. ábra).

A *Helyesírás* fülre a fordítások bevitelénél nagy szükség mutatkozott. Az eredmény, hogy nem csak a fordításokhoz, hanem önálló alkalmazásként is naponta mintegy 1500-2000 felhasználó használja (3. ábra). A helyesírási rendszer természetesen elválasztási segítséget is tud adni, ami ugyan a fordításhoz nem igazán a legszükségesebb, de nagyban növeli egy nyelvi portál „komfortszintjét" (4. ábra).

2009 első felében bevezettük a bemenő szövegnek és fordításának *hangos felolvasását* is, elsőként angol és magyar nyelvű szövegekre. Az angol beszéd a kliens operációs rendszerének hanggenerátorát, a magyar a BME TMIT Profivox TTS rendszerét használja (Olaszy és mtsai, 2000). Meglepő módon, ez a szolgáltatás, mely a gépi fordítási feladathoz csak áttételesen kapcsolódik, rendkívül népszerű lett: látogatóinak havi átlagos száma meghaladja a szótárhasználókét, és így 2009-ben a szövegfordító után a *webforditas.hu* második legnépszerűbb szolgáltatásának számít.

A fordítás minőségét a portál működtetése során a felhasználók bevonásával kívánjuk javítani, így egy a *javaslatok* közzétételére szolgáló felület kialakítását is meg kellett oldani (5. ábra).

A Medián WebAudit szerint a mára elért 65-70 ezer napi látogatójával a *webforditas.hu* bent van az első 70 magyar weboldal között. Ez a gyakorlatban azt jelenti, hogy megelőzi az olyan népszerű tévé- és rádiócsatornák internetes oldalait, mint pl. *Magyar Televízió, ATV, TV2, HirTV, Sláger Rádió, Danubius Rádió, Magyar Rádió,* vagy akár a *Magyar Telekom.* Sok olyan neves újság weboldala is a

webforditas.hu mögött van látogatottságban, mint a *Bors Online,* a *168óra,* a *Magyar Nemzet Online,* vagy a *Népszava*. Sőt, a *webforditas.hu* megelőzi az olyan, széles körben használt szolgáltatásokat is, mint a *Magyar Elektronikus Könyvtár*, a *BKV*, az *Útvonalterv*, a *Vendégváró*, vagy a fiatalok közt igen népszerű *Zeneszöveg* vagy *Teveclub*.

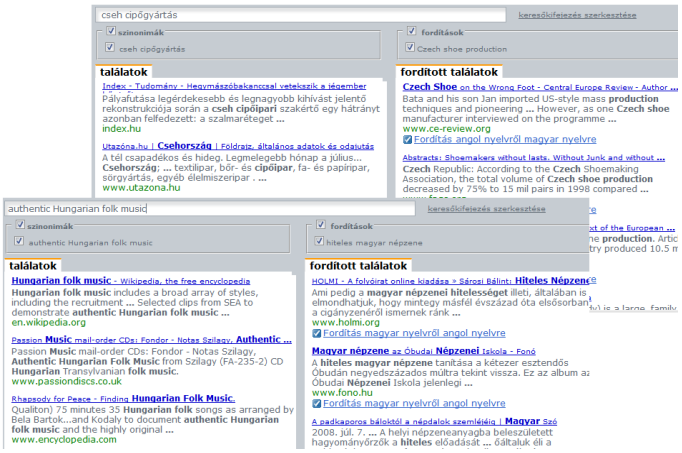
2 Nyelvpárok

A fordítóprogramok legfontosabb, általában egyetlen nyilvános jellemzője a nyelvpárok száma. A nyelvpárok számának belátható növelhetősége meghatározó szempont volt különféle fordítóprogram-technikák kialakításában. Mivel, mint Tihanyi (2007) írja, „nyelvi elszigeteltségünket az angol-magyar és magyar-angol változatok elkészítésével alapvetően feloldottuk”, olyan megoldások után kellett néznünk, amelyek kielégítik a további nyelvek bevonásával kapcsolatban felmerülő igényeket, ám az erre vonatkozó döntéseket az anyagi lehetőségek figyelembe vételével kellett meghozni.

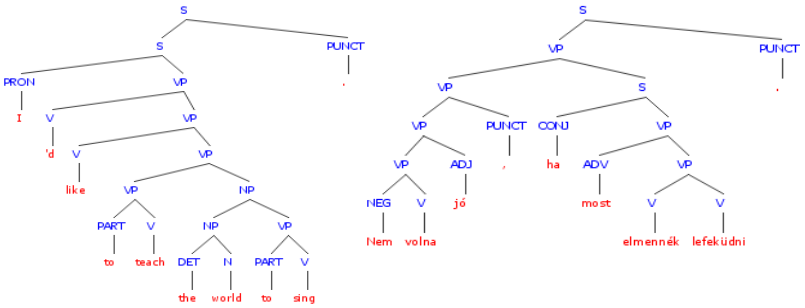
Hamar felmerült az igény az angol mellett további nyelveknek a fordítórendszerbe való esetleges bevonására, azonban ennek lehetőségeit a fejlesztések megtérülése határozza meg. Mivel egy újabb nyelvnek a MetaMorpho rendszerbe való bevonási költségei meglehetősen magasnak tűntek, más megoldást kellett választani. A lehetőségeket az utóbbi időben világszerte elszaporodó internetes nyelvi szolgáltatások sugallták. A fordítóprogramok világában a különféle nemzeti nyelvek fordítórendszerei közel 100%-ban elsőként az angol nyelvre készülnek el. Ezek jelentős része ingyenes webes szolgáltatásként el is érhető. A feladat tehát adott volt: minden X-angol/angol-X nyelvpár esetében ki kellett választani a legjobb minőséget adó fordítórendszert, és meg kellett keresni a technikai és üzleti lehetőséget a *webforditas.hu* angol-magyar/magyar-angol szolgáltatást biztosító MetaMorpho rendszerével való hatékony összekapcsolásra. Az egyes nyelvpárokhoz professzionális fordítók segítségével komoly tesztanyag készült Tihanyi László vezetésével, és az alapos kiértékelés után megindulhatott az újabb nyelvek legjobb fordítóprogramjainak bevonása a *webforditas.hu* rendszerbe: az angol mellett először a legfontosabb európai és világnyelvek, majd ezt fokozatosan kiterjesztve ma már gyakorlatilag minden fontosabb európai nyelv és világnyelv.

A többnyelvűségből adódóan újabb funkciók jelentek meg, mint például a különféle nyelvekhez automatikusan illeszkedő *virtuális billentyűzet*, melyet a szabad elérésű VirtualKeyboard program segítségével valósítottunk meg (6. ábra).

A többnyelvűségre való áttérés másik „hozadéka” a *nyelvfelismerő* modul volt. Ez különösen hasznos azoknak a nyelveknek az esetében, melyek az ezeket a nyelveket nem beszélők számára igen hasonlónak tűnhetnek. Sokszor nem könnyű eldönteni egy lefordítandó szövegről, hogy pl. dán vagy norvég-e, esetleg cseh-e vagy szlovák? A felismerő modul integrálása 2009-ben megtörtént, ám erre csak a fordítandó szövegek kis részénél van szükség, ugyanis a *webforditas.hu* bemenő szövegeinek igen jelentős része az öt európai világnyelv valamelyikén íródott, azon belül is elsősorban angolul (7. ábra). Érdekességként a 8. ábrán látható a további 45 nyelv 2009-es statisztikája.

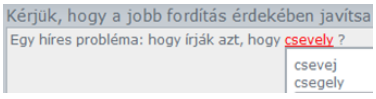


1.ábra

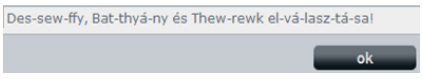


Angol: *I'd like to teach the world to sing. /* Magyar: *Nem volna jó, ha most elmennék lefeküdni.*

2.ábra



3.ábra



4.ábra

Javítások

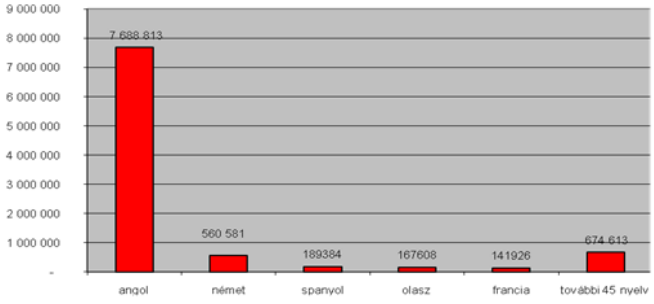
utolsó javaslatok

szűrés: nincs << < 1 2 3 4 5 6 .. > >> rendezés: legújabb elsőnek

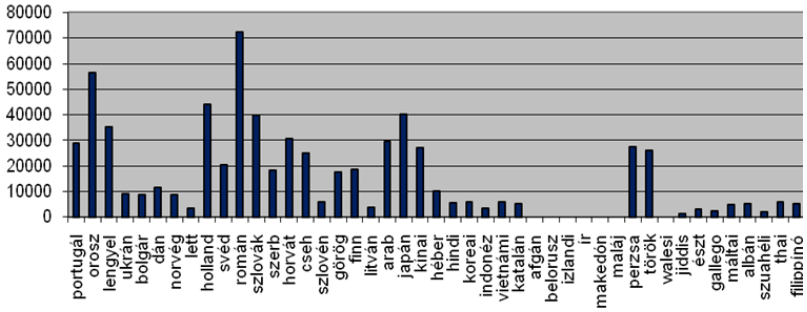
ID	nyelv	fordítandó szöveg	régi fordítás	új fordítás
9845	magyar-angol	elhagyni	to leaf	leave
8771	magyar-angol	Törődj a saját dolgoddal!	Care about your own thing.	Mind your own business.
8744	angol-magyar	masterbatch	masterbatch	mesterkeverék
8716	angol-magyar	showjumping rider	díjugratás lovas	díjugrató lovas
8715	angol-magyar	dressage rider	díjlovaglás lovas	díjlovas
8713	angol-magyar	Yuna! Don't talk like that!	Yuna! Ne beszélj úgy!	Yuna! Ne beszélj így!
8692	magyar-angol	Tisztelt .vki	Respected .sy	Dear .sy
8689	magyar-angol	kiszállási költség	getting out expense	fieldwork expense
8686	magyar-angol	tündecsávó	tündecsávó	elfguy
8685	angol-magyar	paladin	lovag	paplovag

5.ábra

6.ábra



7.ábra



8.ábra

9.ábra

3 Számok, eredmények, tapasztalatok

A *webforditas.hu* éves látogatottsági számai évről évre egyre nőnek, és 2009-ben a portálnak már az első 10 hónap után jóval több látogatója volt, mint 2008-ban, és több mint háromszor annyi, mint 2007-ben. A *webforditas.hu* eddigi látogatóinak összlétszáma meghaladja Magyarország lakosainak számát, ami természetesen nem jelenti azt, hogy minden magyar állampolgár ténylegesen járt volna már az oldalon, hanem sokkal inkább azt jelenti, hogy egyre gyakrabban térnek vissza a felhasználók. Valóban, a visszatérő felhasználók száma egyre nő, és 2009-ben már éves szinten hét százalék körül van (1. táblázat). Érdemes itt megemlíteni, hogy jelenleg az egy hónapon belül visszatérők száma az összes látogatókhoz viszonyítva 30%, az egy héten belülieké 55%, az egy napon belül visszatérők száma 70 % körül van.

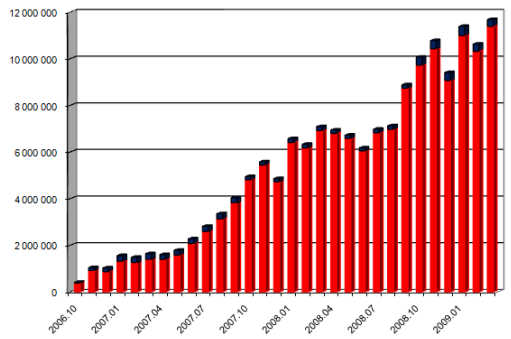
2008 októberétől, azaz amióta elindult a magyarról és magyarra való fordítás az angoltól eltérő nyelvekkel is, a *gemius.hu* szerint hirtelen 600.000 fölé emelkedett a havi látogatószám. Ezt még a Google 2009 februárjában történt bejelentése sem tudta lenyomni, miszerint a Google is elindította a magyar és más nyelvek közötti fordítórendszerét. Sőt, márciusig ez még az érdeklődést is növelte a *webforditas.hu* iránt, hiszen sok cikk megemlítette, hogy létezik ez a fordítási portál is, amit talán e nélkül a bejelentés nélkül kevesebben tudtak volna meg. Néhány hónap stagnálást követően, 2009 augusztusától a látogatószám ismét emelkedni kezdett, és közelíti a havi egymilliót (9. ábra). Ugyanezek a számok egy másik auditrendszer, a *webaudit.hu* számai alapján némiképp alacsonyabbak, ám az mindkét kimutatás alapján figyelemre méltó, hogy a nyitólap és a szövegfordítás szolgáltatás adatai nagyjából azonosak, míg a weblapfordítás átlagos látogatószáma 2008 után visszaesett. Ez egyértelműen a Google korábban említett magyar nyelvi fordítószolgáltatásának megjelenésével magyarázható. Pontosabban: nem pusztán a szolgáltatás megjelenése, hanem annak elérési módja adja a teljes magyarázatot. A Google weblapfordító szolgáltatása azonnal ott található a keresés eredményeként kapott találati lista minden eleménél, míg a *webforditas.hu* oldalra oda kell mennie a felhasználónak. Ezzel szemben a szövegfordítás a Google esetében sem automatikus, hiszen ez a szolgáltatás csak a találati oldaltól eltérő *translate.google.com* oldal fellapozásával válik elérhetővé. Ha mindehhez hozzávesszük, hogy a *webforditas.hu* oldalon a korábban ismertetett kiegészítő szol-

gáltatások kényelmesebb fordítási környezetet adnak, így a látogatók száma itt folyamatosan tovább tudott növekedni. A szöveg- és weblapfordítás gyakorlatilag változatlan egymáshoz viszonyított arányát mutatja – az egyre növekvő lekérdezés-szám mellett is – a 10. ábrán látható grafikon.

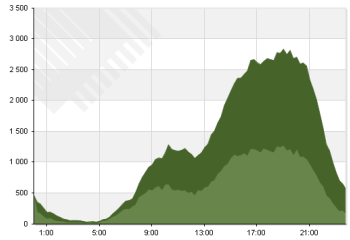
A *webforditas.hu* szövegfordítójának napi átlagos időbeli eloszlása a 11. ábrán látható. Jól látható, hogy a felhasználók igazán este 5 és 9 közt aktívak, és az is látszik, hogy javarészt déli 1 óra körül ebédelnek. Hasonló jellegű kimutatást készítettünk a szótárfelhasználók esetében is (12. ábra). A hozzávetőleges hasonlóság ellenére a két grafikon közti eltérések hamar látszanak: az egyik, hogy a délelőtti szótárhasználat – a szövegfordító-használattal szemben – összemérhető a délutánival; a másik, lényegesebb eltérés a függőleges skáláról olvasható le: a szótármodul felhasználói – sajnálatos módon – nagyságrenddel kevesebben vannak, mint a fordítóprogramé. Ráadásul a *webforditas.hu* szótári szolgáltatásának heti látogatószáma 2009-ben önmagában is visszaesést mutat (13. ábra).

1.táblázat

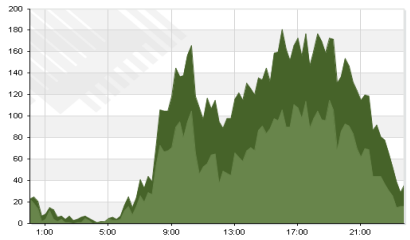
Év	Látogatások	Oldal-letöltések	Látogatók (cookie)	Új		Visszatérő	
				#	%	#	%
2006	642 376	4 877 357	201 313	201 313	100,0	0	0,0
2007	7 263 881	52 755 643	1 712 322	1 668 165	97,4	44 157	2,6
2008	19 010 954	126 714 017	3 997 964	3 831 963	95,8	166 001	4,2
2009 (1-10)	23 456 439	147 660 970	5 305 375	4 934 205	93,0	371 170	7,0
Σ	43 373 650	328 007 987	**11 216 974**	10 635 646		581 328	



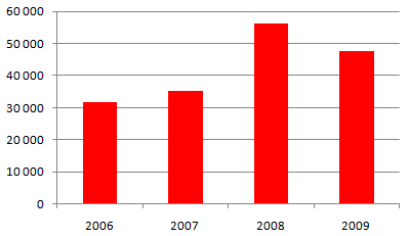
10.ábra



11.ábra



12.ábra

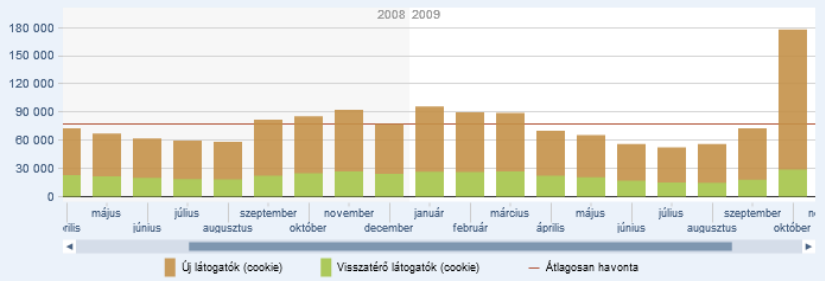


13.ábra

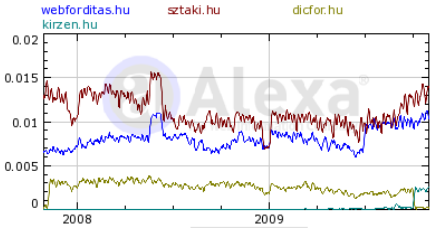
Ezek az alacsony számok – amint elemzéseink kimutatták – több különböző okra vezethetők vissza. Az egyik, hogy a szótárakat a tipikus magyar internethasználó általában a kimondottan erre szolgáló *sztaki.hu* és *dicfor.hu* (illetve ez utóbbi helyett ma már a *kirzen.hu*) oldalakon keresi. Ezeknek ugyan a teljes *webforditas.hu* oldalhoz viszonyított napi elérési statisztikája alacsonyabb – a *sztaki.hu* kivételével, de ott is jelentősen csökkent a különbség az idők folyamán (15. ábra). A felsorolt szolgáltatásokat tehát „dedikált" szótárlapoknak tekinti a tipikus felhasználó, míg a *webforditas.hu* oldalt elsősorban a fordítóprogram miatt használja, ezért itt, ha szótáraznia kell, marad a fordítóprogram ablakában. Ez a második ok, ami csak a fordítóprogram log-fájljainak elemzésekor vált világossá: a felhasználók jelentős része vagy lusta átmenni a szótárfülre, vagy nem érti pontosan a fordítóprogram és a szótárprogram közti különbséget, ezért egy-egy szót ír be a fordítóprogram ablakába, amire a fordítórendszer természetesen megadja az általa legszerencsésebbnek gondolt fordítást. A szótártól való legnagyobb eltérés tehát itt az, hogy nincs mód a „kevésbé jó fordítás" átadására, azaz egyetlen találattal kell beérnie a felhasználónak, míg a szótárfülön több lehetséges értelmezés is megjelenik, ahogy ez a szótáraknál szokás, ráadásul lexikográfiai szempontból is rendezettebb formában. Például a *dog* szóra a fordítóprogram eddig csak annyit mondott, hogy *kutya* – szemben a szótárral, mely főnévi értelemben is hat találatot ad, az *eb*-től a *vaskapocs*-ig. A szótárfül segítségével látható továbbá az igei *nyomon követ* is, valamint a *dog* mintegy ötven kifejezésbeli előfordulása is elérhető.A probléma technikai megoldása tehát az lett, hogy amennyiben a fordítóprogram bemenetén szótári kérdésnek látszó – javarészt egyetlen szóból álló – bemenet jelenik meg, a rendszer azonnal a saját szótári szolgáltatását kínálja fel. Ezen felül még a szótári szolgáltatás sebességét és más minőségi javításokat is bevetettünk, így 2009 októberétől a *webforditas.hu* havi szótárlátogatóinak

száma egyetlen hónap alatt megháromszorozódott (14. ábra). Ez az arány láthatóan tovább javul, mert a változást az említett javítások okozták, méghozzá úgy, hogy a látogatószám egyik napról a másikra a tízszeresére emelkedett, és az azóta eltelt időben nem változott.

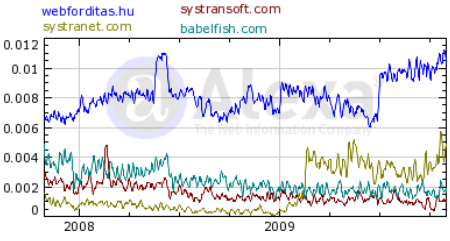
Természetesen egy-egy összehasonlítás „sikere” önmagában nem érték, hiszen nem mindegy, hogy szolgáltatásunkat mivel hasonlítjuk össze. Érdekes viszont, ha megnézzük, hogy viszonyulnak a *webforditas.hu* látogatottsági adatai a hosszú időn át legnépszerűbb fordítórendszer, a Systran – *systransoft. com, systranet.com, babelfish.com* nevű – internetes szolgáltatásaihoz, akkor látjuk, hogy a Systran visszaesése e piacon szembetűnő, hiszen a csak magyar nyelvre specializálódott *webforditas.hu* portált is többen látogatják (16. ábra).



14.ábra



15.ábra



16.ábra

Ennek a visszaesésnek természetesen nem a *webforditas.hu* az oka, hanem a Google internetes nyelvi szolgáltatásainak előretörése, mely a világnyelvek esetében sokkal nagyobb veszteséget okozott a korábban ezzel foglalkozóknak, mint a magyar esetében a *webforditas.hu* oldalnak, legalábbis egyelőre.

Annak az elemzésével is érdemes foglalkozni, hogy elsősorban milyen típusú szövegeket fordítanak a felhasználók a *webforditas.hu* segítségével? Érdemes azért az alábbi, teljességre nem törekvő felsorolásra egy pillantást vetni: *en.wikipedia.org, www.fanfiction.net, edition.cnn.com, www.download.com, servedby.advertising.com www.viamichelin.com, www.wowhead.com, www.myspace.com, www.youtube.com, www.cnet.com, www.amazon.com, www.bbc.co.uk, ad.doubleclick.net, www.fifa.com, i.thottbot.com* stb. Ennek a listának a segítségével a fordítandó szövegek egy meghatározó részének tematikájáról is hamar képet alkothatunk.

Egy másik érdekes kérdés, hogy hogyan jutnak a *webforditas.hu* oldalra a felhasználók. Természetesen, ha már ismerik az oldalt, csak rákattintanak a könyvjelzőre, de ha még nem, akkor mit írnak be keresőjükbe, hogy magyarra vagy magyarról fordítást kaphassanak? A 2. táblázat a leggyakoribb ilyen keresőszavakat mutatja. A második oszlop a találatok Google által becsült számát, a harmadik pedig a *webforditas.hu* oldalnak ebben a Google által visszaadott találati listában elfoglalt pozícióját mutatja.

2. táblázat

fordítás	4 030 000	1
forditás	2 680 000	2
angol fordítás	924 000	1
online szótár	1 600 000	3
szótár	6 880 000	3
forditas	2 690 000	1
online	2 950 000 000	102
fordító program	420 000	1
angol	7 950 000	280
fordito	88 200	3
angol fordító	1 040 000	2
fordítás magyarról angolra	396 000	2
angol magyar	2 620 000	3
angol magyar fordító	462 000	2
magyar fordítás	1 670 000	1
angol szótár	361 000	6
fordító	613 000	3
fordítóprogram	139 000	2

Ezek az adatok elég biztatóak arra nézve, hogy 2009-ben a felhasználó meglehetősen nagy biztonsággal megtalálják a *webforditas.hu* weboldalt, ha erre van szükségük.

4 Összefoglalás és továbblépés

Cikkünkben a MetaMorpho gépi fordítási projekt ingyenes internetes fordítórendszerként való működtetésével foglalkoztunk, és igyekeztünk bemutatni a *webforditas.hu* működtetésének gyakorlati tapasztalatait. Az ezekből levonható következtetéseket arra igyekeztünk felhasználni, hogy hogyan lehetne a meglevő nyelvtechnológiai tudásbázist a *webforditas.hu* rendszerbe még jobban integrálni, és hogy mi módon lehetne a technikai és anyagi lehetőségek ismeretében továbbfejleszteni a meglevő rendszert. Egy ilyen lehetőség a bemutatott know-how-nak, illetve a gyakorlati tapasztalatoknak a beépítése egy hasonló, de már nem pusztán a magyar nyelvre, hanem az EU többi nyelvére hasonló elveken épülő rendszerbe. Egy ezt célzó projektjavaslat kidolgozásra is került, és az elkövetkező években az Európai Bizottság által elfogadott és *iTranslate4* néven magyar vezetéssel és az európai nyelvtechnológiai piac legfontosabb szereplői közreműködésével 2010-ben induló ICT-PSP pályázat keretében erre mód nyílik.

Hivatkozások

1. Olaszy G., Németh G., Olaszi P., Kiss G. Zainkó Cs., Gordos G: Profivox - a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology, *V*ol. 3-4 (2000) 201-215
2. Tihanyi László: A MetaMorpho projekt története. Alexin Zoltán; Csendes Dóra (szerk.) Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai, SZTE, Szeged (2003) 247-253
3. Tihanyi László: A MetaMorpho projekt 2004-ben. Alexin Zoltán; Csendes Dóra (szerk.) A 2. Magyar Számítógépes Nyelvészeti Konferencia előadásai, SZTE, Szeged (2004) 85-87
4. Tihanyi László: A MetaMorpho fordítóprogram projekt 2005-ben. Alexin Zoltán; Csendes Dóra (szerk.) A 3. Magyar Számítógépes Nyelvészeti Konferencia előadásai, SZTE, Szeged (2005) 99-107
5. Tihanyi László, Merényi Csaba: A MetaMorpho fordítóprogram projekt 2006-ban. Alexin Zoltán; Csendes Dóra (szerk.) A 4. Magyar Számítógépes Nyelvészeti Konferencia előadásai, SZTE, Szeged (2006)
6. Tihanyi László: A MetaMorpho projekt 2007-ben – a sorozat vége. Tanács Attila; Csendes Dóra (szerk.) Az 5. Magyar Számítógépes Nyelvészeti Konferencia előadásai, SZTE, Szeged (2007) 179-186

II. Szövegbányászat

Információkivonatolás szabad szövegekből szabályalapú és gépi tanulásos módszerekkel

Miháltz Márton[1], Schönhofen Péter[2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
H-1083 Budapest, Práter utca 50/a
mmihaltz@gmail.com
[2] in4 Kft
1011 Budapest, Bem rakpart 26. III/2.
schonhofen@gmail.com

Kivonat: Bemutatunk háromféle megközelítést egy információkivonatoló rendszerre, melynek célja doménfüggő szöveges információk kinyerése nagy tételben angol nyelvű Wikipédia-szócikkekből. Az első megközelítés mély nyelvi elemzést és manuálisan létrehozott információkinyerő mintákat használ. Ennek kiterjesztése egy olyan módszer, mely képes annotált példamondatok segítségével ilyen mintákat automatikusan megtanulni. A harmadik módszer csupán szófaji egyértelműsítésre támaszkodik és felügyelt gépi tanulást alkalmaz. Mindhárom módszer esetében bemutatjuk azok kiértékelését és összehasonlítását, két különböző doménen (tanulmányi adatok, díjak elnyerése.)

1 Bevezetés

Szeretnénk bemutatni egy saját fejlesztésű információkivonatoló rendszert nagy mennyiségű, megbízható szöveges információ kinyerésére angol szövegekből, mely az iGlue projekt [1] – melynek célja személyek, földrajzi helyek, intézmények stb. adatainak egységesen kezelt, szemantikusan összekapcsolt adattárba gyűjtése – számára készült.

A rendszer bemeneti szövegállománya jelenleg a Wikipédia nyílt tartalmú webes enciklopédia [2] angol nyelvű szócikkeinek halmaza. A fejlesztéshez használt első domén a tanulmányi adatok területe volt. Az egyes személyekhez az alábbi attribútumokat szerettük volna kinyerni a róluk szóló Wikipédia-szócikkek szöveges részéből: oktatási intézmény neve, ahol a személy tanulmányokat folytatott; tanulmányok kezdete és vége (dátumok); fokozatszerzés dátuma; elért tudományos fokozat; tanulmányterület. Például:

In 1977*, he graduated magna cum laude from* Harvard University *with a* B.A. *in* mathematics *and* economics*.*

Intézmény neve: Harvard University
Tanulmányok kezdete: -
Tanulmányok vége: -

Fokozatszerzés dátuma: 1977
Elért fokozat: B.A.
Tanumányterület(ek): mathematics; economics

Az információkivonatoló rendszer működése mély nyelvi elemzésen és az ezeken definiált minták, valamint az egyes attribútumok névelemtípusainak felismerésén alapul. A kivonatoló minták az igei vonzatkereteken (tagmondatok főigéje és annak vonzatai, szabad határozói) alapulnak. A mintakészlet előállítására mind teljesen manuális, mind félig automatikus módszerekkel is kísérletet tettünk. Emellett bemutatunk egy kísérletet a feladat megoldására felügyelt gépi tanulással is.

A cikk következő részében ismertetjük az elsőként alkalmazott nyelvi elemzés és névelem-felismerés fontosabb részleteit, a felmerült problémákra adott megoldásainkat, valamint egy teljesen manuálisan létrehozott mintákkal működő rendszer kiértékelésének eredményeit. A 3. részben ismertetjük a felügyelt gépi tanulásos megközelítést és összevetjük a mintafelismerésen alapuló módszerrel. Végül a 4. részben bemutatunk egy kísérletet a mintaalapú megközelítés részben automatikussá tételére.

2 Mintaalapú információkivonatolás

2.1 Korpuszépítés

Az információkivonatolás forrásául a Wikipédia-szócikkeit választottuk, mivel ezek nagy mennyiségben állnak rendelkezésre, viszonylag egységes, géppel jól feldolgozható enciklopédikus stílust követnek, valamint a nyílt közösségi fejlesztői megközelítés miatt tartalmilag elfogadható pontosság jellemzi őket.

A korpusz alapja a statikus Wikipédia dump [4] 2008 júniusi verziója volt, mely összesen mintegy 2.4 millió szócikket tartalmaz. Ezek között heurisztikákkal korábban sikerült beazonosítani 90.000, nagy valószínűséggel személyekről szóló szócikket, ez képezte a feldolgozás bemenetét. A HTML-oldalak szöveges tartalmát (nyers szöveget tartalmazó bekezdések) elkülönítettük a formázástól, külön megtartva olyan metainformációkat, mint az oldal címe és különböző címváltozatai (egy adott oldalra utaló átirányító oldalak (redirection page) követésével), kategóriacímkéi, a szövegben lévő hiperlinkek stb.

2.2 Nyelvi elemzés

A nyers szöveget a LingPipe mondatszegmentáló eszközével [3] bontottuk mondatokra, majd ezt követte a nyelvi elemzés az *Enju parser* szintaktikai elemzővel [5]. Az Enju egy gyors, valószínűségi HPSG-nyelvtannal működő angol parser, mely képes predikátum-argumentum szerkezetek és frázisstruktúrák azonosítására. A következő lépésben az Enju által létrehozott elemzési szerkezetek eredményeiben azonosítottuk az igei szerkezeteket, majd az utolsó lépésben ezeken működött az esemé-

nyeken (igei szerkezeteken) alapuló információkinyerő modul. Az alábbiakban az utóbbi két modulról lesz bővebben szó.

Az Enju parser kimenete a frázisstruktúra-viszonyokat XML-hierarchiában, míg a predikátum-argumentum viszonyokat és egyéb jellemzőket (pl. morfológiai információk, aspektus, igenem stb.) jegyszerkezetek formájában adja meg az egyes mondatokra.

A feldolgozás során először azonosítottuk a mondatot alkotó VP-k közül azokat, melyek az információkivonatolás számára releváns információkat tartalmaznak (mellérendelt tagmondatok, vonatkozó mellékmondatok, bizonyos határozói mellékmondatok (pl. „miután", „mielőtt")). A tagadott, vagy nem állító módban álló főigéjű VP-ket kihagytuk.

A következő lépésben az egyes VP-ket alkotó összetevőket azonosítottuk: főige (és partikulája), alany, direkt tárgy és indirekt tárgy, valamint a vonzat vagy módosítói szerepet betöltő prepozíciós frázisok.

Az NP-kben csak a fejjel bezárólag vettük figyelembe a tokeneket, illetve a fej után következő appozíciókat és birtokos szerkezeteket. Az NP-k elejéről elhagytuk a determinánsokat, birtokos névmásokat, prepozíciókat stb.

Ha a főige vonzata igei volt, akkor a beágyazott igét és annak vonzatait/határozóit is azonosítottuk.

Minden összetevőben azonosítottuk az azt alkotó tokenek felszíni alakját, lexikai alakját, szófajkódját, valamint mondatbeli pozícióját.

A koordinált összetevőket szétbontottuk és előállítottuk a többi összetevővel való összes kombinációjukat.

Példa:

Input mondat:

After receiving a Bachelor's Degree in mathematics and physics at the University of Michigan, he went on to obtain a Ph.D. in electrical engineering at Harvard in 1998.

Output elemzési szerkezetek (egyszerűsített):

((Verb, "receive"), (Subj, "he"), (Obj, "Bachelor's Degree"), (PP-in, "mathematics"))
((Verb, "receive"), (Subj, "he"), (Obj, "Bachelor's Degree"), (PP-in, "physics"))
((Verb, "go on"), (Subj, „he"), (Verb2, „obtain"), (Obj2, "Ph.D."), (PP-in2, "electrical engineering"), (PP-at2, "Harvard"), (PP-in2, "1998"))

2.3 Névelem-felismerés

Az információkinyerő minták a nyelvi elemzésben azonosított igei szerkezetekre, mint eseményekre, illetve ezek összetevőire, mint „szereplőkre" alapulnak. Az egyes mintákban az eseménykeret különböző szereplőire (oktatási intézmény, elért tudományos fokozat, tanulmányterület, végzés dátuma stb.) a szintaktikai tulajdonságokon felül szemantikai megszorításokat is tettünk. Így például egy lehetséges szabály az alábbi mintának felelhet meg:

```
Subj(PERSON) + V('attain') + Obj(DEGREE) + PP-in(SCHOOL)
+ PP-in(DATE)
```

Vagyis megköveteljük, hogy a VP feje az „attain” ige legyen, az alanyi szerepű igevonzat SZEMÉLY típusú névelem legyen, a tárgy egy TUDOMÁNYOS FOKOZAT típusú NP stb.

A szemantikai megszorítások (névelemtípusok) ellenőrzésére reguláris kifejezéseket és/vagy lokális lexikonokat használtunk fel. A lexikonok minél kimerítőbb összeállításához számos online információforrást és weboldal anyagát felhasználtuk (WordNet, Wikipédia, CrunchBase, univ.cc stb.) Így pl. a lehetséges tanulmányterületek listája mintegy 2.100, az oktatási intézmények listája 34 ezer tételt tartalmazott.

2.4 Mintaillesztés

Az információkinyerő modullal csak azokat az igei szerkezeteket dolgoztuk fel, amelyekben valamelyik meghatározott igevonzat/módosító azonos volt vagy a címszóban megjelenő személynévvel, annak valamilyen névváltozatával, vagy egy (hím- vagy nőnemű) személyes névmás volt, ezzel valószínűsítve azt, hogy a kinyert információ a kérdéses személyre vonatkozik.

A kérdéses eseményszereplőket a főigétől függően kb. 20 összetett szabály (minta) azonosította. A minták hivatkoznak a nyelvi elemzés által azonosított összetevőkre, valamint használják a felismerhető szemantikai kategóriákat (névelemtípusokat).

A minták fejlesztéséhez és folyamatos, iteratív validációjához készítettünk egy fejlesztői korpuszt, melyben humán annotátorok 200 db, véletlenszerűen kiválasztott személy Wikipédia-szócikkében azonosították a releváns tanulmányi attribútumokat. A minták és a mintafelismerés fejlesztéséhez ezen a halmazon végeztünk folyamatosan pontosság- és fedésméréseket, illetve elemeztük a negatív találatokat.

2.5 Problémák

A munka során számos olyan probléma merült fel, melynek során az Enju parser hibás elemzésein kellet korrekciót végezni.

Az első problémát a prepozíciós frázisok illesztési problémája jelentette (PP-attachment problem), a parser ugyanis inkonzekvens módon ugyanolyan típusú PP-ket különböző esetekben különböző összetevőkhöz kapcsolt. Emiatt a VP-kben a PP-ket rendezetlen listaként kezeltük, és speciális szabályokkal vettük őket figyelembe. Így például az időhatározókat (dátum típusú NP-k 'in' vagy 'on' prepozícióval) a mondatbeli pozíciójukat figyelembe vevő szabályokkal azonosítottuk.

Egy másik, igen gyakori problémát a névelemek (named entityk) határainak hibás felismerése okozta. Ennek orvosolására igyekeztünk minél több névelemet az elemzés előtt, a szegmentált nyers szövegen felismertetni és speciális karakterekkel egyetlen input tokenné összevonni, hogy a parser ezután egyetlen (főnévi) entitásként kezelje őket. A névelemek elő-felismerésének legegyszerűbb eszköze az eredeti szövegben nagy kezdőbetűket tartalmazó, wikipédiás hiperlinkkel ellátott szövegdarabok (anchor textek) azonosítása volt, mivel ezek nagy valószínűséggel tulajdonnévi enti-

tásoknak felelnek meg. Szintén felismertük és összevontuk azokat a többszavas névkifejezéseket, melyek többszavas, nagy kezdőbetűs tokeneket tartalmazó Wikipédia-oldal-címekkel voltak azonosak.

Hasonló probléma volt, hogy az elemző koordinációként értelmezett bizonyos, vesszőt tartalmazó névelemtípusokat, például dátumokat, vagy az angolban gyakori intézménynév-vessző-földrajzi összetételű tulajdonneveket (pl. University of California, Berkeley.) Az előbbiek felismerésére reguláris kifejezéseket, az utóbbiakhoz reguláris kifejezéseket és névlistákat használtunk (34 ezer oktatási intézménynév, 2,3 millió földrajzi név).

Egy további, gyakori problémát a többelemű NP-felsorolások hibás, néha koordinációként, néha appozícióként való elemzése jelentett, ezt az Enju kimenetének feldolgozása során külön szabályokkal kellett korrigálni.

2.6 Kiértékelés

A rendszer kiértékeléséhez az annotátorokkal készíttettünk egy újabb, 100 szócikkből álló annotált kiértékelő halmazt. Ezen a mintán kiszámítottuk a tanulmányok doménen működő, kézzel fejlesztett mintákon alapuló információkivonatoló rendszer pontosságát és fedését a kinyert attribútumokra nézve. Pontosságon a rendszer által helyesen megadott értékek és a rendszer által megadott értékek arányát, fedésen a rendszer által helyesen megadott és a referenciaértékek arányát értjük. A pontosság 94,22%, a fedés 60,33% volt ezen a mintán (F-mérték = 73,55%)

3 Információkivonatolás felügyelt gépi tanulással

A tanulmányok domén esetében a rendszer teljesítményének növelésére kísérletet tettünk a szabályalapú megközelítés ötvözésére felügyelt gépi tanulásos módszerekkel. A tanításhoz a Wikipédia-kategóriacímkék felhasználásával, valamint kézi annotációval generáltunk mintegy 200 tanítópéldát, azonban a szabályalapú módszerhez képest csak kevesebb attribútumot tudtunk azonosítani (intézmény neve, tudományos fokozat, fokozatszerzés dátuma).

A példákat csupán mondatszegmentálásnak, tokenizálásnak és szófaji egyértelműsítésnek vetettük alá. A tanulóalgoritmus a maximum entropy módszert használta [6], a felhasznált feature-ök a kérdéses elemet megelőző és az azt követő n-gramok (n=1,2,3 és n=1,2), illetve az azt megelőző legközelebbi ige töve voltak.

A kiértékelő halmaz segítségével elvégeztük a 3 attribútum gépi tanulásos módszerrel történő felismerésének külön-külön kiértékelését (pontosság és fedés), majd egyenként megvizsgáltuk, hogy a szabályalapú módszer kimenetének metszetével (pontosság várható növekedése) vagy uniójával (fedés várható növekedése) érünk-e el jobb eredményeket (1. táblázat.) A legjobb eredményeket az intézménynév és a tudományos fokozat attribútumok esetében, a két módszer eredményeinek uniójával kaptuk. Az intézményneveknél a szabályalapú módszerhez képest a kombinált módszer a fedésen 9,15%-os növekedést %), a pontosságon 1.25%-os csökkenést (91.01%) eredményezett, míg a tudományos fokozatoknál a fedés 18,28%-os növe-

kedést (80,88%), a pontosság 0,24%-os csökkenést (94,01%) mutatott. A hibrid módszerrel így sikerült a teljes rendszer fedését szignifikánsan növelni, miközben a pontosságot is sikerült a kritikusnak ítélt 90%-os küszöb felett tartani.

1. táblázat: A szabályalapú és a gépi tanulásos módszerek, valamint ezek uniójának és metszetének pontossága és fedése az egyes tanulmányi attribútumok felismerésében.

	Intézménynév		Fokozatszerzés dátuma		Tudományos fokozat	
	P	R	P	R	P	R
Szabály-alapú	92,25%	67,29%	100%	54,69%	94,25%	62,60%
Gépi tanulásos	90,81%	40,63%	84,51%	46,88%	91,93%	43,51%
Unió	91,01%	**76,44%**	89,71%	75%	94,01%	**80,88%**
Metszet	100%	10,50%	100%	4,69%	100%	2,29%

4 Mintaáltalánosítás

Az információkinyerő rendszer fejlesztésének következő szakaszában kísérletet tettünk egy olyan változat kifejlesztésére, mely képes annotált példamondatokból jórészt automatikus módon, információkinyerő mintákat önállóan tanulni. A cél egy olyan általános metódus kifejlesztése volt, mely annotált példákból kiindulva, a szükséges humán munkaerő-ráfordítást minimalizálva adaptálható egy-egy újabb IE-doménre akár egyetlen munkanap alatt is. A humán annotátor feladata csupán a rendszer által megtanult minták ellenőrzése, kiegészítése, illetve az esetlegesen előforduló negatív minták felismerése és megjelölése lenne.

4.1 Tanítópéldák előállítása

Annotált tanítópéldák előállításához felhasználtuk a Yago projekt [7] eredményeit, mely a teljes angol nyelvű Wikipédia-szócikkállomány strukturáltan rendelkezésre álló (tehát nem a szabad szöveges részekbe eső, hanem a keretes részekbe (info boxok) tartozó, kategóriacímkékben megjelenő) információit dolgozta fel és szervezte szemantikai hálózatba.

A Yago tudásanyagának egy része 2-argumentumú relációk formájában áll rendelkezésre. A relációkban álló párok a Wikipédia-szócikkekben jellemzett entitások (pl. személyek, intézmények stb.) Az entitások mind WordNet-synsetek, mind Wikipédia-kategóriaosztályok alá vannak rendelve. Feltételezve, hogy bizonyos redundancia várható a Wikipédia-szócikkek strukturált és strukturálatlan részei között, az entitások neveit a Wikipédia-szócikkek szövegében visszakeresve automatikusan előállíthatunk annotált tanítópéldákat egy-egy Yago-relációhoz.

A mintaáltalánosító eszköz fejlesztéséhez a díjadás domént használtuk fel (ki milyen díjat, elismerést, kitüntetést stb. nyert), mivel összehasonlítási alapként ehhez is

rendelkezésre állt már egy manuálisan fejlesztett mintakészlet, illetve egy annotált kiértékelő halmaz. A tanítópéldák előállításához a Yago *hasWonPrize* relációját használtuk, mely személyek és díjnevek között áll fenn. A személyek oldalain a díjneveket visszakeresve mintegy 16 ezer potenciális tanítómondathoz jutottunk. Mivel a *hasWonPrize* reláció zajos volt – a 2. argumentumhelyen nem csak díjnevek, hanem a díjat elnyerő műalkotások (filmcímek) is szerepeltek, – kiszűrtük azokat a példákat, amelyekben a 2. argumentum nem volt díjnév, vagyis nem szerepelt a mintegy 7.400 tételt tartalmazó lokális lexikonban, illetve nem illeszkedett rá az erre a célra létrehozott reguláris kifejezés, így 13 ezer mondat maradt.

4.2 Minták előállítása és általánosítása

A tanítómondatokat az Enju parserrel és az erre épülő, a 2. pontban bemutatott igeiszerkezet-kinyerő modullal dolgoztuk fel.

A következő lépésben az azonosított összetevőkben (alany, tárgy, direkt tárgy stb.) azonosítottuk és annotáltuk a Yago-reláció által megadott argumentumokat. A díjadás domén esetében ez egyrészt a díjat kapó személyre utaló kifejezést, másrészt a díj nevét jelentette. Előbbi azonosításához felhasználtuk a személyről szóló Wikipédia-oldal címét, a személy szócikkének első bekezdésében az első mondatban szereplő kövérrel kiemelt névváltozato(ka)t, valamint ezek token-részsorozatait. Ezek hiányában személyes névmásokat is elfogadtunk a címszeméllyel koreferáló kifejezésként, figyelembe véve azok nyelvtani nemét, ha az megfelelt a személy szócikkében leggyakrabban előforduló nemű névmásoknak.

A két Yago-argumentum annotálása után kiszűrtük azokat a példamondatokat, amelyek csak valamelyiket, vagy egyiket sem tartalmazták, 11 ezer valódi tanítómondathoz jutva így.

A mintaáltalánosítás megkönnyítése érdekében a mondatokban felismertünk és egyszerű tagekre cseréltünk néhány egyszerű, gyakori, reguláris kifejezésekkel könnyen felismerhető névelemet (sorszámnevek, tőszámnevek, dátumok különböző formátumban, hónapnevek, évek, számok).

A nyelvi elemzéssel, egyszerű névelemekkel és Yago-szerepekkel annotált példamondatokat a következő lépésben mintákká alakítottuk. Minden minta *(G, S)* rendezett párok sorozata, ahol *G* valamilyen nyelvtani összetevő elnevezése (*Verb, Subj, Obj, PP-xx* stb.), *S* pedig az összetevőt alkotó tokenek sorozata, mely metatokenekből (Yago-relációk vagy egyszerű névelemek tag-jei) vagy mondatbeli szavak felszíni alakjaiból áll.

A következő lépésben összevontuk az azonos mintákat (megtartva hivatkozásaikat azokra az eredeti szövegbeli mondatokra, amelyekben szerepeltek.) Ezeket a mintákat ezután klasztereztük úgy, hogy egy-egy mintaosztályba (klaszterbe) akkor került 2 minta, ha a) ugyanazt az igét tartalmazták, és b) a 2 Yago-relációargumentumot ugyanazokban a nyelvtani szerepekben tartalmazták. Ezzel a módszerrel a díjadásos mondatokból 376 különböző mintaosztályt kaptunk. A mintaosztályokat a bennük szereplő minták által lefedett mondatok számával rangsoroltuk. 64 olyan osztály volt, amely 2-nél több mondatot fedett le a tanítóhalmazban, ez a tanítópéldák 97%-át jelenti.

Az alábbi példában bemutatjuk a rangsorban 1. és 4. helyen szereplő mintaosztályokat és néhány első elemüket, feltüntetve az elemi minták által lefedett tanítómondatok számát:

```
Class id: 8
Sentences covered by patterns in class:  1092
Patterns in class: 210
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE#'))     548
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', '@CARDINAL@ #PRIZE#s')) 99
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', '@YEAR@ #PRIZE#'))      98
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', '@ORDINAL@ #PRIZE#'))   48
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', 'Daytime #PRIZE#'))     22
(('Verb', 'win'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE#s'))    17
…
Class id: 5
Sentences covered by patterns in class:  406
Patterns in class: 258
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE#'))      27
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE# -winning American actor'))      18
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE# -winning American actress'))    18
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', 'recipient of the #PRIZE#'))     10
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', 'American #PRIZE# -winning actor'))      8
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', 'American #PRIZE# -winning actress'))    7
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', '#PRIZE# winner'))       6
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', 'winner of the @YEAR@ #PRIZE#')) 6
(('Verb', 'be'), ('Subj', '#PERSON#'), ('Obj', '@CARDINAL@-time #PRIZE# winner'))       5
...
```

Látható, hogy a negyedik helyen álló mintaosztályban – amelynek 258 eleme összesen 406 mondatot fed le, és ahol a „be" a főige, az 1. argumentum (díjat elnyerő személy) alanyi, a 2. argumentum (elnyert díj neve) pedig tárgyi szerepben áll – a díjnév környezetében több különböző token is szerepel. Ezek egy része egyértelműen utal arra, hogy az igei szerkezet díjadás eseményt jelöl (pl. *winner, recipient* stb.), egy része viszont nem (pl. *actor, actress* stb.) Célunk ezért az volt, hogy megpróbáljunk automatikusan javaslatot tenni a felismerendő eseménnyel korreláló markerszavakra, elkülönítve őket a többi, zajként értelmezhető szótól.

A markerszavak azonosítására a Pearson-féle (egyoldalú) χ^2-próbát alkalmaztuk. Ehhez szükség volt olyan negatív példákra, amelyek – szemben az előzőekben ismertetett módon előállított pozitív példákkal – nem a vizsgált relációról szólnak. Ehhez a kérdéses személyekről szóló szócikkeknek azokat a mondatait vettük, amelyek nem szerepeltek a pozitív példamondatok között. Mivel ilyen mondatból jóval több volt, mint pozitív mondatból, a tanítóhalmaz kiegyensúlyozottsága érdekében ezek közül véletlenszerűen kiválasztottunk a pozitív minta nagyságának megfelelő számú mondatot.

A χ^2-próbával megvizsgáltuk, hogy melyek azok a (mintákban szereplő) szavak, amelyek függőséget mutatnak a pozitív-negatív besorolással. A teszthez empirikus úton a 25,0 kritikus értéket választottuk (α=0,01-nél a kritikus érték 6,635 lenne).

A mintaáltalánosítás utolsó lépéseként minden egyes mintaosztályhoz generáltunk egy-egy mintajavaslatot, mely a χ^2-próbával azonosított markerszavak felsorolását tartalmazta. Ezeket a javasolt mintákat kellett ezután egy humán annotátornak átnéznie, és a szükséges módosítások után pozitív vagy negatív mintaként megjelölnie. Az eredmény a fenti két mintaosztályra:

```
Class id: 8
+(('Verb', 'win'), ('Subj', '* #PERSON# *'), ('Obj', '* #PRIZE# *'), ('ObjII', '*'))
-(('Verb', 'win'), ('Subj', '* #PERSON# *'), ('Obj', '* #PRIZE# -nomination|nomination|'), ('ObjII', '*'))

Class id: 5
+(('Verb', 'be'), ('Subj', '* #PERSON# *'), ('Obj', '* #PRIZE# -winning|recipient|winner|winning'))
+(('Verb', 'be'), ('Subj', '* #PERSON# *'), ('Obj', 'recipient|winner #PRIZE# *'))
```

Az általánosított minták *(G, S)* rendezett párjaiban *S* tartalmazhat diszjunktív felsorolásokat („|”-jellel elválasztva), ezen elemek közül legalább az egyik megléte kötelező a minta illeszkedéséhez. A „*” jellel tetszőleges token elfogadható az adott pozícióban. A „+”prefix az annotátor által megjelölt pozitív, a „-”a negatív mintákat jelöli, melyekkel kizárhatók bizonyos szerkezetek az információkivonatolásból (pl. a díjadás domén esetében ki akarjuk zárni a díjakra *jelölés* eseményére utaló mondatokat.)

4.3 Információkivonatolás az általánosított mintákkal

A díjátadás területen (Yago *hasWonPrize* relációval generált tanítópéldák) egy humán annotátor az 5-nél több mondatot lefedő mintaosztályokhoz hagyott jóvá automatikusan generált általánosított mintákat (ez körülbelül 1,5 munkaórát jelentett.)

Ezeket a mintákat ezután felismertettük 100 Wikipédia-szócikk szövegén, melyekhez humán annotátorok előzőleg már elkészítették a helyes válaszokat tartalmazó annotációt. A mintafelismerésben ugyanazokat a nyelvtani, névelem- és szerepannotáló algoritmusokat használtuk, mint a minták előállításánál. A negatív mintákkal előzetesen megszűrt tesztmondatokra illesztve a pozitív mintákat, kinyertük a tesztszövegekből a kérdéses attribútumokat. Ezeket a manuális annotációval összehasonlítva a fent megadott módon kiszámoltuk a félautomatikusan elállított mintákkal végrehajtott információkivonatolás pontosságát és fedését. A 2. táblázatban ezek mellett az értékek mellett feltüntettük a díjátadás területre teljesen manuálisan kidolgozott mintákkal történt információkivonatolás kiértékelésének eredményét ugyanerre a teszthalmazra.

2. táblázat: A félautomatikusan előállított mintákkal végrehajtott információkivonatolás összehasonlítása a teljes egészében manuálisan létrehozott mintákkal végrehajtott információkivonatolás eredményével.

	Precision	Recall
Félautomatikusan generált mintákkal	91,66%	36,70%
Manuálisan létrehozott mintákkal	93,97%	50,00%

Hivatkozások

1. http://blog.iglue.hu/
2. http://en.wikipedia.org/wiki/
3. http://alias-i.com/lingpipe/

4. http://static.wikipedia.org/downloads/2008-06/
5. Sagae, Kenji, Miyao, Yusuke, and Tsujii, Jun'ichi: HPSG Parsing with Shallow Dependency Constraints. In Proceedings of ACL (2007)
6. McCallum, Andrew Kachites: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. (2002)
7. Suchanek, Fabian, Kasneci, Gjergji, Weikum, Gerhard: YAGO: A Core of Semantic Knowledge - Unifying WordNet and Wikipedia. In Proc. Of 16th International World Wide Web Conference (WWW 2007) (2007) 697–706

Panaszlevelek automatikus kategorizálása szerkezeti egységek és jellemző kifejezések figyelembevételével

Bártházi Eszter[1] *, Héder Mihály[2,3] **

[1] MTA SZTAKI Géppel Támogatott Megértés Kutatócsoport, barthazi@sztaki.hu
[2] MTA SZTAKI Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[3] Budapesti Műszaki és Gazdaságtudományi Egyetem Filozófia és Tudománytörténet Tanszék

1. Bevezető

2008-ban indult kutatásunk célja, hogy egy rendszert készítsünk, amely egyszerre könnyíti meg valamely hivatal és a hozzá forduló ügyfelek dolgát. A gép közreműködésének lényege, hogy az ügyfél számára egy felületet nyújt, ahol panaszát, hozzászólását (továbbiakban levelét) megfogalmazhatja. A levél írása során az elképzelt rendszer dialógusok formájában kapcsolatot teremt a levélíróval, kérdések segítségével pontosabb információkat kér, megpróbálja eldönteni, hogy a levél milyen hivatali kategóriába tartozik.

Ezen elképzelt rendszer megvalósításához mindenekelőtt kiterjedt alapozó kutatások szükségesek. A jelen cikk ezen kutatásokat, kísérleteket mutatja be, melyeket az Igazságügyi Minisztériumtól kapott, nagyon változatos, közel 900 levélből álló korpuszon (a továbbiakban korpusz) hajtottunk végre. A korpuszért külön köszönetet szeretnénk mondani dr. Vörös Editnek, az Igazságügyi és Rendészeti Minisztérium Társadalmi Kapcsolatok Osztálya vezetőjének, aki nemcsak rendelkezünkre bocsátotta a szövegkorpuszt, hanem gondoskodott a felhasználás jogi és előfeldolgozási körülményeiről.

A korpusszal végzett munka első lépése az előfeldolgozás volt. Ehhez egy alkalmas keretrendszert készítettünk, amely integrálja a magyar nyelvre elérhető különféle elemző eszközök jelentős részét. A 2. fejezet ezt a keretrendszert mutatja be.

A feladat sajátossága, hogy a korpusz szókincse kivételesen terebélyes, inhomogén, hivatalosnak egyáltalán nem nevezhető. A levelek fogalmazása is gyakran hiányos, nehezen értelmezhető, és sok helyesírási hibát, elírást tartalmaz. Mivel az automatikus kategorizáló rendszerek igazán jó teljesítményt csak egy jól behatárolható terület szaknyelvi kontextusában szoktak elérni, kiemelten sok

* PhD-hallgató, témavezető: Németh T. Enikő
** PhD-hallgató, témavezető: Vámos Tibor

energiát kell fordítanunk arra, hogy a hétköznapi és a szaknyelv között kapcsolatot teremtsünk, illetve hogy a levelek által meghatározott túlságosan tág, emiatt nehezen kezelhező kontextust leszűkítsük. Egy kísérlet a kontextus szűkítésére a szerkezeti egységek detektálása és felhasználása a kategorizálási kísérletek során. A szerkezeti egységek jelentősége, hogy segítséget nyújtanak abban, hogy bizonyos típusú információkat hol érdemes keresni. Például a kategorizálás szempontjából lényeges részek többnyire a problémát bemutató szerkezeti egységben találhatók, míg a levélíróról rendelkezésre álló adatok jellemzően a Bemutatkozás szerkezeti egységben keresendők. Ezt az elképzelést részletesebben a 3. fejezet bontja ki.

A következő lépés a rendszer megvalósítása felé a jó eredményekkel működő osztályozási és csoportosítási algoritmusok kipróbálása a korpuszon. Ez egyrészt információt szolgáltat számunkra a meghaladni kívánt pontosságról, másrészt a végleges rendszerben is fel szeretnénk használni a kategorizálás, illetve a kategorizálási javaslatra vonatkozó dialógus megvalósításánál. A részleteket a 4. fejezet tartalmazza.

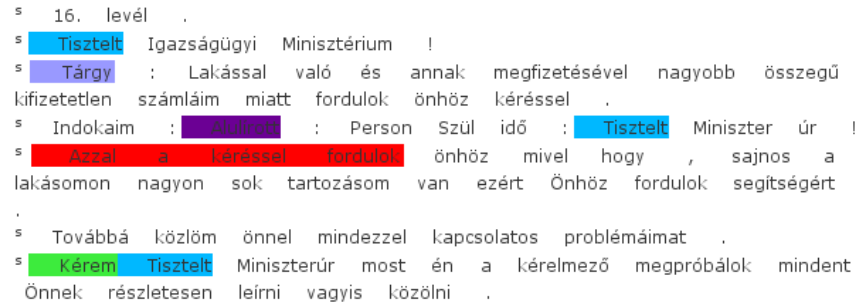
A szerkezeti egységek ismeretében már lehetőségünk van egy speciálisabb, az ügyintéző feltételezett gondolatmenetét modellező kategorizálási eljárás készítésére. Feltevésünk szerint az ügyintéző a levelek feldolgozásánál forgatókönyveket követ. A KATEGORIZÁLÁS forgatókönyv egy általunk formalizált algoritmus, ami megadja a panaszkategóriába való soroláshoz vezető lépéseket. A gép az algoritmust bejárván, miután azonosította az ügyfelet, azokat a szerkezeti egységeket vizsgálja meg, amelyek a kategorizálás szempontjából releváns kifejezéseket taralmazzák. A KATEGORIZÁLÁS forgatókönyv részleteit az 5.1. fejezet fejti ki. A jövőben a besorolástól függően újabb, immár kategóriaspecifikus forgatókönyvekkel is szeretnénk kísérletezni.

Számos, a kategorizálás szempontjából irreleváns, ugyanakkor egyéb – szociológiai, valamint pszichológiai – szempontból fontos információ is található a levelekben. A levelek besorolását, és még inkább a dialógusok alakítását befolyásolja a levélíró pszichés-szociológiai profilja, amelyet a használt kifejezések és fordulatok, a szerkesztési jegyek alapján folyamatosan építünk. A profil meghatározásához egy, A Magyar nyelv értelmező szótárára [1] épülő, a szavak stilisztikai jegyeit tartalmazó listát használtunk. A részleteket az 5.2. fejezetben mutatjuk be.

Ezen kutatás közvetlen előzményének tekinthető Héder diplomaterve [2], amely szemantikus annotációk géppel támogatott elhelyezését tárgyalja webes dokumentumokban. Abból a munkából eszközöket és sok tapasztalatot sikerült átmenteni, de hiányzott belőle a szöveg nyelvi, szerkezeti elemzése és a profil készítése.

2. A használt keretrendszer

Kutatásunkhoz egy egyszerűen használható, általános előfeldolgozó, illetve nyelvi elemző rendszert készítettünk, melynek segítségével sok különféle, kész eszközt homogén módon tudunk kezelni. A fejlesztés fő követelményeiként a könnyű hasz-



1. ábra. Egy DMD-fájl vizuális megjelenése

nálhatóságot, az új eszközök minél egyszerűbb integrációját és a robusztusságot jelöltük meg. Mivel a fő célunk nem eszközfejlesztés, törekedtünk minden elérhető megoldás beépítésére.

Az így elkészült rendszer bemenete egy egyszerű szövegfájl vagy strukturált XML-dokumentum lehet. A kimenet egy úgynevezett Docuphet Mixed Document (DMD) típusú XML-fájl, amelyet több névtérből gyúrtunk egybe, úgy, hogy az lehetőleg minden elképzelhető annotációtípust hordozni tudjon. A DMD saját hordozó névterén kívül definiáltunk egy névteret a projektben létrehozott eszközeink számára is. A többi névtér a felhasznált külső eszközök annotációit reprezentálja. Használtuk a Hitec projekt [3] kapcsán kifejlesztett fulldoc formátum egyes elemeit és a Huntools jól ismert komponenseit, a Huntokent, a Hunmorphot és a Hunpost.

A névterek éles megkülönböztetése révén megpróbáljuk a jövőbeli feldolgozó eszközök számára minél egyszerűbbé tenni az általuk ismert névterek elemeinek kezelését, miközben az ismeretlen névtereket figyelmen kívül hagyhatják. Ezzel egyidejűleg lehetővé tesszük a rendszerünk zökkenőmentes kiterjesztését is.

A DMD-fájloknak van egy egyszerű, informatívnak és tetszetősnek szánt XHTML megjelenítése is (2. ábra). A DMD-fájl XHTML formátumba való konvertálását XSLT 2 transzformációval végezzük.

Korábbi saját fejlesztés [2] az eredetileg névelemek, később a tipikus szerkezeti egységek (lásd 3. fejezet) felismerésére használt JNER rendszer. A java nyelven íródott eszköz szabályok és katalógusok segítségével végzi feladatát.

Az egyes névterekkel jelölt annotációkat különféle szkriptek lefuttatása állítja elő. Vannak a Huntools egyes elemeit, illetve a JNER-t egy-egy fájlon lefuttató szkriptek, mások minden feladatot kötegelten, esetleg egész könyvtárakra hajtanak végre. Készítettünk eszközöket a szó, szótő és egyéb típusú statisztikák gyűjtésére is. Megemlítendő, hogy az integrált, minden elemzést egyben elvégző megoldáshoz webes felületet is készítettünk, ahol a beírt szöveg AJAX technológia segítségével a háttérben feldolgozásra kerül.

3. Szerkezeti egységek annotálása

A vizsgált panaszlevelek esetében megfigyelhető, hogy az állampolgárok jelentős része a hétköznapi szókincsére támaszkodva pontatlanul, hiányosan, sok esetben nehezen érthetően fogalmazza meg a panaszát, és számos, az ügyintézés szempontból irreleváns információt is közöl. Ezzel megjósolhatatlanul tág kontextusba helyezi a levélben megfogalmazottakat. Továbbá a levelek szerkezeti felépítése is igen változatos, ezért pusztán a közigazgatási területekre jellemző terminológiára támaszkodva egy bottom-up megközelítéssel nagyon nehéz jó eredményt elérni. Ennek a problémának a megoldásaként, a leveleket alaposan megvizsgálva tizenkét szerkezeti egységet találtunk, amelyek egyben kontextusként, értelmezési keretként is szolgálnak a bennük előforduló kifejezések interpretálásához. A tizenkét szerkezeti egység a következő:

1. **Megszólítás**: a levélíró valamilyen módon kifejezi, hogy kinek szánja levelét, pl.: *Tisztelt [személynév/titulus/intézménynév/stb.]*
2. **Bemutatkozás**: a levélíró azonosításához szükséges adatokat tartalmazza, pl.: *Alulírott, [személynév], született [évszám], anyja neve [személynév]* stb.
3. **Cél**: a levélíró még a panasza ismertetése előtt kifejezi, hogy milyen területen vár segítséget, pl.: *Tárgy: nyugdíjügy.*
4. **Előzmény**: a jelenlegi problémát megelőző, de ahhoz kapcsolódó események ismertetése, pl.: *Kértem a miniszter urat, hogy...*
5. **Probléma**: a levélíró a problémáját részletezi, pl.: *Az alábbi problémámra várnám a segítséget.*
6. **Javaslat**: a Probléma szerkezeti egység alternatívája, amikor a levélíró nem egy megoldásra váró problémával fordul a minisztériumhoz, csupán egy javaslatot tesz valamivel kapcsolatban, pl.: *A következő javaslattal fordulok Önökhöz...*
7. **Vádaskodás**: a levélíró indulatait, kétségeit fejezi ki, erősen emocionális módon, pl.: *Hol itt a törvény?*
8. **Elismerés**: a levélíró elismerését fejezi ki a levél címzettjének eddigi tevékenységével szemben, pl.: *Engedje meg, hogy gratuláljak.*
9. **Egyéb körülmények**: a levélíró a problémájához szorosan nem vagy egyáltalán nem kapcsolódó egyéb problémáját, életkörülményeit, egészségügyi állapotát stb. ecseteli, pl.: *Az igaz hogy jobb kezem az ujjam hegyétől a vállamig és az egész törzsem a derekamtól a fejem hegyéig zsibog a jobboldalamon* – egy öregségi nyugdíjemelési kérelemről szóló levélben.
10. **Elvárás**: a levélíró azt fogalmazza meg, hogy milyen viselkedést, intézkedést vár el az ügyintéző részéről, pl.: *A fentiek alapján kérem...*
11. **Köszönet**: a levélíró megköszöni az eddigi intézkedést, türelmet, illetve előre is megköszöni a további intézkedéseket, pl.: *Előre is köszönöm, hogy válaszlevelével megtisztel.*
12. **Lezárás**: a levélíró egy adott formulával befejezi a levelét, pl.: *Minden jót.*

Az egyes szerkezeti részek sorrendje levelenként eltérő lehet, és természetesen nem minden szerkezeti egység található meg minden levélben. Az azonban,

hogy mely szerkezeti egységek fordulnak elő egy adott levélben, valamint az is, hogy milyen sorrendben, további információval szolgálhat a levélíróval kapcsolatban. A szerkezeti egységeknek köszönhetően az információkinyerés egyszerre bottom-up (jellemző kifejezések figyelembevétele a kategorizálás során) és top-down folyamatok eredménye (egy bizonyos kontextusban/értelmezési keretben történik), ami azért is fontos, mivel a humán megértés során a kontextus ismerete éppúgy irányítja az interpretációt, mint az egyes kifejezések jelentése (a kompozicionalitás és a kontextualitás elvének együttműködése, lásd [4]).

A szerkezeti egységek felismeréséhez a levelek 10%-ának manuális elemzésével elkészítettük az egyes egységeket tipikusan jelölő definitív kifejezések listáját. A lista alapján a JNER segítségével annotáltuk a leveleket, az annotációk megjelenítéséhez színkódokat használtunk (lásd a 2. ábrát).

A megoldás tesztelése azt az eredményt hozta, hogy a lista még kiegészítésre szorul, ugyanis sok levélben csak kevés szerkezeti egységet találtunk így. Ennek oka feltételezhetően kettős: egyrészt az általunk vizsgált 89 levél valószínűleg nem reprezentálja a teljes korpuszt megfelelően; másrészt a levelekre jellemző szóhasználat sokkal változatosabb annál, mint amit ezzel az egyszerű módszerrel jelenleg kezelni tudunk. Ugyanakkor jó eredménynek tartjuk, hogy a felismert szerkezeti egységek többnyire helytállóak. A helyesen felismert szerkezeti egységek százalékos arányát megfelelő tesztadatok hiányában egyelőre nem tudjuk megállapítani.

4. Osztályozási és csoportosítási kísérletek

4.1. A korpusz

A kutatás alanyául szolgáló korpusz az Igazságügyi Minisztériumhoz beérkezett 888 levél digitális, anonimizált verziójából állt. A levelekről általánosan elmondható, hogy igen szerteágazó témakörökben és nagyon változó stílusban, illetve helyesírással íródtak. Továbbá sok levél nyilvánvalóan felfokozott érzelmi állapotban (düh, elkeseredettség) íródott, értékes alapanyagot szolgáltatva ezáltal a levélírók különféle profiljainak meghatározásában.

A közel kilencszáz levélből Kabai Dóra munkája [5] nyomán 210-hez rendelkezésünkre állt kategóriainformáció is. Ezen levelek 10 kategóriába voltak besorolva. Némely levél több kategóriába is tartozott egyszerre, így a kategóriabesorolások összesített levélszáma 330 volt.

4.2. Szűrés

A korpuszon először különféle szűrési eljárásokat próbáltunk ki. Feltételezésünk szerint a szűrésnek nagy szerepe lehet a csoportosítás és osztályozás hatékonyságának növelésében, de még nagyobb a gépi megértést nem befolyásoló, vagy zavaró zaj csökkentésében.

Az egyes szűrési eljárásokkal eredeti szóalakokat tartalmazó, illetve csak szótöveket tartalmazó tanulóadat-verziót is előállítottunk. A szótöveket minden esetben a HunMorph segítségével állapítottuk meg.

A legegyszerűbb szűrésünk azon szóalakok kihagyása volt, amelyek a levelek több mint 50 %-ában szerepelnek. Épp 50 ilyen szóalakot találtunk. Ide soroltuk továbbá az egyéb karaktereket is. Ezt a szűrési típust a továbbiakban H betűvel jelöljük.

Az egyszerű, ökölszabályszerű H szűrés mellett kézzel is készítettünk egy szűrőlistát. A lista elkészítése során figyelembe vettük a szavak eloszlását is a levelekben, ebben a Weka rendszer volt segítségünkre [6]. Ez a lista a H listával szemben nem szóalakokat tartalmaz, hanem 235 szótövet (pl.: ha, mert, stb.) , illetve a HunMorph különféle morfológiai elemzési kimenetei közül 111-et (pl.: DET, ART, PUNCT, stb.). Ezt a szűrőlistát elsősorban a csoportosítás, illetve osztályozás hatékonyságának növelésére szántuk. Az információkinyerés és megértés szempontjából alkalmazásuk nem feltétlenül célszerű, mert kiszűri többek között a tagadószavakat, számneveket, illetve a létigéket is, ezáltal információveszteséget termelve. Ezt a szűrést a továbbiakban K-nak nevezzük.

A H és K globálisan alkalmazott szűrési eljárások mellett nagy reményekkel kísérletezünk egy, az egyes leveleken külön-külön kiértékelendő szűrési metódussal is. Ennek során megkíséreljük azonosítani a levelek szerkezeti egységeit, és a kategorizálás szempontjából irreleváns mondatokat – jelenleg: Megszólítás, Lezárás, Vádaskodás – teljes egészében kivesszük. A strukturális elemek azonosításáról a 3. fejezet szól. Ezen szűrést a továbbiakban S-nek nevezzük.

A szűretlen levelek összesen 425 ezer tokenből állnak – így kb. 450 token/levél adódik. Ha kiszűrjük a többszörös előfordulásokat, 53 ezer különböző tokent kapunk. Szótövekre ugyanezen számok 318 ezer (az egyéb karaktereknek, mondatvégi jeleknek nincs szótöve, ezért a különbség) és 13,5 ezer. Az egyes szűrések alkalmazásával az összméret kevesebb mint a felére csökkenthető, és az egyedi szóalakok, illetve szótövek száma is csökken. Külön kiemelendő, hogy az S szűrés kb. 30 ezerrel csökkenti az összesített méretet, de még az egyedi számokat is csökkenti néhány százzal.

Az adathalmazokból a Weka által feldolgozható Vektor (arff) fájlokat készítettünk. Itt két további szűrést alkalmaztunk: elhagytuk a kevesebb mint háromszor szereplő elemeket, illetve összevontuk a kicsi és nagybetűs verziókat. Más korlátozást a vektor dimenzióinak méretére (az arff attribútumok számára) nem tettünk.

4.3. Kategorizálás

A célunk az volt, hogyan megvizsgáljuk, javítható-e a kategorizálás pontossága és hatékonysága a különféle szűrések segítségével. A kísérleteket a 210 kategorizált levéllel végeztük, úgy, hogy a levelek kétharmadát tanításra, a fennmaradó egyharmadot tesztelésre használtuk fel. Két elterjedtnek mondható algoritmust is kipróbáltunk.

A Naive Bayes [7] és az SVM [8] eljárásokat a Weka keretrendszer által alapértelmezettként felkínált paraméterekkel futtattuk. SVM implementáció gyanánt a libsvm rendszert vettük igénybe a Wekán keresztül.

A kísérleteket elvégeztük a szűretlen leveleken, illetve a szűrők H, H+K, S, S+H, S+H+K kombinációival átrostált leveleken is. Minden tesztet elvégeztünk

a szótő, illetve szóalak vektorokon is. A 1. táblázat tájékoztat az egyes lefutások időigényéről, illetve a helyesen kategorizált levelek százalékáról, zárójelben a pontos levélszámmal.

Elmondható, hogy bár a pontosságot nem befolyásolta lényegesen a szűrések alkalmazása, a legjobb eredményeket döntően az összes szűrő együttes alkalmazásával kaptuk. Eközben a futási idők jelentősen csökkentek. Az is kiemelendő, hogy ezen a korpuszon a szóalak és szótő vizsgálata között az összes szűrés együttes használata mellett nincs különbség. Érdekesség ugyanakkor, hogy a Bayes-módszernél sokkal jobb pontosságú SVM kiegyensúlyozott eredményt hoz a szóalakok esetén a szűréstől függetlenül, de érzékeny a szűrés hiányára a szótövek esetében. A Bayes-módszer ezzel szemben a szóalakok esetében jobban működik, ha erős szűrést alkalmazunk, a szótövek esetében viszont épp fordítva: gyengülő teljesítményt mutat a szűrések hatására. A szóalak vektorok dimenzióinak száma kb. kétszer nagyobb, mint a szótövek dimenzióinak száma, ami feltehetően szerepet játszik a tapasztalt eltérésben.

Összegzésként elmondható, hogy – bár a kategorizált levelek kis száma nem engedi meg nagyon erős általánosítások tételét –, a szűrés semmi esetre sem rontott az osztályozás pontosságán, ugyanakkor a futási időket és a feldolgozandó adatmennyiséget jelentősen csökkentette.

1. táblázat. Az osztályozási kísérletek eredményei

Típus	Szűréstípus	Naive Bayes	Futási idő(s)	SVM	Futási idő(s)
Szóalak		14.2857 % (16)	8.65	29.4643 % (33)	2.06
Szóalak	H	14.2857 % (16)	8.49	30.3571 % (34)	2.07
Szóalak	H+K	16.0714 % (18)	6.05	30.3571 % (34)	1.13
Szóalak	S	15.1786 % (17)	7.66	29.4643 % (33)	2.4
Szóalak	S+H	15.1786 % (17)	7.49	30.3571 % (34)	2.13
Szóalak	S+H+K	16.9643 % (19)	5.32	30.3571 % (34)	1.26
Szótő		16.9643 % (19)	4.47	25.8929 % (29)	1.59
Szótő	H	16.9643 % (19)	4.1	28.5714 % (32)	2.09
Szótő	H+K	16.9643 % (19)	3.49	30.3571 % (34)	1.11
Szótő	S	16.9643 % (19)	4.02	25.8929 % (29)	1.57
Szótő	S+H	16.0714 % (18)	3.85	28.5714 % (32)	1.42
Szótő	S+H+K	15.1786 % (17)	3.08	30.3571 % (34)	1.17

4.4. Csoportosítás

Néhány kísérletet elvégeztünk az X-mérték [9] csoportosítási algoritmussal is. Ezen algoritmus sajátossága, hogy a csoportok számát is képes adaptívan meghatározni, ugyanakkor a vágáshoz egyszerű K-mérték eljárást használ. A számunkra érdekes kérdés az volt, hogy a sok helyen előforduló, de lényegi információt nem tartalmazó szavak/mondatok szűrése segíti-e a csoportok elkülönülését. Ezért az algoritmust minden esetben tíz iterációban futtattuk, a kapott csoportok számát 2 és 30 közé limitálva.

Ahogy a 2. táblázatban látható, a csoportok számát a szűrések nem módosítják, ellenben a csoporthozzárendelésekre jelentős hatásuk van. Ezen jelenség okát az attribútumeloszlások és csoport-hozzárendelések emberi vizsgálata tárhatná fel.

2. táblázat. A csoportosítási kísérletek eredményei

Típus	Szűrés típus	Csoportok száma	eloszlások
Token		4	104(12%), 329(37%), 209(24%), 245 (28%)
Szótő		4	67(8%), 214(24%), 271(31%), 335(38%)
Szótő	H	4	79(9%), 297(33%), 207(23%), 304(34%)
Szótő	H+K	4	106(12%), 299(34%), 188(21%), 294(33%)
Szótő	S+H+K	4	97(11%), 285(32%), 213(24%), 292(33%)

5. Forgatókönyv és profil

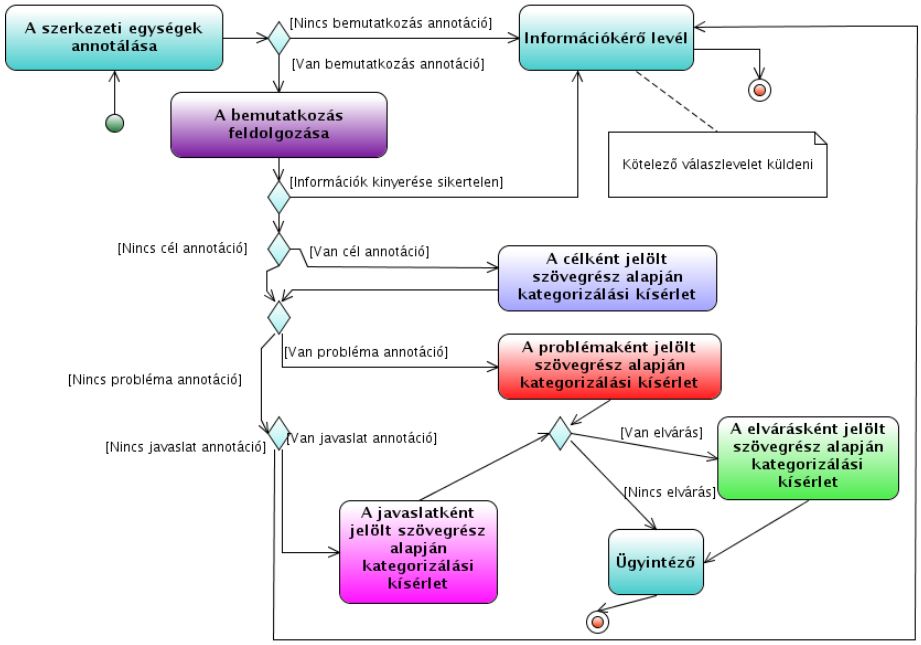
Az egyes panaszlevelekből két célból szeretnénk információt kinyerni. Az egyik cél az, hogy megállapítsuk, hogy melyik panaszkategóriába tartozik az adott panaszlevél, a másik, hogy az azt beküldő ügyfélről egy profilt alakíthassunk ki, az ügyfél aktuális érzelmi állapotáról, szociális körülményeiről tehessünk megállapításokat, amelynek majd a későbbiekben, a dialógusokban lesz fontos szerepe.

5.1. A kategorizálás forgatókönyv

A kategorizáláshoz elképzelésünk szerint a következő szerkezeti egységeket kell figyelembe venni: Bemutatkozás, Cél, Probléma, Javaslat, Elvárás. Ezek a levélnek azon részei, amelyek a panaszkategória megállapításához szükséges definitív kifejezéseket tartalmazzák, tehát azokat a nyelvi elemeket, amelyek alapján eldönthető, hogy az adott levél írója milyen kategóriájú panasszal fordul az ügyfélszolgálathoz. A Bemutatkozás szerkezeti egység figyelembevétele pedig azért alapvető, hogy az állampolgár egyértelmű azonosítása lehetővé váljon.

A kategorizálás általános forgatókönyvét a 5.1. ábrán látható aktivitás diagram mutatja be. Az algoritmus először a Bemutatkozás szerkezeti egységet keresi, hogy ezt feldolgozva kinyerhesse azokat az információkat, amelyek segítségével egyértelműen azonosítható az ügyfél. Ezen szerkezeti egység azonosítása a jellemző definitív kifejezések alapján történik.

A definitív kifejezések kétfélék lehetnek: kategóriasemlegesek vagy kategóriaspecifikusak. A kategóriasemleges kifejezések kizárólag az adott szerkezeti egység beazonosításában játszanak szerepet. A kategóriaspecifikus kifejezések szintén segíthetnek az adott szerkezeti egység beazonosításában, de nem ez az elsődleges feladatuk, hanem az, hogy az egységen belül a kategorizáláshoz szükséges, tartalmi szempontból releváns információkat hordozzák. Másrészt a kategóriaspecifikus kifejezések megtalálását a kategóriasemleges kifejezések segíthetik.



2. ábra. A KATEGORIZÁLÁS forgatókönyv aktivitás diagramja

Ha a rendszer nem talál bemutatkozás annotációt, azaz a Bemutatkozás szerkezeti egységet nem sikerül azonosítani, illetve ha az azonosítás sikerült, de az információk kinyerése sikertelen, akkor egy levél kerül kiküldésre az ügyfélhez, amely egy arra vonatkozó kérést tartalmaz az ügyfél felé, hogy pótolja a hiányzó adatokat. Az információkérő levél küldésével egyben az az elvárás is teljesül, miszerint a hivatalnak kötelező válaszlevelet küldeni minden egyes panaszlevélre egy meghatározott időn belül.

Amennyiben a gépnek sikerült azonosítania az ügyfelet, a következő lépés a Cél szerkezeti egység keresése. Ha a rendszer talál cél annotációt, azaz a Cél szerkezeti egységre jellemző kategóriasemleges- vagy kategóriaspecifikus definitív kifejezések alapján képes beazonosítani azt, akkor az ezen a szerkezeti egységen belüli kategóriaspecifikus definitív kifejezések segítségével (ha vannak ilyenek) azonosíthatóvá válik a panaszkategória. Minden lehetséges esetben, azaz ha a Cél szerkezeti egység hiányzik, vagy ha a szerkezeti egységet sikerült ugyan azonosítani, de kategorizálás szempontjából releváns információt nem sikerült belőle kinyerni (azaz a rendszer nem talált kategóriaspecifikus definitív kifejezéseket), vagy harmadik lehetőségként, ha a panaszkategóriát sikerült azonosítani, a keresés a Probléma szerkezeti egységgel folytatódik.

Amennyiben a Probléma szerkezeti egységnek az azonosítása megtörtént, akkor, feltéve, hogy a definitív kifejezések között talál kategóriaspecifikusakat, a gép újra elvégzi a kategorizálási lépést, most már ebben az egységben talált definitív

kifejezések figyelembevételével. Ez a panaszkategória lehet azonos az előzővel, de lehet ettől eltérő is.

Miután sikerült a levélhez panaszkategóriát rendelni, a keresés az Elvárás szerkezeti egységgel folytatódik tovább. Ezt a szerkezeti egységet szintén a kategóriasemleges, valamint a kategóriaspecifikus kifejezések figyelembevételével azonosítja a rendszer, és csakúgy mint az előző esetekben, a kategóriaspecifikus kifejezések alapján újra egy panaszkategóriát rendel a levélhez.

Az eddigiek alapján tehát az algoritmus ezen pontján a következő esetek állhatnak fenn: a Cél, a Probléma, valamint az Elvárás szerkezeti egységek alapján a gép egy, kettő vagy három különböző panaszkategóriát rendelt a levélhez. Az első esetben az algoritmus következő lépése, hogy a levelet a megállapított panaszkategóriában jártas ügyintőzőhöz továbbítja, míg a második és harmadik esetben, hogy a két-, illetve három, az adott panaszkategóriában jártas ügyintézőkhöz kerül a levél továbbításra. Azok a levelek, amelyek több ügyintézőhöz is eljutnak, tartalmazzák azt az információt, hogy kik a címzettek, hiszen ez az ügyintézők számára releváns lehet.

Abban az esetben, ha az algoritmus a Probléma szerkezeti egységre utaló kategóriasemleges és kategóriaspecifikus kifejezéseket nem talál a levélben, úgy megvizsgálja, hogy annak alternatívájaként Javaslat szerkezeti egységet talál-e. Amennyiben igen, úgy a kategóriaspecifikus kifejezések alapján kikalkulált panaszkategória megállapítása után a folyamat az Elvárás szerkezeti egység keresésével folytatódik. Amennyiben nem, úgy információkérő levél kerül kiküldésre az ügyfélnek, amelyben kérik, hogy tisztázza, hogy pontosan milyen ügyben fordult a minisztériumhoz.

Abban az esetben, ha a rendszer nem talál elvárás annotációt a levélben, vagy nem sikerül abból releváns információt kinyernie, a levél a valamelyik korábbi szerkezeti egység alapján megállapított panaszkategóriában jártas ügyintézőhöz kerül továbbításra.

Az ábra és a fentiek alapján is látható, hogy amennyiben egy levélből kinyerhető információ arra vonatkozóan, hogy az állampolgár milyen közigazgatási kategóriának megfelelő panasszal fordult az ügyfélszolgálathoz, úgy azt a gép az adott témában szakértő ügyintézőhöz juttatja el, aki válaszol arra a megszabott határidőn belül, ellenkező esetben pedig a rendszer automatikusan is generálhat egy információkérő levelet. Hogy milyen információ hiányzik a levélből, az megállapítható annak alapján, hogy az algoritmus milyen lépéseket járt be, mielőtt az információkérő levél ponthoz ért volna.

A kategorizálás forgatókönyv tesztelése eddig a kategóriasemleges definitív kifejezések figyelembevételével történt, azaz azt teszteltük, hogy az algoritmus, illetve a kategóriasemleges kifejezések listája alapján a gép milyen mértékben képes beazonosítani a kategorizáláshoz szükséges szerkezeti egységeket (Bemutatkozás, Cél, Probléma, Javaslat, Elvárás), és jut el az Ügyintéző pontig (lásd 5.1. ábra). Az eredmény, hogy a 888 panaszlevélből 156 levél esetében a levél az ügyintézőig jut, a többi esetben azonban valamelyik szerkezeti egység annotációjának hiányában az algoritmus információkérő levelet küld ki. Feltételezésünk szerint a számos információkérő levél küldésének egyik fő oka a szerkezeti egysé-

gekről szóló fejezetben is említett definitív kifejezések listájának a hiányossága. Elvásáraink szerint a lista bővítésével, pontosításával az eredemények jelentősen javulni fognak.

5.2. A profil

A profil kialakításához az összes szerkezeti egységet figyelembe kell venni, azaz a Megszólítást, az Elismerést, az Előzményt, az Egyéb körülményeket, a Vádaskodást, a Köszönetet és a Lezárást, valamint a problémakategória megállapításához figyelembe vett szerkezeti egységeket. Ezeknek a jelenléte, illetve a hiánya önmagában is árulkodó lehet, valamint egymáshoz viszonyított sorrendjük, arányaik is hordozhatnak fontos információkat. Ugyanakkor az ezekben az egységekben előforduló kifejezések stilisztikai jellemzői is értékesek lehetnek. Természetesen nem állítjuk, hogy egy pontos szociológiai, illetve pszichológiai profilt lehet ezek alapján az információk alapján felállítani az illető ügyfélről, azonban elvárásaink szerint bizonyos következtetések levonhatóak.

Az ügyfél aktuális (a levél írásának pillanatában fennálló) érzelmi állapotára következtethetünk a szótárunkban durva vagy bizalmas stílusúként jelölt kifejezések használatából. A levelek ilyen célú vizsgálata után azt mondhatjuk, hogy ha egy levél legalább egy durva vagy bizalmas stílusjegyű kifejezést tartalmaz (bármely szerkezeti egységben), akkor az ügyfél aktuális érzelmi állapota zaklatott. Amennyiben a levélben előforduló kifejezések legalább 0,5%-a, de legfeljebb 1%-a durva és/vagy bizalmas kifejezéseket tartalmaz, akkor az ügyfél erősen zaklatott érzelmi állapotban írta a levelet, ha pedig ez az érték 1% fölötti, akkor a levélíró aktuális érzelmi állapota szélsőségesen zaklatottnak tekinthető.

Az ügyfél szociológiai és pszichológiai profiljának felállítása a dialógusok kialakítása során lesz majd nagyon fontos, hiszen ha a használt kifejezésekből, a levél szerkezetéből, illetve tartalmi jellemzőkből tudunk következtetni az ügyfél életkörülményeire, iskolázottságára, vagy épp az aktuális érzelmi állapotára, az megszabhatja a kérdések és a válaszok formáját, illetve tartalmát egyaránt.

6. Összefoglalás

Jelen cikkben egy összetett cél elérése érdekében folytatott kutatás első eredményeit tárgyaltuk. Ezek közül az első egy egységes, robusztus előfeldolgozó keretrendszer és az ehhez kapcsolódó formátum elkészítése volt. Az eszköz és a formátum lehetővé teszi, hogy különféle, már korábban rendelkezésre álló nyelvfeldolgozó eszközöket, illetve a saját fejlesztéseinket egységesen kezeljük és a későbbiekben újabb komponensekkel egészítsük ki. Meglátásunk szerint az igazi értéke a rendszernek a szolgáltatásként igénybe vehető interfészekben, formátumokban rejlik. Hosszú távon tervezzük az implementáció UIMA [10] alapokra való helyezését.

A nyelvhasználattal, fogalmazással kapcsolatos problémák leküzdésében részeredményeket értünk el a szerkezeti egységek beazonosítása és a szűrésben való

felhasználásuk által. A szerkezeti egységek jelölése – amennyiben létrejön – kellően pontos. Azonban javítanunk kell még a felismerés hatékonyságán, amelyet a definitív kifejezések listájának bővítésével remélünk elérni.

A szerkezeti egységek felismerésének másik hozadéka az, hogy lehetővé teszik a levelek bizonyos részeinek elhagyását, és ezáltal megkönnyítik a kategorizálást. Ez a megközelítés túlmutat az egyszerű, egész korpuszra jellemző stopszólisták használatán, mivel ez a szűrés minden levélre külön-külön elvégezhető. Feltehetőleg a kategorizált levelek kis száma miatt a szűrés csak minimális javulást hozott a pontosság terén. Másrészt a különféle szűrési eljárásaink jelentős futási idő megtakarítást eredményeztek.

Fontos iránynak tartjuk a profilok építését a használt kifejezések alapján, valamint ezeknek a dialógusokban történő alkalmazását. Ezen a korpuszon alapvetően a profil szociológiai és pszichológiai dimenziójának felépítését tervezzük. A profilt nem csak egyszerű adatgyűjtési igények kielégítése miatt építjük. Fontos szerepet szánunk neki a levélíróval folytatott dialógus paraméterezésében: a kérdések és válaszok nyelvezetének meghatározásában, és az ügyfél várható reakcióinak megbecslésében. Ezen reakcióktól függővé tehetjük azt is, hogy felteszünk-e egyáltalán egy adott kérdést. A profilok segítségével szélsőségesen zaklatott felhasználók gyakran zavaros leveleit is felismerhetjük és megfelelően kezelhetjük.

Az itt bemutatott eredmények szándékaink szerint csupán az alapját képezik egy hosszabb kutatómunkának, amely során a rendszerünk kísérleti alkalmazását szeretnénk elérni. A munka része lesz más, eltérő tulajdonságokkal rendelkező korpuszok kipróbálása, valamint egy géppel támogatott megértést szemantikus keretek és hálók segítségével megvalósító eszköz elkészítése is.

Hivatkozások

1. Bárczi, G., Országh, L.: A Magyar Nyelv Értelmező Szótára (CD). Arcanum Adatbázis Kft (1994)
2. Héder, M.: Szemantikusan annotált dokumentumok létrehozása szövegfeldolgozó eszközök támogatásával (2009)
3. Hitec. (*categorizer.tmit.bme.hu/trac/wiki*)
4. Rott, H. In: Words in Context: Fregean Elucidations. Volume 23. (2000) 621–641
5. Kabai, D.: Automatikus tartalmi kódolás és osztályozás kidolgozása az igazságügyi minisztérium ügyfélszolgálatára beérkező állampolgári levelekre (2006)
6. 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)
7. Szafron, D., Greiner, R., Lu, P., Wishart, D., Macdonell, C., Anvik, J., Poulin, B., Lu, Z.: Explaining naive bayes classifications. Technical report (2003)
8. Busuttil, S.: Support vector machines (2003)
9. Pelleg, D., Moore, A.: X-means: Extending k-means with efficient estimation of the number of clusters. In: In Proceedings of the 17th International Conf. on Machine Learning, Morgan Kaufmann (2000) 727–734
10. Ferrucci, D., Lally, A.: Uima: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. **10** (2004) 327–348

A. További példák annotált levelekre

s 1. levél .
s Hozzá nem értésük már nagyon dühítő .
s Mit képzelnek maguk tulajdonképpen Mi a maguk munkája .
s Törvény szerint legkésőbb 30 napon belül válaszolni kell .
s Megmagyaráznák .
s Majd segítek .
s Amit maguk már művelnek egyenlő a Szervezett . bűnözéssel . .
s Maguk nagyon sok kárt okoznak .
s Ez már egyértelmű Bűnpártolás .
s Kértem két fontos címet 3 napon belül itt legyen vagy kénytelen leszek magukat is feljelenteni szándékos károkozásért .
s Nem lopom a pénzemet .
s Az az egyetemeken szokás .
s . Bűnöző . disznók miatt én nem fogom veszni hagyni a pénzemet ami jogosan jár .
s Mit képzeltek ti .
s Köztörvényes gazemberek Az erdőkből kitakarodni .
s Maguk csak . . hülyék . .

s 95. levél .
s Igazságügyi Minisztérium Budapest V. ker. Kossuth tér 2/4 Tisztelt Name Miniszter Úr Tisztelettel Kérem a Miniszter Úr Segítségét hogy a boszorkány ügyembe segítsen panasszal élek a Name ellen mert felbérelt kettő boszorkányokat ellenem lakása Place utca 22 szám nagyon rossz így élni ettől az orr viszketéstől nagyobb fájdalmat és szenvedést nem lehet okozni mert ez a legidegesítőbb az egész világon azért csinálják a boszorkányok élvezik azt hogy szenvedést okoznak ez a boszorkány ügyem nehéz úgy mert nem tudom bizonyítani és nem lehet mert nem ismerem azt a kettő boszorkányokat akik üldöznek csak a Name ismeri őket .
s Tisztelettel kérem azt a hivatali személyt aki az ügyemmel foglalkozik hogy próbálja valahogyan . szóra . bírni .
s A Name hogy mondja meg a kettő boszorkányoknak nevüket és . lakcímüket . még akkor is ha letagadja és hazudik és nem ismeri el azt hogy boszorkányokkal üldöztet engem .
s Tisztelt Minisztérium Tudatom önnel Miniszter úr és önökkel azt is hogy leveleket írtam a Kék Fény Szerkesztőségének a Magyar Rádió Szerkesztőségének a TV RTL Klub Fókusz Szerkesztőségének és annyi Tiszteletet nem érdemlek meg hogy a panasz leveleimre válaszoljanak nyilvánosság elé is akartam vinni az ügyemet de nem sikerült mert a három Szerkesztőségek nem veszik komolyan az ügyemet tudom hogy tele vannak panaszos ügyekkel de azért válaszolhattak volna csak egyedül az Ombudsman válaszolt a levelemre és . nyilvántartásba . vették

Magyar szövegek véleményanalízise[1]

Szaszkó Sándor[1], Sebők Péter[1], Kóczy T. László[1, 2]

[1] Budapesti Műszaki Egyetem, Távközlési és Média Informatikai Tanszék
1117, Budapest, Magyar tudósok körútja 2.
{Szaszko, Sebok, Koczy}@tmit.bme.hu
2 Széchenyi István Egyetem, Jedlik Ányos Gépész-,
Informatikai és Villamosmérnöki Intézet
9026, Győr, Egyetem tér 1.

Kivonat: A témaalapú osztályozásokban ismert módszerek hatékonyságát mutatjuk be a dokumentumok orientációjának eldöntésére. Ehhez összeállítottunk 240 dokumentumos tanító korpuszt. Az angol eredményekhez hasonlóan a klasszikus megoldások közül az SVM a leghatékonyabb, de ennek a teljesítményén is javít az eddig e célra nem használt RRM osztályozó. A Fuzzy-IDF súlyozás bevezetésével a kis felidézésű régióban a pontosságot tovább javítottuk.

1 Bevezetés

A dokumentumosztályozási feladat az egyik legismertebb szövegbányászati kutatási terület. Megoldásával hosszú idő óta foglalkozik a tudományos közösség, mára számos ipari alkalmazás igen jó hatékonyságú eredményt ad. A megoldások hátterében a legerősebb faktor, hogy az osztályokra jellemző szó halmazt gépi tanulási módszerrel közelítjük.

A véleményanalízis egy olyan két kategóriás osztályozási feladat, ahol a dokumentumok témája azonos. A különbséget a szöveg és a téma viszonyában keressük. Kutatásunk során filmekről szóló kritikákat elemeztünk, célunk a kritika pozitív vagy negatív beállítottságának eldöntése volt. A feladat nehézségét jól mutatja, hogy – bár jelentős előnnyel járna – a pozitív vagy negatív minősítés számszerűsítésére a szakirodalomban nem találtunk példát.

A téma a legújabb kutatási területekhez tartozik. Magyar nyelvű szövegek véleményanalízisével – legjobb tudásunk szerint – eddig csak egy szerzőpáros foglalkozott, Berend és Farkas jellemzően rövid, egymásra reagáló fórumbejegyzések alapján eredményesen jósolta a részvevők véleményét egy választási referendumról [9]. A mi vizsgálatunk tárgyát képező, egymástól független, hosszabb szövegek vizsgálata egészen más eszközök bevetését igénylik.

[1] OTKA K75711 számú támogatási szerződés keretében végzett kutatás.

1.1 Irodalmi eredmények

Hatzivassiloglou és McKeown (1997) melléknevek orientációjának meghatározását végezték el, majd ezek előfordulásának függvényében döntöttek [1]. Az általuk javasolt módszer képes arra, hogy a dokumentumokból kinyert melléknevek orientációját bizonyos halmazon 78%-ot meghaladó pontossággal becsülje meg. Szavak orientációját használja még [2] és [3] is.

Pang és társai sok további kutatásnak adtak irányt, amikor bebizonyították, hogy a gépi tanulási módszerekkel jobb eredmény érhető el, mint a priori módszerrel [4]. Naive Bayes, Maximum Entrópia és szupport vektor gép (SVM) módszerek teljesítményét hasonlították össze, ahol az SVM-et találták a leghatékonyabbnak. Ehhez hasonlóan a termékkritikák minősítésével foglalkozó [5] eredményei is az SVM (76%) elsőbbségéről tanúskodik.

A legjobb eredményeket a kivonatolás és gépi tanulási módszerek kombinációjával érték el. A módszer lényege, hogy a szövegeknek csak a szubjektív tartalmú mondatait használjuk fel, ezek alapján építjük a szeparációt végző modellt. A kétlépéses metódussal 86,4%-os eredményt értek el a korábban is említett angol mozikritika adatbázison [6]. Sajnos jelentős hátrány jelent, hogy a szubjektív mondatok kereséséhez nagyméretű példa mondatbázist kell felépíteni.

2 A korpusz

Véleményelemzés mindig egy jól behatárolható központi témakör köré épülő szövegvilág alapján történhet (témák pl.: politika világa, banki szolgáltatások, színházi előadások stb.). Külföldi gyakorlatot követve központi témakörnek a mozifilmek világát választottuk.

Az általunk épített polaritás adatbázis olyan magyar nyelvű kritikákat tartalmaz, amelyeknek a témája a műsorra tűzött mozifilmek tisztán szöveges tartalmi minősítése, nem pedig valamilyen meghatározott skála alapján vett kategorizálása (pl. „ötcsillagos” értékelés stb.). A szöveges értékelések döntő részét a port.hu, illetve az index.hu gyűjtőportál témát érintő moduljairól válogattuk össze. Az elemzési célnak megfelelően az összeállított tanuló-tesztelő polaritás adatbázis két kategóriából tevődik össze: egyik osztály a negatív (NEG), míg a másik a pozitív (POS) kritikákat tartalmazza.

A megépítendő korpusznak mennyiségi és minőségi kritériumoknak is eleget kellett tennie. A végső cél a korpusz méretét illetően az volt, hogy legalább 120 pozitív és ugyanennyi negatív kritikát fel tudjunk használni a módszerek vizsgálatához. A későbbi összehasonlítás reményében a szükséges anyagok összeválogatásánál törekedtünk a külföldi kutatásokban felhasznált angol nyelvű kritika-gyűjteményhez[2] hasonlatos korpusz felépítéséhez. Az általunk megfogalmazott minőségi kritérium szerint próbáltunk eleget tenni annak az elvárásnak, miszerint stílusában, méretében is olyan kritikákat válogassunk össze, mint amilyenek az angol nyelvű korpuszban is találhatók.

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

Az összeválogatott korpusz összesen 240 kritikát tartalmaz. A korpusz kiegyenlített a két kategória méretét illetően, továbbá az egyes kritikák hossza átlagosan 250 szó. A korpuszt a stopszó[3] szűrés révén 29181 darab szóból álló szótár jellemzi.

A polaritás korpusz építése igen időigényes és fáradságos munkát igénylő folyamat. A kritikák kézzel történő annotációja mellett szűk keresztmetszetet jelentett számunkra, hogy viszonylag csekély számú forrásból gyűjthettünk mintákat, ugyanis kevés magyar nyelvű kritika hozzáférhető az interneten. A külföldi kutatóknak több lehetőségük volt korpuszépítésre, mivel jóval nagyobb angol nyelvű adatbázis állt rendelkezésükre, mint amilyen az imdb.com is.

3 Szózsák modell, fuzzy-IDF bevezetése

A jelenlegi számítási kapacitások szövegbányászati feladatok megoldását leginkább csak szózsák alapú dokumentum reprezentáció esetén teszik lehetővé. A vélemény analízis esetén fel kell vállalnunk azt a tetemes információ veszteségét, amit a szavak sorrendje tartalmaz.

A veszteségek mérséklésére alakították a különböző súlyozási sémákat, melyek különböző módon veszik figyelembe

- szó előfordulásának számát
- dokumentum méretét
- dokumentum csoport, korpusz számosságát

A legáltalánosabban használt mérték a TF-IDF, amely a szó-dokumentum mátrix egyes értékét a következő módon állítja elő:

A *TF(t,d)* kifejezés egy adott szó (*t*) előfordulási gyakoriságát adja meg a vizsgált dokumentumban (*d*):

$$TF(t,d) = \frac{c_{t,d}}{\sum_i c_{i,d}}$$

ahol $c_{i,d}$ (count) az i-edik szó előfordulásának száma a *d* dokumentumban. A kifejezésből adódik, hogy a súlyozás a dokumentumvektorokat egységnyi hosszúságra normálja.

Az IDF súlyozás csökkenti a korpuszban a nagyobb támogatottságú szavak súlyát, míg a kevesebb dokumentumban előforduló szavak súlyát növeli:

$$IDF(j) = \log\left(\frac{N}{DF(j)}\right)$$

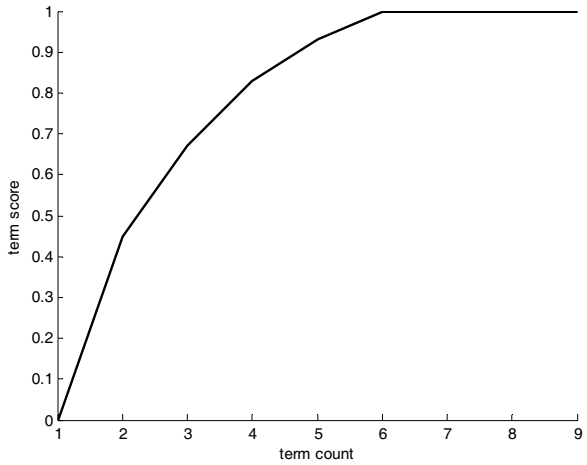
ahol *N a* korpusz dokumentumainak száma, míg *DF(j)* a j-edik szó *támogatottsága* a korpuszban (megadja, hogy a vizsgált szó hány dokumentumban szerepel).

Könnyen belátható, hogy más mértékben módosítja a dokumentum és a szó összetartozását, ha egyről kettőre nől az előfordulás, mint ha 5-ről 10-re. A TF súlyozás esetén ugyan az a hatás jön létre (duplázódik az érték). Ennek a megközelítésnek a logikájában vezettük be a fuzzy súlyozási sémát.

[3] A stopszó lista 743 darab szóból áll.

$$FS(j,d) = sigm(c_{j,d})$$

Az *FS(j,d)* kifejezés a j-edik szóhoz rendel értéket az ábrán látható módon.



1. ábra. Fuzzy súlyozási séma.

A sigmoid függvény telítődő jellege világosan kifejezi, hogy a vizsgált *d* dokumentumban az adott *j* szóhoz rendelt fontosság mértéke nem egyenesen arányos a szó dokumentumbeli előfordulásával (term count). A transzformáció praktikusan a kérdéses szó dokumentumhoz tartozásának erősségét fejezi ki a szóhoz rendelt fuzzy tagsági értékkel. Fuzzy reprezentációval szélsőséges esetben - sigmoid helyett lépésfüggvényt alkalmazva - a szavak *presence* információját kaphatjuk meg.

4 Robosztus Kockázat Minimalizáló alkalmazása

Újítással éltünk az osztályozó kiválasztásakor. A véleményelemzés témakörben ismereteink szerint eddig nem vizsgált megközelítést, a robosztus kockázat minimalizáció (RRM, Robust Risk Minimization) elvét alkalmaztuk a dokumentumok polaritás alapú osztályozásának megvalósítására [9][10][11][12].

Az elv gyakorlati jelentősége röviden összefoglalva abban áll, hogy az osztályozást meghatározó hipersík paramétereit egy *redukált keresési térre* korlátozva határozza meg az eljárás. Egy regularizációs paraméter révén szűkíthetjük a keresési tér méretét, ami egyúttal szabályozza a túltanulásra való hajlam érvényesülését is. Robosztusabbá válik tehát az osztályozó a tanulómintákon jelentkező túlilleszkedéssel szemben, amellett, hogy egyidejűleg minimalizálja a minták rossz osztályba sorolásának kockázatát.

Az általunk implementált algoritmus eltéréseket mutat a [11]-ben leírtaktól, mivel az értelmezése során az említett publikációban néhány cikkben – valószínűsíthetően elírást – fedeztünk fel. Sajnos a szerzőkkel nem sikerült felvennünk a kapcsolatot, de

az általunk megvalósított algoritmus eredményei igazolják az eljárásról alkotott elképzelésünk helyességét.

A fejezetben RRM rövid bemutatása mellett vázoljuk azt az értelmezést, amelyre támaszkodunk a dokumentumok polaritás alapú osztályozásában, illetve amely későbbi módosításaink alapját is képezi. A dolgozatban később ismertetésre kerülő kísérleteinkhez is a fejezetben ismertetett algoritmust és módosított változatait alkalmaztuk.

4.1 Az osztályozási feladat modellje

A kiindulási feladat megegyezik egy szokványosnak tekinthető szövegosztályozási feladattal.

A dokumentumainkat n elemű bináris feature (szó) vektorokkal ($\underline{x}=[x_1,..,x_n]$) reprezentáljuk, ahol n a szótár mérete, a szótárban lévő j-edik terminushoz (x_j) *bináris* értéket rendelünk aszerint, hogy a kérdéses szó előfordul-e a vizsgált dokumentumban vagy sem.

Feladatunk eldönteni azt, hogy a dokumentumvektor mely kategóriába tartozik. A becslést kizárólag annak alapján végezzük el, hogy a vizsgált dokumentum mely terminusokat tartalmazza, illetve melyeket nem. Az egyes osztályokat ($c \in C$) relevanciájuk szerint rangsoroljuk. A legrelevánsabb kategória (c^*) lesz a vizsgált dokumentum osztálya. A megoldást a maximum a posteriori hipotézis adja:

$$c^* = \arg\max_{c \in C} \{P(c \mid \underline{x})\}$$

A becslési feladat matematikailag a Naive-Bayes formulával fogalmazható meg, amely a dokumentumok szózsák alapú reprezentációjára épül. A terminusok előfordulásának bináris ábrázolásával a formulát átírhatjuk a következő formára:

$$\Pr(c \mid \underline{x}) = \frac{\Pr(c)\prod_j \Pr(x_j = 0 \mid c)}{\Pr(\underline{x})} \prod_j \left(\frac{\Pr(x_j = 1 \mid c)}{\Pr(x_j = 0 \mid c)} \right)^{x_j}, \quad c \in C, \quad x_j \in \{0,1\},$$

Ha vesszük a jobb oldali kifejezés természetes alapú logaritmusát, a következő formában is felírhatjuk az osztályozási problémát:

$$\Pr(c \mid \underline{x}) = \frac{1}{\Pr(\underline{x})} \exp\left(\overbrace{\sum_j w_j x_j + b}^{hipersík} \right)$$

A formulában felfedezhető a hipersík egyenlete, ahol a sík paramétereit a következő összefüggések adják:

$$w_j = \ln \frac{\Pr(x_j = 1 \mid c)}{\Pr(x_j = 0 \mid c)}$$

$$b = \ln \Pr(c) + \sum_j \ln \Pr(x_j = 0 \mid c)$$

A $\underline{w}=[w_1,...w_n]$ súlyvektor a tanulómintákból becsülhető, és az adott osztályozási problémát jellemző leíró. A súlyvektor w_j együtthatója azt fejezi ki, hogy az x_j szó

mennyire jellemző a vizsgált kategóriára. Annak az esélyét (odds[4]) fejezi ki, hogy a kérdéses szó a *c* osztályhoz tartozik. A fenti kifejezésekből lehetőségünk van visszavezetni a becslési feladatot a súlyvektor közvetlen meghatározásának problémájára. A szakirodalom arról a tapasztalatról számol be, miszerint Naive Bayes esetében jobb eredményeket lehet elérni a becslésben, ha nem max likelihood alapján számoljuk az osztályok relevanciáját, hanem közvetlenül a szavak súlyait próbáljuk meghatározni valamilyen lineáris döntési modellel [12]. A becslési feladat ilyen megközelítése a lineáris osztályozó módszerek egyik másik értelmezéséhez vezet: a lineáris súlyozás alapú módszerekhez.

4.2 Lineáris súlyozáson alapuló osztályozás – a dokumentum polaritása

Az osztályozási probléma újszerű megközelítésével tehát a szavak súlyozása révén kifejezhetjük, hogy a kategorizálás szempontjából milyen mértékben meghatározóak az egyes szavak. A stratégiát alkalmazhatjuk véleményanalízisre is.

Polaritás alapú osztályozás esetén koncepcionálisan két osztályt alakítunk ki: egyik halmazban a negatív (C_{neg}), míg másikban a pozitív (C_{pos}) véleményt hordozó kritikákat tároljuk. A koncepcióból eredően és (36) alapján tehát a szavakhoz rendelt súlyokra úgy tekintünk, hogy azok a szó által kifejezett vélemény orientációjának a mértékét fejezik ki. Az elgondolásunk alapján tehát a szóhoz rendelt súlyt megkapjuk:

$$w_j = \ln \frac{\Pr(x_j = 1 \mid C_{pos})}{\Pr(x_j = 0 \mid C_{pos})} \quad ahol \quad \Pr(x_j = 1 \mid C_{pos}) = \frac{\#d}{\left|C_{pos}\right|}$$

A kifejezésben $\#d$ azon pozitív kritikák számát adja meg, amelyekben az x_j szó előfordul. A súlyok előjele adott polaritású orientációt kapcsol a szóhoz, a súly nagysága a szó által kifejezett vélemény polaritásának erősségét fejezi ki. Az osztályozás során az ismeretlen polaritású dokumentum által képviselt eredő orientációt a dokumentum szövegében előforduló szavakhoz rendelt súlyok összegeként határozzuk meg:

$$polarity\_score(\underline{x}) = \sum_j w_j x_j + b = \underline{w}^T \underline{x} + b$$

A dokumentumra adott eredő súly polaritása határozza meg a teljes szöveg orientációját.

A fenti kifejezés rávilágít az osztályozási feladat egy más megközelítésű értelmezési lehetőségére, miszerint a vizsgált dokumentum alapján az egyes osztály címkék rangsorolása közvetlenül a dokumentumban lévő szavakhoz rendelt súlyok lineáris kombinációjával is meghatározható egy helyesen becsült $\underline{w}$ súlyvektor ismeretében. A feladatunk tehát az, hogy a tanulóminták alapján megbecsüljük a helyes döntéshez szükséges súlyvektort.

Korábbi fejezetben már ismertetésre került néhány lineáris döntési modellt megvalósító algoritmus. Korábbi kísérleteink azt támasztották alá, hogy érdemes regularizált

4 Odds: angol szakirodalomban terjedt el, jelentése: p/(1-p).

osztályozókkal kísérletezni a modell paramétereinek meghatározásában. A súlyvektor meghatározásához alkalmazott RRM algoritmus alapját T. Zhang és F. J. Oles által kidolgozott keretrendszer alkotja [9].

Az Information Retrieval folyóiratban megjelent tanulmányuk arra keresi a választ, hogy az SVM dokumentumok osztályozásában nyújtott teljesítménye vajon csak az SVM tervezés sajátossága-e, vagy talán alkotható egy olyan egységes matematikai keretrendszer („Regularized Linear Systems"), amelyet alkalmazva más lineáris osztályozók esetében is jó teljesítmény lenne elérhető. A keretrendszer meghatározza a regularizált osztályozási feladatok megoldásához vezető utat a feladat megfogalmazásától kiindulva. A megoldáshoz numerikus módszereket is ajánlanak, amelyek a szövegbányászat nagydimenziós terében képesek hatékonyan megoldani a feladatot.

4.3 Robosztus kockázat minimalizálás elve

Az RRM algoritmus a regularizált lineáris osztályozók csoportjába tartozik. A következőkben az algoritmus ismertetése mellett egyúttal betekintést nyújtunk a regularizált osztályozók alapjaiba is.

Az algoritmussal való ismerkedésünk kiindulási pontja az osztályozási feladat megfogalmazása. Felügyelt tanulási módszerről lévén szó tanulómintákat $\{(x, y)\}_{i=1}^{N}$ alkalmazunk a modellépítési fázisban. A tanulási feladatban a korpuszt alkotó dokumentumok vektoros reprezentációja ($\underline{x}_i$) bemeneti, míg a dokumentumokhoz rendelt osztály azonosítója ($y_i \in \{-1,+1\}$) kimeneti változóként jelenik meg. Az osztályozási feladat koncepcionálisan a következő kényszerekkel fogalmazható meg:

$$\left(\underline{w}^T \underline{x}_i + w_0\right) y_i > 0 \quad \forall i \quad (39)$$

$$\left\|\underline{w}\right\|^2 + w_0^2 \leq A \quad (40)$$

Ahol a (39) feltétel biztosítja, hogy minden $\underline{x}_i$ dokumentum megfelelően legyen osztályozva, míg a (40) regularizációból adódó kényszer korlátozza a lehetséges hipersík paraméterek keresési terének a méretét. A megfogalmazott feladat értelmében tehát keressük a lineáris döntési modell azon paramétereit ($\underline{w}, w_0$), amelyek kielégítik az előírt feltételeket.

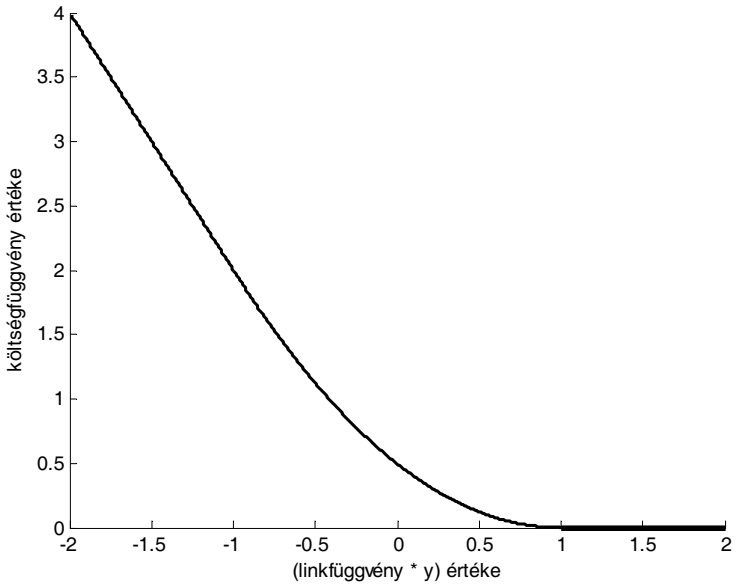
A keresési térben a legjobb modell megtalálásához költségfüggvényt alkalmazunk. A költségfüggvény révén matematikailag kezelni tudjuk az osztályozási problémát, ellenőrizhetjük a modellünk illeszkedését[5], azaz mérhetjük, hogy a modell milyen jól írja le a mintáinkat. Célunk a modell paramétereinek meghatározása úgy, hogy közben az általunk használt költségfüggvényt minimalizáljuk. A függvény optimalizálása általában valamilyen iteratív numerikus módszerrel végezhető, ahol a függvény mentén történő minimalizálás során történik a modellünk lépésenkénti finomítása. Az optimumot eredményező pontban kapjuk azt a modellt, amely a legjobban megfelel a függvény által kifejezett elvárásainknak (osztályozási hiba legyen minimális).

[5] Büntetjük a modell pontatlanságát.

Célunk meghatározni azokat a $(\underline{w}, w_0)$ paramétereket, amelyekre $y_i = \phi(\underline{w}, w_0, \underline{x}_i) \quad \forall i$, ahol $\phi(.)$ az úgynevezett „link függvény” (esetünkben a hipersík matematikai kifejezése). Adott modellparaméterek mellett az illeszkedés mértékét a $L(.)$ költségfüggvénnyel (loss function)[6] határozzuk meg. Numerikus okokból kifolyólag a link függvényt úgy választjuk meg, hogy a keletkező költségfüggvény $L\left(\phi(\underline{w}^T, w_0, \underline{x}_i), y_i\right)$ konvex legyen.

A költségfüggvény révén büntetjük az osztályozás hibáját. A súlyvektort a bemeneti mintákból határozzuk meg az illeszkedés *várható* költségének minimalizálása mentén. A megoldáshoz a következő költségfüggvényt használjuk RRM esetén [12]:

$$L\left(\phi(\underline{w}^T, w_0, \underline{x}_i), y_i\right) = \begin{Bmatrix} -2\phi(\underline{w}^T, w_0, \underline{x}_i) y_i & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i < -1 \\ \frac{1}{2}\left(\phi(\underline{w}^T, w_0, \underline{x}_i) y_i - 1\right)^2 & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i \in [-1,+1] \\ 0 & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i > 1 \end{Bmatrix} \quad (41)$$



2. ábra. Robosztus költségfüggvény.

A mintákra illeszkedő optimális döntési modell paramétereit a robusztus költségfüggvény minimumában kapjuk meg.

6 Magyar szakirodalomban a veszteségfüggvény elnevezés is használatos.

4.4 RM algoritmus pszeudokódja

Előfeldolgozás: bináris szó-dokumentum mátrix generálása a korpuszból

$$\underline{x}^j = [x_1^j,...,x_m^j] \quad x_k^j = \begin{cases} 0 & \text{k. szó} \notin \text{j. dokumentum} \\ 1 & \text{k. szó} \in \text{j. dokumentum} \end{cases}$$

Bemenet: tanuló minták $(\underline{x}^1, y^1),.....,(\underline{x}^N, y^N)$

Paraméterek: K, A, η

Kimenet: súlyvektor $\underline{w} = [w_1,....,w_m]$, w_0

Inicializálás: $\alpha_i = 0$ (i = 1...N), $\underline{w} = \underline{0}$, $w_0 = 0$

for k = 1 to K **do**
  for i = 1 to N **do**
    p = $(\underline{w}^T \underline{x}^i + b) y^i$
    gradient$_i$ =
      $\max(\min(2A - \alpha_i, \eta((A - \alpha_i)/A - p)), -\alpha_i)$ (*)
    $\underline{w} = \underline{w} + \text{gradient}_i \underline{x}^i y^i$
    $w_0 = w_0 + \text{gradient}_i y^i$
    $\alpha_i = \alpha_i + \text{gradient}_i$
  end for
end for

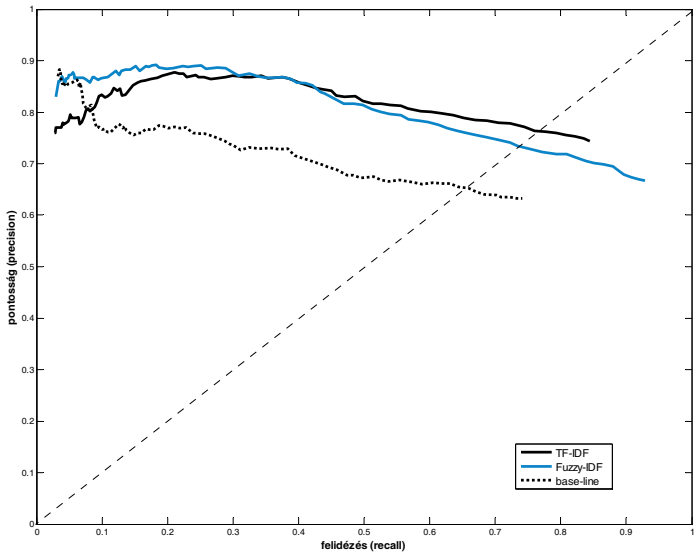
3. ábra. Az RRM algoritmus pszeudokódja.

A pszeudokódban előforduló változók jelentése a következő: *K* paraméter az iteratív algoritmus maximális lépésszáma, *A* paraméter a keresési tér méretét korlátozza ($A = \frac{1}{\lambda N}$), a η paraméter a tanulási ráta, amely az iteráció során a gradiens irányába tett lépésünk nagyságát határozza meg. Az online módszer értelmében a duális α_i változók egy-egy tanulóminta párashoz kapcsolódnak. A gradiens kifejezésénél (*) a maximálás, illetve a minimálás biztosítja, hogy a duális változó az előírt $\left[0, \frac{2}{\lambda N}\right]$ intervallumban maradjon. Ha egy bemeneti $\underline{x}_i$ mintához tartozó duális változó értéke meghaladja az előírt intervallumot, akkor a változót nullázzuk.

5 Eredmények

Méréseink elsősorban arra irányulnak, hogy megvizsgáljuk a szó-dokumentum mátrixon alkalmazható különböző súlyozási séma osztályozás pontosságára gyakorolt hatását az RRM esetén. A tanuló modellt a korábbi fejezetben felsorolt sémák alapján súlyoztuk, majd az algoritmus korpuszra hangolását és tanítását követően 50 mérés-

ből álló tesztsorozat átlagából meghatároztuk az alábbi ábrán látható felidézés – pontosság görbéket.



4. ábra. Felidézés - pontossággörbék magyar korpuszon

Vizsgálataink fókuszában elsősorban a zérus döntési küszöbhöz tartozó felidézés - pontosság értékpárok álltak, vagyis az osztályozás kiértékelését az *összes dokumentumra* hozott döntés figyelembevételével végeztük. Ezen értékpárokat mindig a vizsgált súlyozási sémára vonatkozó görbe maximális felidézés értékéhez tartozó pontosság értékének kettőse alkotja.

Az eredményül kapott görbékből leolvasható, hogy a base-line módszer (pontozott vonal) alkalmazása esetén közel 64%-os pontosság mellett mindössze a POS kritikák 74%-át találtuk meg. Az alacsonyabb felidézési érték mellett az elért pontosság kedvezőtlen hatású az osztályozás találati arányára nézve. Gyakorlatilag azt jelenti, hogy a becslés során a negatív kritikák egy jelentős részét is hibásan becsülte az algoritmus, miközben a pozitív kritikák közel háromnegyedét becsülte csak helyesen.

Az ábrán az is látható, hogy a base-line módszer felidézésén a szó-dokumentum mátrix Fuzzy-IDF súlyozásával nagymértékben sikerült javítani. Az ábrázolt görbe alapján megállapítható, hogy a tesztmintákban a POS kritikák 92,8%-át megtaláljuk közel 67%-os pontosság mellett. Fuzzy-IDF súlyozás alkalmazásával a tesztdokumentumokon elérhető pontosság értéke szinte változatlan maradt ugyan, de a nagyobb felidézés érték azt igazolja, hogy képesek vagyunk szinte az összes pozitív kritikát megtalálni a mintahalmazban.

Az ábra tanulsága szerint a szó-dokumentum mátrixon kipróbált súlyozási módszerek közül a TF-IDF súlyozás bizonyul a leghatékonyabbnak. TF-IDF súlyozású mátrixon tanított algoritmus képes arra, hogy 74,34%-os pontosság mellett megtalálja a

teszthalmazban a POS kritikák több mint 84%-át. Megfigyelhető azonban, hogy az osztályozó pontossága alacsony felidézés mellett csökkent, ami arra enged következtetni, hogy a szeparáló felülettől távol lévő dokumentumok címkéjét pontatlanabbul becsüli az algoritmus.

Az eredmények alapján megállapítható, hogy a tanuló modell különböző súlyozásai révén sikerült javítani az eredeti base-line módszer magyar korpuszon elérhető hatékonyságán. Azt a következtetést vonhatjuk le, miszerint a különböző súlyozási konvenciókkal minden esetben pozitív irányban befolyásoltuk az algoritmust: a Fuzzy-IDF súlyozás hatására hasonló pontosság mellett jobb felidézést értünk el mint a base-line módszer. A tanuló modell TF-IDF súlyozása nagyban javít mind az elérhető pontosságon, mind a felidézésen, továbbá egyben a legnagyobb BEP értékű osztályozót eredményezi.

6 Összefoglalás

Korábbi munkák során az általunk készített magyar filmkritika korpuszon megvizsgáltunk több osztályozó módszert. Ezek illetve a legjobban teljesítő, az e tanulmányban ismertetett RRM módszer eredményeit mutatja az 1. táblázat.

Külön vizsgáltuk a „nem" jelentésmódosító hatását. Pl. „nem jó" szereplése esetén a „nem" stop szót figyelmen kívül hagyjuk és csak a „jó" kerül be a szózsák modellbe. A táblázat utolsó sorában szereplő „NOT TAGGING" esetben a „nem" szó és az általa módosított szó együtt képez tókent.

1. táblázat: Eredmények magyar korpuszon.

	Naive Bayes	Percep-tron	neurális hálózat	SVM	RRM
helyes osztályozási arány	0.63	0.65	0.697	**0.715**	**0.76**
helyes osztályozási arány (NOT-TAGGING)	–	0.645	0.662	0.703	–

A NOT-tagging módszer láthatóan nem segíti a magyar filmkritika korpusz véleményanalízisét.

A magyar nyelvű véleményanalízisre újszerűen alkalmazott RRM az általunk javasolt fuzzy-IDF súlyozással jelentős javulást hozott az eddigi legjobb SVM-mel szemben is.

Módszerünk az angol korpuszon 78,8%-ot ér el, ami hasonló az 1.1-ben olvasható eredményekhez.

Hivatkozások

1. Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL. Madrid, Spain, July 1997. Association for Computational Linguistics (1997) 174–181
2. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) 2003, 21 (4) (2003) 315-346
3. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis. In: Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, Bremen, DE (2005) 617-624
4. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002 (2002)
5. Na, J-C., Khoo, C., Horng Jyh Wu, P.: Use of negation phrases in automatic sentiment classification of product reviews. Library Collections, Acquistions & Technical Services, 29 (2005) 180-191
6. Pang, L., Lee, A.: Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL 2004 (2004)
7. Damerau, F. J., Zhang, T., Weiss, S. M., Indurkhya, N.: Text categorization for a comprehensive time-dependent benchmark. Information Processing & Management, 40 (2004) 209-221
8. Berend, G., Farkas, R.: Opinion mining in Hungarian based on textual and graphical clues. In: Proceedings of the 4th Intern. Symposium on Data Mining and Intelligent Information Processing, Santander (2008)
9. Zhang, T., Oles. F. J.: Text categorization based on regularized linear classification methods. Information Retrieval, 4 (2001) 5-31
   http://www-cs-students.stanford.edu/~tzhang/papers/ir01_textcat.pdf
10. Zhang, T.: On the dual formulation of regularized linear systems. Machine Learning 46 (2002) 91-129 http://www-cs-students.stanford.edu/~tzhang/papers/ml02_dual.pdf
11. Damerau, F. J., Zhang, T., Weiss, S. M., Indurkhya, N.: Text categorization for a comprehensive time-dependent benchmark. Information Processing & Management, 40 (2004) 209-221 http://www-cs-students.stanford.edu/~tzhang/papers/ipm04-new_reuters.pdf
12. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.: Text Mining - Predictive Methods for Analyzing Unstructured Information. Springer, ISBN: 978-0-387-95433-2 (2005)
13. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publisher Inc., ISBN: 978-1-60198-150-9 (2008)

Az [origo] automatikus címkézési projekt tapasztalatai

Farkas Richárd

MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: A cikkben bemutatjuk az [origo] hírportál archívumának automatikus címkézésére irányuló projektet. Címkézés alatt azt az eljárását értjük, ami az egyes dokumentumokhoz egy olyan kifejezéshalmazt rendel, amely annak tartalmát jól reprezentálja. A cikkben bemutatásra kerülnek az újságarchívumok címkézésére vonatkozó irányelvek, az automatikus címkézési megoldásunk, az elért eredmények és tárgyalunk olyan nyitott számítógépes nyelvészeti problémákat, amelyek megoldása nagyban hozzájárulhat a címkézés sikerességéhez. Az [origo] archívumának automatikus címkézése manuális kiértékelés alapján a dokumentumok 77,5 százalékát megfelelőnek minősítette, ami meghaladta az eredeti célkitűzéseket.

1 Bevezetés

Az egyik legismertebb Web 2.0-ás technológia az úgynevezett *címkézés* (tagging), aminek keretében az internetes közösség tagjai címkéket rendelnek az elektronikus tartalmakhoz (blogbejegyzésekhez, képekhez, URL címekhez stb.) [1]. A címkék egy vagy néhány szavas természetes nyelvű kifejezések, amelyek célja általában az adott tartalom tömör leírása, jellemzése. Egy nagyméretű, címkézett adathalmazban a keresés, rendszerezés jóval hatékonyabbá válik mind a címkéző felhasználó, mind az egész közösség számára. Ezen felül az úgynevezett *címkefelhő* segítségével az egész adathalmaz mindenki számára azonnal értelmezhető tartalmi reprezentációja is megvalósítható. Napjainkban címke-hozzárendelések vagy címkefelhő(k) szinte minden közepes és nagy weboldalon megtalálhatók.

Az [origo] internetes hírportál 2009 márciusától vezette be cikkeinek manuális címkézését[1]. A portál ezt megelőzően már üzemeltette azt a szolgáltatást, amelyben a videókat[2] a felhasználók (látogatók) szabadon címkézhették. A videócímkézés legfőbb tapasztalatai azok voltak, hogy a címkék hasznosak ugyan, de mivel a felhasználók saját szemszögükből (szubjektív címkék, saját kategóriarendszer) rendelik hozzá a címkéket, azok gyakran nem alkalmasak az adott tartalom témájának azonosítására (hasonló következtetéseket von le [2] is). Ezen tapasztalatok alapján az [origo] híreinek címkézésére egy köztes megoldást vezetett be: a cikkeket a szerkesztők közössé-

[1] http://www.origo.hu/techbazis/internet/20090312-tagging-a-hirportalon-az-origo-bevezeti-a-cikkek-cimkezeset.html

[2] http://videa.hu

ge (körülbelül 50 fő) együttesen címkézi, a híreket annak szerzője látja el címkével. Azonban nincsen előre rögzített taxonómia, a címkézés teljesen szabad.

Egy hírportál számára több szempontból is igen hasznos a teljes híranyagának felcímkézése. Ezzel minden olyan témának, melyről gyakran írnak, önálló oldala lehet, tulajdonképpen automatikusan önálló rovatoldalak keletkeznek. Az archívum címkézése lehetőséget biztosít arra is, hogy az [origo]-ról amúgy eltűnt tartalmainkat újra elérhetővé tegyék és segít abban is, hogy a különböző oldalakon megjelenő, de tematikában megegyező tartalmak egy helyről legyenek elérhetőek, ezáltal növelve az egyes termékeink közötti keresztolvasottságot.

A címkék automatikus tartalmakhoz rendelése csak az utóbbi években vált intenzíven kutatott témává, mind a számítógépes nyelvészet, mind egyéb tudományágak (képfeldolgozás, zene/videó címkézés stb.) területén. Az [origo] manuális címkézésével egyidejűleg a Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportjával közösen elindult az archívum automatikus felcímkézését célzó projekt is. Az origo.hu oldalon 1998. december 1-jén jelent meg az első hír, és a kézi címkézés elindulásáig 380 ezer cikk látott napvilágot. Az archívumcímkézési projekt célja az volt, hogy ezen hírekhez automatikusan, azok tartalmát jól reprezentáló címkéket rendeljünk, oly módon, hogy az egész dokumentumhalmazhoz tartozó címkekészlet koherens legyen. A következőkben bemutatásra kerülnek a címkézés irányelvei, a megoldási mód váza, illetve tárgyalunk olyan nyitott számítógépes nyelvészeti problémákat, amelyek megoldása nagyban hozzájárulhat a címkézési probléma megoldásához.

2 Címkézési útmutató

A projekt kezdetén az [origo] munkatársai elkészítettek egy címkézési útmutatót, ami elsősorban a szerkesztők (manuális címkézés) számára tartalmazott irányelveket, de ezeket az automatikus címkézésnél is követtük.

Címkeadásnál a közvetlen cél nem az, hogy a cikk témáját hogyan tudjuk kulcsszavak segítségével absztrahálni, hanem az, hogy megtaláljuk azokat a címkegyűjtőoldalakat, amelyek alatt szeretnék a felhasználók a cikket viszontlátni. Címkét négy kategóriában lehet megadni (*téma*, *személy*, *intézmény*, *földrajz*), ebből egyedül a téma kategóriánál kötelező megadni legalább egy címkét. Személynévnek kell tekinteni az állatneveket és fiktív élőlények neveit is. Ha több közszereplőre is vonatkozhat egy személynév, mindig meg kell toldani egy olyan kifejezéssel, mely egyértelműen csak rá vonatkozik, például: „*Csányi Sándor színész*”. A földrajzi nevek megkülönböztetése igen fontos, ezáltal lehetővé válik a hírek mellé megjeleníteni azok geográfiai pozícióját és az olvasónak lehetősége nyílik a helyi közösségéhez kapcsolódó hírek közt keresni. Földrajzi nevek közé tartoznak az egész univerzum egységei (bolygók stb.) is.

Az entitásokkal kapcsolatosan általánosságban követendő, hogy csak akkor vehetjük fel őket címkeként, ha azok nemcsak eseti alapon kerülnek bele a nemzetközi/hazai hírfolyamba, hanem viszonylag rendszeresen kerülnek szóba, és meghatározóak a cikkben elmondottakkal kapcsolatban.

A téma kategóriába kerülhetnek elvont fogalmak, jelenségek (pl. „*koalíciós válság*”), sportágak, ligák, események (pl. „*Sziget-fesztivál*”), tudományos fogalmak, szakkifejezések és egyéb entitások (pl. márkanevek) amelyek leírják a cikk főbb témáját. A téma kategória alá felvett fogalmaknál csak olyan címkék adhatóak meg, melyek önmagukban jól definiálnak egy korszakot, helyzetet, viszonyrendszert. A címkéknek jól kell definiálniuk egy területet, érdeklődési vagy fogalomkört, azaz vannak olyan emberek, akik kifejezetten csak arról a témáról akarnak majd olvasni.

A címkék megfelelő számát a cikk témája határozza meg. Koncentrált témájú cikkeknél általában 2-4 címke elegendő, általános elemzések, átfogó cikkek esetében több, maximum 8-10 címkét is adhatunk. A címkék csak főnévi szerkezetek lehetnek, és kerülendő a szleng, a zsargon, az átvitt értelmű szavak, a metafora, a humor és a parafrázis.

Ezek az irányelvek számos ponton eltérnek a szokásosnak tekinthető címkézési követelményektől. Ilyen például a cikk fontos szereplőinek (személyek, szervezetek, földrajzi helyek) kiemelt szerepe, legfeljebb három szó hosszúságú címkék, rokon értelmű címkék konzekvens használata stb.

3 Kapcsolódó munkák

A hírarchívum-címkézési probléma több szempontból is újszerű. Egyrészt nem illeszthető a létező automatikus címkézési megközelítések közé, mert azok vagy egyetlen dokumentum kulcskifejezéseinek megtalálására törekednek (keyphrase extraction) [3, 4], vagy hasonló dokumentumok címkéit emelik át (tag recommendation) [5, 6]. Előbbi megközelítés a dokumentumból kiemeli a potenciális kulcsszavakat, majd azok közül kiválaszt néhányat úgy, hogy azok a dokumentum tartalmát lefedjék, de ne tartalmazzon redundáns elemet. Ez a megközelítés nem alkalmazható a mi esetünkben, mert csak egyetlen dokumentumra fókuszál, az archívum címkézésénél pedig kiemelt szempont az egész dokumentumhalmazon vett konzisztens címkézés.

A másik megközelítésben rendelkezésre áll egy kielégítő méretű címkézett dokumentumhalmaz, és a fókusz egy címkézetlen dokumentumhoz hasonló dokumentumok megtalálására irányul, ui. a hipotézis az, hogy a hasonló témájú tartalmakról a címkék egy az egyben átemelhetőek. Az ilyen rendszereket általában címkeajánlásra használják blogbejegyzésekhez, ahol rendelkezésre áll nagyszámú címkézett dokumentum, a blog szájt összes korábbi bejegyzése [5]. Ez a megközelítés sem alkalmazható közvetlenül a mi esetünkben, habár hozzáfértünk a szerkesztők által 2009 márciusa és májusa közt címkézett hírekhez, azok nem tartalmazhatnak minden hozzárendelendő címkét (a 2009-es hírek nem kapnak például „*Tocsik-ügy*” címkét).

Az archívumcímkézés folyamán a két bemutatott módszert ötvözve kell alkalmazni, ahol alapvetően a címkéket a szövegekből kell származtatni (ezáltal biztosítani az új témák, entitások felismerését), de tekintettel kell lenni az egész dokumentumhalmaz koherens címkézésére is (témájukban megegyező hírek kapjanak közös címkét).

Amellett, hogy – legjobb tudomásunk szerint – ez az első munka, ami egy hírportál automatikus címkézését célozta meg (annak specialitásaival), ez az első megoldás magyar nyelvű automatikus címkézésre is.

4 Automatikus címkézés

Az automatikus címkézés során azt a soros feldolgozást választottuk, hogy első lépésben kiemeljük a szövegben egzaktul előforduló potenciális címkéket, majd ezeket megpróbáljuk absztrahálni, ami újabb címkék felvételét eredményezi. Végül a címke jelöltek halmazát leszűkítjük egy megfelelő méretűre, és ezt tekintjük végleges címkézésnek.

4.1 Szövegbeli címkejelöltek gyűjtése

A címkézési útmutató alapján csak főnévi csoportok szerepelhetnek címkeként. Három különböző módon gyűjtöttünk főnévi csoportokat: automatikus tulajdonnév-felismeréssel, szófaji kódok alapján derivációval és szótárillesztéssel.

A tulajdonnevek automatikus felismerése és szemantikai kategorizálása (személynév, földrajzi név, szervezetnév, egyéb) felügyelt tanulási keretben tulajdonképpen megoldottnak tekintett (habár bizonyos esetekben a módszerek pontossága a 70%-ot sem éri el [7]). Azonban ha nem áll rendelkezésre megfelelő méretű, karakterisztikájában a jelölendő szöveggel megegyező tanító adatbázis, a pontosság drasztikusan csökken. Az [origo] hírei témájukat, karakterisztikájukat tekintve igen diverzek, ezért manuálisan annotálásra került az *autós*, *itthon*, *nagyvilág*, *sport*, *szórakozás* és *techbázis* kategóriák körülbelül 200-200 híre. Ezeken az adatbázisokon Conditional Random Fields-et[3] tanítottunk a korábban gazdasági hírekre kidolgozott jellemzőkészlet [8] felhasználásával. A felsorolt főkategóriákon kívüli hírek esetében az egész annotált dokumentumhalmazon tanított modell predikcióját használtuk fel.

A nagyméretű dokumentumhalmaz lehetőséget biztosított az automatikus tulajdonnév-felismerés hibáinak javítására (utófeldolgozására) és normalizálására. Hibajavítás alatt az automatikusan jelölt tulajdonnévi frázisok határainak korrekcióját értjük (azaz összeragadt tulajdonnevek szétbontását, hozzáragadt tokenek eltávolítását, illetve a határok kiterjesztését). A normalizáció elsődleges célja a tulajdonnevek szótövesítése volt – ami nem oldható meg a standard morfológiai elemzők segítségével, hiszen itt a lehetséges szótövek felsorolása nem lehetséges – (például a „*Pannon”* szótöve „*Pann”*?). Emellett egyszerű szabályokkal kísérletet tettünk a rövidítések feloldására és az egyes kifejezések egységes szemantikai kategorizálására (leggyakoribb szerep) is. Ez utóbbi csak egyes főkategóriákon belül értelmes, hiszen például a *Kecskemét* a sporthíreken belül általában szervezetként szerepel (mint egy csapat), míg a belföldi hírek esetében földrajzi entitásként. Ezen utófeldolgozási lépéseket a korpusz automatikus jelöléséből nyert statisztikák alapján végeztük el a [9]-ben bemutatott eljáráshoz hasonlóan. Itt a fő hipotézisünk az volt, hogy egy tulajdonnév ragozatlan alakjának gyakorisága szignifikánsan nagyobb, mint bármely ragozott alakjáé.

A cikkek témájának felismeréséhez főnévi csoportokat (NP) is gyűjtöttünk a szövegből. Ehhez kísérleteztünk a hunpars-szal [10], de azt találtuk, hogy egy POS-tagger eredményeit felhasználva egyszerűbben és kevésbé zajosan tudunk NP-ket kiemelni. A megoldás során NP-nek tekintettük az egyszavas főneveket, melléknév-

[3] implementácó: http://mallet.cs.umass.edu/

főnév párokat, főnévi birtokos szerkezeteket és az igéből és melléknévből képzett főneveket. Az igékből és melléknevekből történő főnévképzésre, valamint a főnévi birtokos szerkezetek összetett szavakká alakítására (például „*üzemanyagok ára*"-ból „*üzemanyagár*") egyszerű átírási szabályokat alkalmaztunk.

A tulajdonnevek és NP-k azonosítása távol van a tökéletestől, ezért külső tudásbázisok is beépítésre kerültek a rendszerbe. Külső tudásbázisnak használtuk a Wikipédia[4] szócikkeinek címeit és annak gyűjtőoldalait (amelyek címe „listája"-ra végződik). Az így nyert listákat illesztettük a szövegre a ragozási és tőhangváltási lehetőségek figyelembevételével.

A tulajdonnév-kinyerés, főnévicsoport-azonosítás és listaillesztés eredményeit a következőképpen aggregáltuk: az azonos helyről érkező – pontosabban átfedő – találatokat (például egy azonosított tulajdonnév a szótárban is szerepelhet) elhagytuk, hiszen az, hogy két módszer is azonosította, nem implikálja, hogy kétszeres súlyt kapjon. Végül a halmazt egy paraméterezett tfidf metrika felhasználásával sorba rendeztük. A metrika figyelembe vette azt is, hogy a vizsgált találat a dokumentum melyik zónájából érkezett (cím, összefoglaló, képaláírás stb.), illetve vonatkozik-e rá formázási információ (például dőlt, kiemelt). A tfidf optimális paraméterezését a rendelkezésünkre álló kézzel jelölt cikkek alapján határoztuk meg.

4.2 Absztrakt címkézés

A szövegben egzaktul előforduló címkejelölteken felül általában szükség van ún. absztrakt címkék felvételére is, amik a dokumentum tartalmát általánosabb módon írják le (például a „*kékfény*" szó általában nem szerepel a cikkekben). Az ilyen jellegű absztrakciók elvégzésére két módszert dolgoztunk ki. Az első módszer a Wikipédia linkstruktúrájának kiaknázásával, a potenciáliscímke-halmaz alapján gyűjt össze absztrakt címkéket (ez a megközelítés általánosságban kerül bemutatásra a [11] publikációban).

A másik módszerben felügyelt tanulási problémaként fogalmaztuk meg egyes címkék felvételének lehetőségét. Ehhez statisztikai jellemzők és szemrevételezés útján kiválasztottunk 243 darab absztrakt témát jelölő címkét és 243 osztályozási modellt építettünk, ami a dokumentumhoz rendelt potenciális címkék alapján (azokat használva jellemzőkészletként) hivatott eldönteni, hogy a szóban forgó címkével kelle bővítenünk a címkejelöltek halmazát. Első pillantásra ezeknek a nagyon absztrakt témákat jelölő címkéknek egyszerűen következniük kellene az adott dokumentum kategóriájából (pl. „*foci*" kategória). Azonban annak ellenére, hogy az [origo] kategóriahierarchiája több mint ezer elemet tartalmaz, ezek nem egyenszilárdságúak (vannak köztük, amelyek több tízezer hírt tartalmaznak) és ráadásul a hierarchia időben evolválódott. Például a *kosárlabda* kategória 2001-ben került bevezetésre, az 1998 és 2001 közt a *kosárlabdával* foglalkozó hírek a *csapatsport* kategóriába kerültek. Ezért úgy döntöttünk, hogy az általunk fontosnak ítélt magas szintű absztrakciót képviselő címkéket gépi tanulási módon keressük meg.

A tanításhoz pozitív példaként egy magasabb szintű kategórián belül azokat a dokumentumokat használtuk, amelyeknél a kérdéses címke szerepelt a potenciáliscím-

[4] http://hu.wikipedia.org

ke-halmazban. Negatív példaként az ugyanezen időszakból származó kategórián kívüli dokumentumok szolgáltak. A kategóriabeli megkötésre például azért volt szükség, mert a „*Manchester*” és „*Liverpool*” potenciális címkék csak a sporthíreken belül implikálhatják a „*Premier League*” absztrakt címkét.

4.3 A címkehalmaz szűrése

A szövegből kiemelt tulajdonnevek, főnévi csoportok, szótárillesztések és az absztrakt címkék után előálló potenciális címkék halmazának átlagos mérete túl magas (17,3 az elvárt 4-5-tel szemben), ezért a legfontosabbak kiválogatását meg kellett oldanunk. Ehhez figyelembe vettük a 4.1 fejezetben röviden bemutatott címkerangsort – vegyük észre, hogy az absztrakt címkékre nem értelmezett a tfidf alapú rangsoroló metrika –, a címke forrását (pl. listaillesztés vagy Wikipédia-alapú absztrakt), a cikk fő kategóriájára vonatkozó specialitásokat és az útmutató egyéb megkötéseit (például legalább egy *téma* címkének mindig szerepelnie kell, és csak olyan címkék használhatóak, amelyek legalább három dokumentumhoz hozzá lettek rendelve).

Ezen jellemzők alapján manuálisan konstruáltunk döntési szabályokat arra vonatkozólag, hogy mely címkék szerepeljenek a dokumentum végső címkehalmazában. Ezek a szabályok csak a felsorolt szintaktikai jellemzőkre épültek. A legfontosabb jövőbeli kutatási iránynak a szemantikai információk felhasználását tekintjük ebben a szűrésben. Ehhez a címkejelöltek közt páronként tervezzük a szemantikai kapcsolat numerikus értékkel történő jellemzését (például a Wikipédia-alapú heurisztikák felhasználásával [11]) majd az így kialakuló súlyozott teljes gráf elemzésével (például hubok vagy communityk azonosítása) kialakítható egy reprezentatív, de koherens szűrt címkehalmaz.

5 Kiértékelés

Az archívum végső címkézésben 59.364 különböző címke került felhasználásra, ami összesen 1.885.427 címke-cikk összerendelést eredményezett (átlagosan 4,98 címke hírenként). A címkék átlagos hossza 1,45 token.

Egy címkézés kiértékelése igen nehéz (és főképp szubjektív feladat), mert meg kell ítélni a kiválasztott címkék megfelelő számát, azok relevanciáját és koherenciáját. Ez nem végezhető el automatikus módon (ahhoz az egyes fogalmak közt ismernünk kellene a pontos szemantikai kapcsolatot, aminek birtokában tulajdonképpen az egész címkézési probléma sem lenne nyitott) csak manuális szemrevételezéssel. A projekt végén az [origo] munkatársai 1000 véletlenszerűen választott cikk automatikus címkézését manuálisan ellenőrizték. A véletlen választás biztosította, hogy a kiértékelő halmaz mind időben, mind cikk-kategóriában kövesse azok valós eloszlását.

A végső kiértékelési metrika dokumentumszintű volt, azaz minden dokumentumról született egy bináris – jó/rossz – döntés. A cikkhez automatikusan rendelt címkehalmazt manuálisan öt különböző szempont szerint értékelték:

- helyesen kiválasztott, valid címkék száma (súly +1),
- olyan címkék száma, amelyek nem kapcsolódnak szorosan a cikk témájához (súly -1),
- olyan címkék száma, amelyek ugyan kapcsolódnak a témához, de valamilyen egyéb szempontból érvénytelenek, például túl absztrakt, túl szűk fogalmak, elírások, összeragadt entitások (súly -0,2),
- a szerkesztő által hiányzónak ítélt címkék száma (súly -0,7)
- helytelen típusba sorolások száma (pl. személynév helyett földrajzi kategória) (súly -0,5).

Egy dokumentumot akkor tekintünk jónak, ha a fenti pontszámok súlyozott összege pozitív. Az egyes típusok súlyai a kiértékelés előtt rögzítésre kerültek és az Origo Zrt. elvárásainak figyelembevételével lettek kialakítva.

A kiértékelés alapján a dokumentumok 77,5 százalékának címkézése megfelelő minőségű lett, ami az eredeti célkitűzéseket meghaladja.

6 Konklúzió, nyitott kérdések

A cikkben bemutattuk, hogy egy újság archívumának automatikus címkézése kielégítő eredményt képes elérni. Címkézési módszerünk számos számítógépes nyelvészeti és statisztikai megoldást használt fel. A problémát több részproblémára bontottuk fel. Ezen részmodulok közül néhány már eléri a már jónak tekinthető szintet (pl. tulajdonnevek azonosítása, dokumentumzónák súlyozása), azonban van számos, amire idő és magyar nyelvtechnológiai erőforrások hiányában csak egy alapmegoldást adtunk. Ezeken a jövőben tovább dolgozunk.

Végezetül felsoroljuk azokat a szükséges számítógépes nyelvészeti módszereket, amelyek megléte a címkézés szempontjából nagy jelentőséggel bírna:

- A tulajdonnevek (ill. minden szótárban fel nem sorolt frázis) szótövesítése a morfológiai elemzési (guessing) megközelítések [12] és a korpuszstatisztikai módszerek [9] kombinációjaként kellene, hogy működjön.
- A főnevek képzése melléknevekből, igékből igen fontos lépés. A jelenlegi egyszerű átalakítási szabályok helyett szükség lenne egy morfológiailag megalapozott derivációra. A címkézés keretében elégséges lenne azt megvizsgálni, hogy egy adott kiindulási szóból lehetséges-e képezni egy szótár valamely elemét (azaz feltehetjük, hogy ismerjük a lehetséges címkék halmazát). Megjegyezzük, hogy a morphdb.hu természetesen már tartalmazza ezeket az átalakítási szabályokat, valószínűleg azok kiegészítése és invertálása lenne a célravezető.
- Jelenleg a szövegből kiemelt potenciális címkék szövegkörnyezetét nem vizsgáljuk. Ha a címke rangsorolásnál figyelembe vennénk például a címkék és az igék közötti viszonyt (vagy csak magát a vonatkozó igét) egy jóval szofisztikáltabb módszert kapnánk. Az igei vonzatkeretek és egyéb függőségi viszonyok automatikus azonosításában nagy előrelépést eredményezhet a Szeged TreeBank függőségi nyelvtan változatának elkészülése [13].

- A szemantikai kapcsolatok felderítése területén a legfrissebb kutatások a részben strukturált, hatalmas méretű nyers korpuszok (elsősorban Wikipédia) kiaknázására építenek. A magyar nyelvtechnológia szempontjából igen kedvezőtlen, hogy a magyar Wikipédia mérete mindössze 4%-a az angolénak, így az onnan kinyerhető információ is kevesebb. Véleményünk szerint itt nem lesz elégséges az angolra bevált módszerek alkalmazása, hanem újszerű megközelítésekre lesz szükség, amelyek képesek szemantikai kapcsolatokat kinyerni ilyen jellegű erőforrásokból.

Köszönetnyilvánítás

Szeretnék köszönetet mondani az Origo Zrt. munkatársainak (Krich Balázs, Kárpáti András, Cserti Gergely) – akik nélkül ez a valós életbeli kutatási projekt el sem indulhatott volna – a konstruktív és inspiráló eszmecserékért, valamint a projektben részvevő kollégáknak (Almási Attila, Berend Gábor, Hegedűs István, Vincze Veronika) áldozatos munkájukért.

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

Hivatkozások

1. Golder, S. A., Huberman, B. A.: Usage patterns of collaborative tagging systems. Journal of Information Science, Vol. 32, No. 2 (2006) 198-208
2. Kipp, M. E.I.: Tagging for Time, Task and Emotion. In: Proceedings of the 8th Information Architecture Summit, Las Vegas (2007)
3. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to Find Exemplar Terms for Keyphrase Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2009) 257-266
4. Mihalcea R., Tarau, P.: Textrank: Bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)
5. Sood, S. C., Owsley, S. H., Hammond, K. J., Birnbaum, L.: TagAssist: Automatic tag suggestion for blog posts. In: Proceedings of the International Conference on Weblogs and Social Media (2007)
6. Tatu, M., Srikanth, M., D'Silva, T.: RSDC'08: Tag Recommendations using Bookmark Content. In: Proceedings of the ECML PKDD Discovery Challenge (2008)
7. Hasan, K. S., Rahman, A., Ng, V.: Learning-Based Named Entity Recognition for Morphologically-Rich, Resource-Scarce Languages. In: Proceedings of the 12th Conference of the European Chapter of the ACL (2009) 354-362
8. Szarvas, Gy., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: DS2006, LNAI 4265 (2006) 267-278
9. Farkas, R., Vincze, V., Nagy, I., Ormándi, R., Szarvas, Gy., Almási, A.: Web-based lemmatisation of Named Entities In: TSD2008 LNCS Volume 5246 (2008) 53-60
10. Babarczy, A., Gabor, B., Hamp, G., Rung, A.: Hunpars: a rule-based sentence parser for Hungarian. In: Proceedings of the 6th International Symposium on Computational Intelligence (2005)

11. Berend G., Farkas R.: A Wikipédia felhasználása az absztrakt címkézési feladatban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 93-103
12. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D.: Hunmorph: open source word analysis. In: Proceeding of ACL (2005)
13. Vincze V., Szauter D., Almási A., Móra Gy., Alexin Z., Csirik J.: A Szeged Treebank függőségi fa formátumban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 127-138

A Wikipédia felhasználása az absztrakt címkézési feladatban

Berend Gábor[1], Farkas Richárd[2]

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

Kivonat: Az elektronikus, azon belül is az online tartalmak méretének robbanása újszerű megközelítést tesz szükségessé kategorizálásukra. Egy ilyen újszerű és elterjedt módszer az ún. címkézés, amely során dokumentumainkat azokat tömören és jól leíró kulcskifejezésekkel látjuk el. Ezek egy része egzaktul a szövegben is megtalálható, de kulcskifejezések lehetnek absztrakt címkék is, amik a dokumentumban nem fordulnak elő, mégis szemantikus kapcsolatba hozhatók a leírtakkal. Az [origo] hírportál archívumának automatikus felcímkézése során egyik részfeladatunknak a cikkekhez való absztrakt címkék hozzárendelését tekintettük, melyhez napjaink legnagyobb egységes formátumú, szabadon hozzáférhető tudásbázisát, a Wikipédiát használtuk föl.

1 Bevezetés

Az online tartalmak mennyiségének rohamos növekedésével egyre nehézkesebbé válik azok használata, katalogizálása. [4] szerint a 2007-ben 281 exabájtosra (281 milliárd gigabájtosra) becsült digitális univerzum mérete 2010-re várhatóan eléri az 1 zettabájtos határt, így nem is lehet kérdéses, hogy újszerű megközelítések szükségesek az online adatok rendszerezésére. Noha az egyszerű szöveges dokumentumok teljes digitális univerzumbeli részesedése csökkenő tendenciát mutat a multimédiás tartalmak térhódításának köszönhetően, fontosságukról így sem szabad megfeledkeznünk, hiszen mennyiségük így is változatlanul exponenciálisan nő. Ezt a növekedést támasztja alá [5] is, mely szerint a blogszféra mérete 5 havonta megduplázódik, naponta pedig átlagosan 30-40 ezer új blog kerül létrehozásra.

Éppen ezért a tartalmak kategorizálásának megkönnyítésére és a szövegekben történő könnyebb navigálás, keresés érdekében az utóbbi években – eleinte éppen a blogokon – bevezették az ún. címkézési (tagging) eljárást. Ezen Web2.0-ás eljárás során minden dokumentum szerzője az általa leírt tartalmat legtömörebben összegezni képes, néhány elemből álló kifejezéshalmazzal látja el írásait, amely alapján aztán könnyebben találhatjuk meg a minket érdeklő információkat. A módszer eredményességének láttán az eljárást időközben szinte minden tartalomszolgáltató bevezette, így a hírportálok is, mint például az [origo], amely szerkesztői 2009 eleje óta friss cikkei-

ket a bennük leírtakat legjobban megragadó kulcsszavakkal látják el. Egy ilyen megoldás hasznos szolgálatot nyújt mind a keresőoptimalizálás, mind pedig a weboldalakon megjelenő hirdetések egyes célcsoportokhoz való eljuttatása terén is.

A címkézés automatizálására – felhasználói megerősítés mellett – több megoldási kísérlet [6, 9, 12] született a korábbiakban, hiszen segítségükkel kiküszöbölhető lenne a korábban föl nem címkézett, nagy mennyiségű adathalmazok emberi erővel történő fölcímkézése mindamellett, hogy ezzel az egyes, tipikusan emberi címkézésre jellemző hibáktól [12] is mentesíteni lehetne a jelölést. A korábbi megoldások jellemzően kézi címkékkel ellátott dokumentumok alapján ajánlottak címkejelölteket a címkézetlen dokumentumoknak.

A dokumentumokhoz elvárhatóan rendelendő címkék egy része a szövegben is fellelhető – még ha esetleg nem is egységes formátumban (pl. a rövidítések vagy éppen toldalékolás miatt), vagy csupán implicit módon (*foci – labdarúgás*) –, más részük egyáltalán nem: hiszen például egy motorsportról szóló cikk esetében nem feltétlenül kell szerepeljen maga a *motorsport* kifejezés is a szövegben. Utóbbi kifejezéseket absztrakt címkéknek nevezzük. Az absztrakt címkék esetenként alkalmasabbnak bizonyulnak nem absztrakt társaikhoz képest, hiszen jóval informatívabbnak találjuk egy adalékanyagokkal foglalkozó dokumentum esetében az *élelmiszer-adalékanyagok* címke használatát (még ha az konkrétan nem is került megemlítésre a dokumentumban), mint a ténylegesen megemlített adalékanyagok listáját (pl. *tartrazin*, *gellángumi*, *nátrium-tartarát*, *csontfoszfát*).

Az előzőekben leírt okok miatt cikkünk az ilyen, ún. absztrakt címkék problémájára ad megoldási javaslatot, felhasználva napjaink legnagyobb egységes formátumban fellelhető, szabadon felhasználható elektronikus tudásbázisát, a Wikipédiát. Eljárásunkkal, amely a cikkekben előforduló releváns kifejezések Wikipédia-szócikkeire támaszkodik, tovább javítható a címkézés minősége: a fedésen, valamint a pontosságon túl a címkefelhő kohéziója egyaránt.

Munkánk során a cikkek szövegeiben előforduló potenciális címkék Wikipédia-szócikkeinek tartalmát éppúgy fölhasználtuk, mint a szócikkek közt hiperlinkek formájában megtestesülő kvázi-szemantikus viszonyokat. Az egyes szócikkekkel gyakran együtt előforduló egyéb fogalmak (szócikkek), valamint az egyes oldalakra mutató és belőlük kifelé irányuló relációk (linkek) vizsgálata éppúgy hasznosnak bizonyult, akárcsak a szócikkek közötti átirányítások (redirect) figyelembevétele.

2 Kapcsolódó munkák

A számítógépes nyelvészeti munkák közül leginkább az automatikus címkézéssel, valamint a termek közötti szemantikus relációk Wikipédia segítségével történő automatikus föltérképezésével foglalkozó irodalomra támaszkodtunk.

2.1 Automatikus címkézés

Az eddigi automatikus címkézésről szóló munkák két fő irányvonalba sorolhatók. Az egyik megoldási módozat, az ún. címke- vagy kulcsszókinyerés (*tag / keyphrase*

extraction) során a fölcímkézendő cikkek szövegéből nyerik ki a címkejelölteket, akárcsak [3]-ban. Egy hátulütője az efféle kulcsszókinyerő rendszereknek, hogy ezek csak a dokumentumokban ténylegesen is előforduló címkék szövegből történő kiemelésére alkalmasak.

Absztrakt címkézési megközelítésünkhöz legközelebb álló megoldások a [9]-hez hasonló, ún. címke-hozzárendelő (*tag assignment*) rendszerek. Ezek a megoldások a fölcímkézendő dokumentumokhoz hasonló, kézi jelöléssel már ellátott dokumentumok címkéinek hozzárendelésével oldják meg a címkézési feladatot, így ezek a megoldások is absztrakt címkézésként foghatók föl, ugyanis egy dokumentumhoz olyan címkék is hozzárendelhetők, melyek annak szövegében nem fordulnak elő. Az ilyen módszerek hátránya azon túl, hogy a hozzárendelt címkék megőrzik az emberi címkézés esetlegességeit, hogy a dokumentumokhoz rendelt címkék egy zárt halmazból kerülhetnek csupán ki, vagyis a tárgyalt témákban az időben végbe menő változásokat nem tudják naprakész, friss címkékkel követni. Ezzel szemben az általunk javasolt rendszernek nincs szüksége kézi címkékkel ellátott dokumentumokra, az absztrakt címkék meghatározása során pedig a hasonló dokumentumok keresésén túlmutató, szemantikusan kapcsolódó címkéket javasol.

2.2 Szemantikus viszonyok vizsgálata

Az automatikus címkézés során hasznos, ha képesek vagyunk meghatározni kifejezések között fönnálló szemantikus viszonyokat: segítségükkel ki lehet szűrni egy dokumentum kulcsszójelöltjei közül azokat, melyek nem koherensek a többivel, vagy épp ellenkezőleg, a jelöltek közötti kohézió megtartása mellett újakkal lehet kiegészíteni azokat. A szemantikus relációk vizsgálata során az utóbbi években többen is a legnagyobb, részben strukturált online tudásbázist, a Wikipédiát használták föl szemben a korábbi megközelítésekkel [10], amelyek ontológiákra vagy különféle korpuszokon mért kifejezések együttes előfordulásának kiszámítására támaszkodtak.

[11] a szövegekben előforduló többértelmű tulajdonnevek (pl. *Kennedy (repülőtér) – Kennedy (személy)*) egyértelműsítésére használta föl a Wikipédiát. [1, 7] egyaránt termek között fönnálló szemantikus viszony erősségét meghatározó rendszert mutatnak be, melyek a szócikkek által kifeszített vektortérben vett hasonlósági mértékek alapján hoznak döntést.

Munkánkhoz legközelebb az előbbi munkákra is támaszkodó [6] áll, mely egy dokumentum szavaihoz egyértelműsítés után rendelt Wikipédia-szócikkek közül gráfanalízist használva választja ki azokat, amelyek leginkább képesek lehetnek az eredeti dokumentum tartalmának megragadására.

3 Módszerek

Absztrakt címkéző eljárásunk az egyes cikkek szövegeiből kinyert, abban egzaktul előforduló kifejezések halmazát várja bemenetül, majd ezekhez rendeli hozzá a velük vélhetően szemantikus relációban álló Wikipédia-szócikkek halmazát. A bementként szolgáló címkejelölteket a cikkekből a [2]-ben leírtak szerint nyertük ki. Ezután a

szövegből kinyert címkeaspiránsokhoz meghatároztuk azon Wikipédia-szócikkeket, amelyek egy az egyben megfeleltethetők a címkejelöltek halmazának legalább egy elemével. Olyan szócikkek esetében, amelyek egyértelműsítő lappal rendelkeztek, nem választottuk ki a szócikk egyik egyértelműsítő lapját sem, elkerülendő ez által az esetleges rossz választásokból adódó zajt a továbbiak során.

Az absztrakt címkék megtalálására alkalmazott módszereink egyaránt támaszkodnak a hírportál cikkeiből kinyert címkejelöltek Wikipédia-szócikkeinek szöveges tartalmára, valamint a közöttük meglévő gazdag linkstruktúrára. A következő fejezetek ezeket az eljárásokat mutatják be részletesen.

3.1 Átirányítások figyelembevétele

A Wikipédia felépítéséből adódóan azonos tartalmak több szócikk alól is elérhetők. Így például akár az *USA*, akár pedig az *Amerikai Egyesült Államok* szócikkekre keresünk rá, egyazon oldalt kapjuk találatul. Ezen ún. átirányító (*redirect*) Wikipédia-oldalak szinonimák, illetve asszociációk meghatározására, rövidítések feloldásai valamint korlátozott mértékig elíráskezelésre egyaránt alkalmazhatók (például 1. táblázat). Segítségükkel kanonikus alakra tudunk hozni eltérő formában előforduló, de azonos jelentéssel bíró címkejelölteket, amivel a teljes címkézés kohézióját javíthatjuk (mivel azonos jelentésű címkék nem fordulnak elő több formában, mint nyereség – profit).

1. táblázat: A Wikipédiában szereplő Amerikai Egyesült Államok szócikkre irányuló átirányítások listája.

Amerikai	Amerikai Egyesült Államok
Amerikaiak	Amerikai Egyesült Államok
Amerikai egyesült államok	Amerikai Egyesült Államok
Egyesült államok	Amerikai Egyesült Államok
Egyesült Államok	Amerikai Egyesült Államok
United Stated	Amerikai Egyesült Államok
United States of America	Amerikai Egyesült Államok
US	Amerikai Egyesült Államok
USA	Amerikai Egyesült Államok

Absztrakt címkéző módszerünk a címkeaspiránsokhoz rendelt Wikipédia-szócikkek közül lecseréltük mindazokat, amelyek más szócikkre voltak irányítva. Ezen a ponton az automatikus címkézés eredményeképp előálló címkefelhő kohézió növelése volt a cél, mivel így elkerülhető volt az eltérő alakban álló, de ugyanazzal a szemantikus jelentéssel bíró címkék alkalmazása.

3.2 Definíciók kinyerése

Ebben a lépésben a Wikipédia oldalnak megfeleltethető címkejelöltekhez rendeltünk definíciókat, amelyek aggregálása után újabb címkejelöltet voltunk képesek javasolni

a már meglévők mellé. Az ilyen módon nyert definíciók jól megragadják az egyes szócikkekben leírt fogalmak hiponim relációit: a *krizoin*ról például megállapítható, hogy az egy *adalékanyag*.

Megfigyelhető, hogy a Wikipédia enciklopédikus jellegéből adódóan az egyes oldalak elején megtalálható a bennük tárgyalt fogalom definiálása. Úgy jártunk el, hogy minden egyes címkejelölthöz meghatároztuk annak Wikipédiáról automatikusan kinyert definícióját, és amennyiben egy definíció címkejelöltek egy adott halmazán több esetben is alkalmasnak bizonyult, úgy azt absztrakt címkeként javasoltuk.

Egy szócikk által leírt fogalom potenciális definícióinak kinyeréséhez elsőként meg kellett határozzuk azt a mondatot, amelyből az kinyerhető lehet. Megközelítésünkben ez a mondat minden esetben az volt, amelyik elsőként megemlítette a szócikket magát, vagy amennyiben nem szerepelt ilyen az egész oldalon, úgy a szócikk első bekezdésének első mondatát tekintettük ilyennek. Az ily módon kinyert szócikkmondat megfeleltetésekre példákat a 2. táblázat hoz.

2. táblázat: Wikipédia-szócikkekből kinyert definíciót tartalmazó mondatok.

Erdős Pál	**Erdős Pál**, a 20. század egyik legkiemelkedőbb *matematikusa*, az *MTA tag*ja.
Gottlob Frege	**Friedrich Ludwig** Gottlob Frege, *német matematikus*, *logikatudós*, *filozófus*, a modern matematikai logika és analitikus filozófia megalapítója, művelője.
Maffiózók	A **Maffiózók** egy *amerikai TV-sorozat*, amelynek David Chase a kitalálója és producere.

Az előzőek szerint generált potenciálisan definíciót tartalmazó mondatokból következő lépésként magukat a lehetséges definíciókat nyertük ki. Ezen lépés során a mondaton belüli szövegkörnyezetet figyelembe véve, továbbá morfológiai és szintaktikai megfontolásokat alkalmazva határoztuk meg az adott szócikkhez tartozó definíciókat, melyeknek vagy önmaguknak is vagy pedig tagonként önálló Wikipédia-szócikk-címeknek kellett lenniük. (Így lett alkalmas definíció az *amerikai TV-sorozat*, ahol az *amerikai* és a *TV-sorozat* külön szócikként szerepel a Wikipédiában.) A leírtak alapján nyert szócikk-definíció párosokra a 3. táblázatban láthatók példák.

3. táblázat: Példa definíciógenerálásra.

Erdős Pál	*matematika*
Gottlob Frege	*matematika, német, filozófia*
Maffiózók	*producer, amerikai TV-sorozat, TV-sorozat*

Átfedő definíciójelöltek esetén (pl. *amerikai*, *TV-sorozat* és *amerikai TV-sorozat*) a leghosszabb szupersztringet választottuk (*amerikai TV-sorozat*). Végül egy dokumentum címkejelöltjeihez akkor rendeltünk hozzá definíciókat is absztrakt címkeként, ha az több címkejelölt esetében is relevánsnak lett minősítve, vagyis például egy olyan esetben, ahol egy dokumentum címkejelöltjei között szerepelt *Erdős Pál* és *Gottlob Frege* is, ott fölvettük a *matematika* szót is mint címkejelöltet, hiszen az mindkettő esetében értelmes definíciónak lett titulálva.

3.3 A linkstruktúra kiaknázása

Adott dokumentumból kinyert címkejelöltekhez rendelhető absztrakt fogalmakat a Wikipédia linkstruktúrája szempontjából is vizsgáltuk: megkerestük azokat a további szócikkeket, amelyek jellemzően együtt fordulnak elő egy potenciális címkéhez rendelt szócikkel, vizsgáltuk azokat a szócikkeket, amelyekre egy hírdokumentumhoz rendelt szócikkek közül több is hivatkozott, illetve megkerestük azokat a szócikkeket, amely egy dokumentum címkejelöltjeihez generált szócikkek halmazát a leginformatívabban tartalmazzák.

Együtt-előfordulás vizsgálata
Ebben az esetben minden egyes címkejelölthöz, melyhez hozzárendeltünk Wikipédia-szócikket, megkerestük azon egyéb szócikkeket, amellyel együtt az gyakran előfordul. A vizsgálat elvégzését csak olyan szócikkek esetében végeztük el, amely legalább 10 és legfeljebb 150 oldalon lett hivatkozva. Ennek oka az volt, hogy a 10 esetnél kevesebbet hivatkozott szócikkek nem tűntek eléggé relevánsnak, a 150-nél többször előfordulók pedig túl általános gyűjtőoldalaknak bizonyultak.

Az olyan szócikkekre, amelyekre a hivatkozások száma az előbb említett két korlát között volt megkerestük azokat a szócikkeket, amelyek legalább az esetek felében ugyanúgy megfigyelhetők voltak a hivatkozó oldalakon linkek formájában. Így például, mivel *Sébastien Loeb* raliversenyző *rali-világbajnokság* szócikkel való együttes előfordulása 0.7073 volt, a Sébastien Loeb nevét tartalmazó cikkhez a rali-világbajnokság címke is fölvételre került.

A kimenő linkek vizsgálata
A kimenő linkek esetében azokat a szócikkeket kerestük, amelyek relevánsnak tekinthetők szócikkek egy adott halmazára nézve. Ehhez vettük a bemeneti szócikkhalmaz egyes elemeiből kifelé irányuló megbízható linkekhez tartozó szócikkeket. Megbízhatónak tituláltunk egy linket, ha az általa hivatkozott oldal tartalmazott visszaélt a hivatkozó dokumentum irányába, vagy a hivatkozó oldal linkjeinek legalább 25%-át a másik oldalra való hivatkozás tette ki, és ezen linkek száma legalább 3 volt (kivéve a portál – és kategória gyűjtőoldalakra mutató linkeket, mivel azok a szerkesztési konvenciókból adódóan az oldalak alján egy példányban szerepelnek többnyire).

Az előbbiek szerint minden egyes Wikipédia-szócikkel rendelkező címkejelölthöz az általuk hivatkozott szócikkek közül azokat tartottuk ténylegesen is relevánsnak a teljes hírcikkre nézve, melyekre nem csupán egy szócikkből mutatott relevánsnak titulált link. Például egy cikk esetében, amely címkejelöltjei között szerepelt a *BUX* és a *Budapesti Értéktőzsde* is, egyúttal implikálta a *Magyarország gazdasága* címke fölvételét is, mivel arra mindkét oldalhoz tartozó Wikipédia-szócikk referál.

Tartalmazások vizsgálata
Az eddigieken túl szemantikus kapcsolatok tárhatók föl szócikkek egy halmaza és egy további szócikk között, ha megvizsgáljuk, hogy egy potenciális absztrakt címkének megfeleltethető szócikk az inputként kapott szócikkhalmaz elemeit milyen mértékben tartalmazza.

A termhalmazok és az absztrakt címkejelöltként funkcionáló szócikkek közötti tartalmazás mértékének számszerűsítésére a tf-idf metrikát adaptáltuk. A bemenetként szolgáló címkeaspiráns-halmaz alapján meghatároztuk azokat a szócikkeket, amelyek legalább egyet is tartalmaznak közülük link formájában. Ezek után az összes szócikk előző feltételnek eleget tevő részhalmának minden elemére kiszámítottuk az adott bemeneti szócikk halmazra vett átlagos tf-idf értékükett, amely ha adott küszöbérték feletti volt, akkor absztrakt címkeként kezeltük a továbbiakban az adott szócikket.

4 Eredmények

Absztrakt címkézési eljárásunk kiértékelésére az [origo] hírportál dokumentumainak kézi címkézésének megkezdése óta keletkezett, január és február hónapokból választott 600-600 dokumentumát választottuk ki. A kiértékelést két annotátorra bíztuk, a 600-600 dokumentumból pedig 100 mindkét annotátor esetében azonos volt, így összesen 1100 különböző cikk került kiválasztásra. Az 1100 dokumentumból azonban csak 1073 esetében állt rendelkezésünkre az absztrakt címkéző eljárásunk inputjaként szolgáló, a cikkek szövegéből kinyert címkejelöltek halmaza, aminek az oka az, hogy az [origo] specifikációja alapján a film-blog csatornájukba tartozó dokumentumaik címkézését nem kellett elvégezzük (a kérdéses 27 dokumentum pedig ebbe a csatornába esett). Így legvégül 584, illetve 589 dokumentum automatikus absztrakt címkézésének kiértékelése történt meg.

Az annotátorok feladata az volt, hogy minden dokumentum esetében a Wikipédia 2009. szeptember 14-i tartalma és struktúrája alapján az egyes hírcikkekhez rendelt absztrakt címkékről döntsék el, hogy azok az adott cikk esetében elfogadhatók-e, valamint hogy határozzák meg, hogy az automatikusan generált absztrakt címkék megfeleltethetők-e a manuális címkézés egy vagy több cikkben ténylegesen elő nem forduló elemével. A végső pontosságot az alkalmasnak talált absztrakt címkézési eljárással nyert címkék arányának (pontosság) és a manuális címkékhez viszonyított fedés értékekének kombinált értékeiből számított F-mértékkel határoztuk meg.

A vizsgált dokumentumokhoz az [origo] munkatársai összesen 1192 alkalommal rendeltek a szövegben elő nem forduló kifejezéseket címkeként, ami dokumentumonként átlagosan 1,11 absztrakt címkét jelent. Az 1192 alkalommal összesen 554 különböző absztrakt címkét használtak. Az annotálás során azt tapasztaltuk, hogy egyes esetekben a cikkek szövegben elő nem forduló címkeként használt termek szinonímája (pl. *gazdasági válság – recesszió*) már megtalálható volt, és ezt az absztrakt címkézést megelőző lépésekben eredményesen ki is nyertük. Más esetekben pedig csupán az absztrakt címke kézi hozzárendelése során történő elírások (pl. Sony Ericcson – Sony Ericsson) tettek absztrakttá (vagyis a cikk szövegében elő nem fordulóvá) egyes kifejezéseket, így az automatikus absztrakt címkék fedésének vizsgálata során az ezekkel való pontos egyezést nem követeltük meg. Ezen „kvázi-absztrakt” címkék figyelmen kívül hagyásával összesen 1114 ténylegesen is absztrakt címke található az 1073 dokumentumból álló teszthalmazon (dokumentumonként átlagosan 1,038), melyek dokumentumok szerinti eloszlását a 4. táblázat tartalmazza.

4. táblázat: Hírdokumentumok és a manuálisan meghatározott absztrakt címkék eloszlása.

Absztrakt címkék száma	Dokumentumok száma	Címkék mennyisége
0	339	0
1	465	465
2	184	368
3	65	195
4	18	72
5	1	5
9	1	9
Összesen	1073	1114

Az 1073 vizsgált dokumentum esetében összesen 13689 címkeaspiránst nyertünk ki az absztrakt címkézést megelőző lépésekben, amelyekhez 5239 esetben voltunk képesek Wikipédia-szócikket rendelni. Az egyedi címkeaspiránsok száma 6578 volt, közülük 1766-hoz (26,85%) határoztunk meg Wikipédia-szócikket, melyek segítségével 5014 alkalommal rendeltünk hozzá összesen 2028 különböző automatikus absztrakt címkét cikkekből kinyert címkeaspiránsok halmazaihoz. A dokumentumok eddigiek alapján vett eloszlásai az 5. táblázatban szerepelnek, melyből az is kitűnik, hogy 32 dokumentum egyetlen címkeaspiránsához sem tudtunk Wikipédia-szócikket kötni.

5. táblázat: Dokumentumok eloszlása a hozzájuk rendelt kezdeti címkeaspiránsok/ Wikipédia-szócikkek/ absztrakt címkék száma szerint.

	Dokumentumok száma n darab		
	szövegből származó címkeaspiránssal	**Wikipédia-szócikk-hozzárendeléssel**	**automatikus absztrakt címkével**
n = 0	0	32	157
0 <n <=5	72	669	639
5 <n <=10	388	320	174
10 <n <=20	509	51	73
n> 20	104	1	30
Összesen	1073	1073	1073

Az 5014 absztrakt címke 5733 címke-hozzárendelésnek volt köszönhető, mely azzal magyarázható, hogy bizonyos absztraktcímke-jelöléseket egyszerre több módszer is javasolt, az egyes módszerek közötti eloszlás pedig a 6. táblázatban látható.

6. táblázat: Az absztrakt címkéző eljárások közötti eloszlás.

Módszerek	**Címke-hozzárendelések száma**
Átirányítás	1155 darab (20.146%)
Definíciók	1471 darab (25.658%)
Együttes előfordulás	1998 darab (34.676%)
Kimenő linkek	558 darab (9.733%)
Tartalmazó szócikkek	551 darab (9.611%)
Összesen	5733 darab (100%)

Mind az 5733 hozzárendelést külön módszerenként vizsgálva, a pontosság értékére a 7. táblázatban lévő adatokat kaptuk.

7. táblázat: Az egyes módszerek által bevont absztrakt címkék pontossága.

Módszerek	Címke-hozzárendelések száma	Elfogadott hozzárendelések	Pontosság
Átirányítás	1155	836	0.7238
Definíciók	1471	414	0.2814
Együttes előfordulás	1998	697	0.3488
Kimenő linkek	558	227	0.4068
Tartalmazó szócikkek	551	90	0.1633
Összesen	5733	2264	0.3949

Az absztrakt címkézés kiértékelésének végső eredményét a két annotátor döntései alapján a 8. táblázat tartalmazza.

8. táblázat: A kézi kiértékelés végső eredménye.

	Pontosság	**Fedés**	**F-mérték**
1. annotátor	0.3933	0.1057	0.1666
2. annotátor	0.3848	0.1077	0.1683
Összesítve	0.3891	0.1067	**0.1675**

5 Konklúzió

Módszerünket az [origo] hírportál címkézetlen archívumán teszteltük, a Wikipédia segítségével bevont absztrakt címkék fölvételével pedig sikerült javítanunk a legvégül előálló címkefelhő minőségén.

Az eredmények figyelembevételénél fontos szem előtt tartani, hogy az automatikus absztrakt címkézés fedésének értéke a cikkekhez ténylegesen hozzárendelt címkékhez lett mérve, ami pedig olyan fogalmakat is tartalmazott, amelyekre a magyar Wikipédiában egyáltalán nem létezik szócikk (pl. *gyárbezárás*), vagy pedig helyességük megkérdőjelezhető (*"Hearts, FTC"* vagy a *"fogászat, árak"* [mindkettő egybe, egy címkeként]). Az ilyen címkék Wikipédia fölhasználásával történő cikkekhez rendelése pedig nemcsak, hogy nem lehetséges, de esetenként nem is lenne célszerű.

Módszerünkre jellemző, hogy eredményessége függ a bementként kapott címkeaspiránsok halmazától, így fontos, hogy azok minősége megfelelő legyen. Ezen túl, ahogy az az 5. táblázatban is látható, 32 dokumentum esetében egyáltalán nem tudtunk Wikipédia-szócikket társítani a bemenetként kapott címkejelöltekhez, így ezekben az esetekben nem is volt lehetőség absztrakt címkék bevonására (a legtöbb módszer ugyanis legalább kettő, a cikk szövegéhez kapcsolódó szócikk címének meglétét igényli). Ezért úgy gondoljuk, hogy tovább javítható lenne módszerünk, amennyiben az eddigiekben figyelmen kívül hagyott (szócikkel nem rendelkező) címkejelöltekhez is társítani tudnánk Wikipédia-oldalakat. További javítási lehetőség látunk még az egyes szócikkeken előforduló linkek alkalmas súlyozásában is, annak megfelelően, hogy azok mekkora mértékben kötődnek az adott szócikkben tárgyaltakhoz.

Ugyan a kézi címkézés során alkalmazott 554 különböző absztrakt címkének megközelítőleg 20%-a bír csak Wikipédia-szócikkel, ezek közül 58-at sikerült pontosan, vagy legalább egy közeli szinonimájával meghatároznunk módszereink valamelyikével. Az esetlegesen tévesen kiválasztott absztrakt címkéket pedig a későbbi címkeszűrési lépések során igyekeztünk eredményesen eltávolítani, amit a teljes címkéző rendszerünk eredeti várakozásainkat meghaladó végső 77.5%-os értékelése is alátámaszt.

Eljárásunkról az is elmondható, hogy a Wikipédia többnyelvűségéből fakadóan más nyelvekre is könnyűszerrel adaptálható, eredményessége pedig várhatóan az adott nyelven elérhető Wikipédia szócikkeinek számától, valamint az oldalak szerkesztésének (a köztük lévő linkstruktúra) minőségétől is függ.

6 Köszönetnyilvánítás

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

Hivatkozások

1. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis (2007)
2. Farkas R.: Az [origo] automatikus címkézési projekt tapasztalatai. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 84-92
3. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G.: KEA: Practical Automatic Keyphrase Extraction
4. Gantz, J. F. et al.: The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011. http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf (2008)
5. Kim, J. W., Selçuk Candan, K., Tatemura, J.: CDIP: Collection-Driven, yet Individuality-Preserving Automated Blog Tagging (2008)
6. Grineva, M., Grinev, M., Lizorkin, D.: Extracting Key Terms From Noisy and Multi-theme Documents. (2009)
7. Strube, M., Ponzetto, S. P.,: WikiRelate! Computiong Semantic Relatedness Using Wikipedia. Americal Association for Artificial Intelligence (2006) 1419-1424

8. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT 2007 (2007) 196-203
9. Sood, S. C., Owsley, S. H., Hammond, K. J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. 1th International Conference on Weblogs and Social Media (ICWSM'2007)
10. Patwardhan, S., Banrjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. CICLing 2003, LNCS 2588 (2003) 241-257
11. Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007) 708-716
12. Waltinger, U., Mehler, A., Heyer, G.: Towards Automatic Content Tagging: Enhanced Web Services in Digital Libraries Using Lexical Chaining. 4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08) (2008) 231-236

Szóhasonlóság mérése analógiás megközelítésben

Rung András[1]

[1] MTA Nyelvtudományi Intézet, Elméleti Nyelvészet Tanszék, Benczúr utca 33., 1068 Budapest, Magyarország
rungandras@gmail.com

Kivonat: A magyar szavak, elsősorban a főnevek hasonlóságát meghatározó tényezők leírására törekszem. Ebben a szabályalapú nyelvtanok helyett a mentálisan reálisabbnak és rugalmasabbnak tűnő analógiás keretrendszert tekintem kiindulási alapnak. Munkámban a számítógépes nyelvészet eredményeire és módszereire támaszkodom. Számításaim, megállapításaim kizárólagosan korpuszból vett adatokon alapszanak. Kutatásomnak közvetlen hozadéka is lehet a nyelvtechnológia számára a szótárbővítés és –karbantartás területén, mivel a hasonlóságot mérő algoritmusom 95%-os pontossággal ismeri fel főnévi tövek hangkivető voltát, amely már lehetővé teszi az ilyen szavaknak akár automatikus besorolását is.

1 Bevezetés

A szabályalapú nyelvtanok sokszor jó közelítő leírást adnak az alaktani viselkedésről, azonban képtelenek olyan nyelvi jelenségeket magyarázni, mint például a fokozatosság, nyelvi ingadozás, a gyakoriság hatása a nyelvi változásra [8]. Ezekben az esetekben az analógiás nyelvtan jobban közelíti a pszichológiai realitást, azaz a valós nyelvi működést. Ha bármely nyelv analógiás nyelvtanát kívánjuk megírni, annak egyik alapfeltétele az, hogy tudjuk mely fonémák-hangok, alakok [1], konstrukciók [2] hasonlóak az adott nyelvben, és ezek hasonlósága milyen mértékű, min alapszik.

Feltételezésem szerint az egyes nyelvi elemek nem elszigetelten léteznek, hanem szoros és állandó interakcióban vannak egymással, amelynek egyik legfontosabb mozgatórugója az analógia. Ha a nyelvben valahol változás következik be, akkor az az erőviszonyokra azonnal hatással van, és a rendszer egészének változásához vezet. Ezt a legtöbb 20. századi nyelvelmélet el is ismeri. Ennek megfelelően az egyes állapotok leírásával foglalkozik a szinkrón nyelvészet, míg az ezek közti átmenetek vizsgálatával a diakrónia.

Ez a megközelítés azonban kimondva vagy kimondatlanul azt közvetíti, hogy vannak mindig stabil, önmagukban megfigyelhető állapotok. A nyelv változik, de maga a változás nem alapvető minősége. Az analógiás nyelvtan által megfigyelt tények, jelenségek azonban cáfolják ezt a szigorú és merev szétválasztást. A változás és az állapot nehéz szétválaszthatóságából következik, hogy a rendszer a maga statikusságában nem létezik, vagy legalábbis olyan absztrakt fogalom, amely a nyelvvel való valós munkára alkalmatlanná teszi.

Ha a nyelvi jelenségeket szorosan összefüggőnek vesszük és változásukat lényegi elemüknek tekintjük, akkor felvetődik a kérdés, hogy egyáltalán lehet-e és értelmes-e a nyelvnek bármely részjelenségét leírni anélkül, hogy más részleteit ne vennénk figyelembe, hisz az összefüggések feltárása nélkül, a jelenség értelmezhetetlen vagy csak részlegesen értelmezhető lesz.

A leírás ebben az esetben valóban nem lesz tökéletes, de mivel a nyelvi változás teljességét az arról való meglehetősen korlátozott és szerény tudásunk miatt semmiképp sem tudjuk megragadni, így mégis kénytelenek vagyunk csak egyes darabjait vizsgálni. A hagyományos megközelítésekkel ellentétben azonban nem állítom, hogy ezeket a részleteket önállóan és pontosan le tudjuk írni, hanem úgy vélem, hogy újabb leírások fényében majd kiegészítésre szorulnak a későbbiekben. Adatainkat folyamatos változásukból kifolyólag sosem tudjuk megragadni, de ez nem is gond, hisz a nyelvészet célja nem feltétlenül a leírás, hanem a leírást meghatározó nyelvi folyamatok megértése és pszichológiailag reális feltárása.

2 Célkitűzések

A szavak összehasonlításában azok felszíni szerkezetét és alakjaik gyakoriságát veszem alapul, mivel az analógiás keretrendszer a mögöttes szerkezeteket inkább gátló, semmint hasznos elméleti konstrukcióknak tartja [1]. Ezekre támaszkodva az ingadozással és fokozatossággal szorosan összefüggő analógiás kiegyenlítődés is jól megragadható jelenséggé válik [5]. Feltételezésem szerint az analógiás alapú változásokat több további szempont is meghatározza (használati mód, jelentés, stb.), de ezekből a legfontosabb a hangtani/fonológiai hasonlóság.

A szavak fonológiai hasonlóságának meghatározására keresek egy egyértelmű algoritmust, amellyel modellálhatom ezt az analógiában egyik legfontosabb szerepet játszó komponenst. Amennyiben kitűzött célomat elérem, az a magyar nyelvtechnológia számára is értékes hozadékkal jár, hisz egy pontos, a valós folyamatokat jól megragadó algoritmus segítségével lehetővé válik a szavak hatékony szótárba sorolása (megfelelő jegyeikkel), illetve a már meglévő szótári anyag frissítése, karbantartása is, amely komoly kihívást jelent, ha csak emberi erőre hagyatkozunk. A szótárak automatikus bővítésével lehetővé válik akár rétegnyelvi (szleng, szaknyelv, stb.) szövegek hatékonyabb elemzése is.

Leggyakrabban az edit distance algoritmust [6] használják két szó, karakterlánc hasonlóságának eldöntésére. Ez a megközelítés azonban számos problémát vet fel. Egyrészt az algoritmus az összehasonlítást betűk és nem fonémák alapján végzi. Két betűt azonosnak vagy teljesen különbözőnek vesz, így az *o* és az *ó* ugyanannyira különböző az algoritmus számára, mint az *o* és a *k*. Továbbá az algoritmus feltételezi, hogy a törlés, beillesztés és megfordítás pont ugyanakkora változást okoz a szón belül, és ezeknek a beavatkozásoknak a helye is lényegtelen. Korábbi [8] és jelenlegi vizsgálatom is megmutatja, hogy ez az algoritmus emberi nyelvek szavainak (legalábbis a magyar nyelv szavainak) hatékony és megbízható összehasonlítására nem alkalmazható, így legfeljebb csak kiindulási pont lehet olyan kifinomultabb megközelítések számára, amelyek jobban megragadják az emberi nyelvi rendszer sajátságait és működését.

3 A hasonlóság mérésére használt algoritmus

A szavak közti hasonlóság mérésére egy python programnyelvben megírt algoritmust használok, amely egy hasonlósági mátrix alapján végzi számításait. Ez a mátrix adja meg, hogy két fonéma mennyire hasonlít egymáshoz. A hasonlóság értéke a 0 és 1 közti skálán helyezkedik el. Így két fonéma nem csak azonos vagy eltérő lehet, hanem az analógiás nyelvi megközelítéssel összhangban több, bár diszkrét fokozatban adható meg hasonlóságuk.

A fonémákat alapul véve korrigálom az edit distance algoritmus azon hiányosságát, hogy az összehasonlításban nem a szavakat, hanem azok grafémikus leképzését veszi alapul. A fonémák kiválasztásában és jegyeik meghatározásában a Magyar Strukturális Nyelvtant [4] és az Ipa (International Phonetic Alphabet) leírásait tartottam irányadónak.

Mivel az analógiás megközelítés egyes irányzatai [1] szerint a szavakat jelentésük mellett konkrét hangalakjukkal is tároljuk, érdemes lett volna a hasonlóságot fonetikus alapon (is) számolni. Ettől a lehetőségtől azonban kénytelen voltam eltekinteni, mivel jelenleg a fonetika nem tudja egyértelműen meghatározni, hogy két hang mikor és mennyire hasonlít, valamint egy ilyen vizsgálathoz szükséges beszéd korpusz vagy az ez alapján készített beszélt nyelvi gyakorisági szólista sem áll rendelkezésre. Így ezek hiányában maradtam a fonémák összehasonlításánál, amelyek még ha meglehetősen durva összehasonlítást is tesznek lehetővé, mégis legalább közelítik a nyelvi realitásokat.

A fonémák hasonlóságának mértékét a fonémák megkülönböztető jegyei alapján számolom. A magánhangzók esetében a nyíltságot, ajakkerekítést, hosszúságot, elölképzettséget, a mássalhangzók esetében pedig a zöngésséget, a képzés helyét és módját veszem figyelembe. Minden eltérő jegy esetén az összehasonlított két fonéma hasonlóságát felére csökkentem. A mássalhangzók és a magánhangzók egymáshoz számított hasonlósága rendszeremben 0, nincsenek közös jegyeik. A későbbiekben tervezem köztes kategóriák bevezetését bizonyos hangokra (*j* és *v*), illetve a hangok egységes kezelésének érdekében azonos mássalhangzók kapcsolatait nem geminátának venném, hanem e megoldás helyett a mássalhangzóknak is lenne hosszúság jegye.

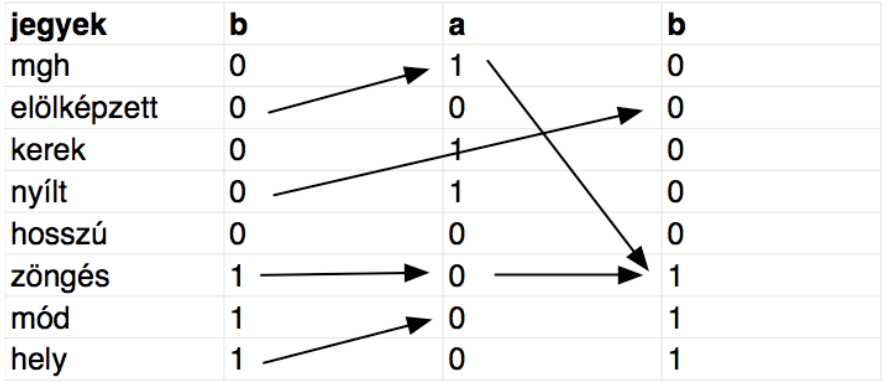
Ezek alapján az *o* fonéma hasonlóságának mértéke egy másik *o* fonémához 1, az *ö*-höz és az *ó*-hoz 0,5 ($1:2^1$, mivel egy jegyben az elölképzettségben, illetve a hosszúságban különböznek), míg az *ő*-höz 0,25 ($1:2^2$ hiszen két jegyben az elölképzettségben és a hosszúságba különböznek). A magyar nyelvleírás hagyományát követve az edit distance algoritmussal ellentétben nagyobb súlyt adok a szóvégek hasonlóságának. A fonémahasonlóság súlya a szóvégétől a szó eleje felé logaritmikusan csökken. Az algoritmusomban 1,8-es alapú logaritmust használok, mivel korábbi vizsgálataimban ez bizonyult a leghatékonyabbnak[7]. A szavak önmagukkal vett hasonlósági értéke 1. Programom számítása alapján a *bab* és a *púp* szavak hasonlósága a következőképp alakul:

b:p = 0,5 (eltérő jegy: zöngésség)
a:ú = 0,25 (eltérő jegy: nyíltság, hosszúság)
b:p = 0,5 (eltérő jegy: zöngésség)

Hasonlóság kiszámítása a logaritmikusságot is figyelembe véve (a könnyebb átláthatóság kedvéért 2-es alapú logaritmussal):

((0,5*1)+(0,25*2)+(0,5*4))/7=0,5

A fentebb leírt algoritmus mellett egy másik algoritmus tesztelését is megkezdtem, amely a hasonlóság számításában az edit distance-hez hasonlóan nem ad súlyt annak, hogy a szavak mely részei hasonlóak. A hasonlóságot az alapján határozza meg, hogy a két szó fonémáinak jegyeiből épített mátrixokban hány közös részgráf van. Így a *bab* és a *púp* tartalmazza a CVC a CV, VC, CC, C és V illetve a zárhang-hátulképzett-zárhang, zárhang-hátulképzett, stb. láncokat. A CC példából látható, hogy a gráfokban megszakításokat is megengedtem, hisz például a magánhangzó harmónia esetén a magyar nyelvben is a releváns összetevők nem közvetlenül követik egymást.



1. ábra. Néhány lehetséges gráf, amely az összehasonlítás alapját képezheti.

Ez az algoritmus az edit distance algoritmussal ellentétben azonban hangsúlyt ad a fonémák hasonlósági jegyeinek. Elvárásaim szerint jó teljesítményt hozhatott volna, mivel algoritmusom fő gyengéjének korábban a magánhangzó harmónia és a szótag-szerkezet iránt mutatott kisebb érzékenység mutatkozott. Azonban a sorrendiséggel szemben való semlegessége olyan hátránynak bizonyult, amelynek köszönhetően még az edit distance algoritmusnál is rosszabbul teljesített.

4 Adataim

Korpuszalapú vizsgálataimban a Szószablya webkorpusz [3] gyakorisági adatait használtam fel, amelyet azért választottam, mert jelenleg ez a legnagyobb, mintegy 19,1 millió szóalakot tartalmazó korpusz, amely 3,493 millió weboldal és 1,486 milliárd szó alapján készült. Tövek helyett szóalakokat, elnagyolt gyakorisági kategóriák helyett pedig pontos gyakoriságú számokat tartalmaz, amely egyedülálló módon

alkalmassá teszi nyelvészeti és nyelvtechnológiai kutatásokra. A Szószablya webkorpusz további nagy erénye, hogy válogatatlan, sokszor a beszélt nyelvhez közelibb nyelvhasználatot és írásmódot rögzítő anyagokat (fórumok, blogok) is nagy arányban tartalmaz, így az ez alapján tett megállapításaink is jobban közelíthetik a magyar nyelvi valóságot.

Vizsgálatomhoz a hangkivető főneveket választottam. Választásom több okból kifolyólag esett rájuk. A magyar hangkivető tövek habár zárt osztályt alkotnak, meglehetősen nagyszámúak, így a viselkedésükből levonható következtetések nem szórványos, egyedi és ritka adatokon alapszanak. A hangkivető főnevek viselkedésére jellemző a fokozatos ingadozás, amelyről még a szabályalapú megközelítések engedékenyebb változatai sem tudnak teljesen számot adni.

A hangkivető töveket kiindulási pontnak az is kiválóan alkalmassá teszi, hogy korábban alapos leírást készített róluk Rebrus Péter[7] a kormányzásfonológia eszköztárát használva, amely jó viszonyítási alapot képez vizsgálódásaimhoz.

A Szószablya webkorpuszon túl vizsgálatomban még a BME MOKK morphdb.hu szótárára támaszkodok, amely jelenleg a legnagyobb, ingyenesen is hozzáférhető nyelvi adatbázis (130 ezer tő, [10]), ahonnan összesen 1205 hangkivető főnévi tövet választottam ki. Ezekből összesen 1097 volt a szótárban hangkivetőként megjelölve, amelyekből kivettem a *kelet, sportberkek, sodor, terem* szavakat. A *kelet* szó egyértelműen a *kelte* szóalak miatt került be hangkivetőként rögzítve, ahelyett, hogy már ragozott főnévként vették volna fel a szótárba. A *sportberkek* már szóalak, és nem tő, helyesen a szótárban *sportberek*-ként kellene szerepelnie, amelynek hiányos a paradigmája. A *sodor* és a *terem* szavak valóban hangkivető főnevek, de mivel alanyesetük és számos további ragozott alakjuk egybeesik a náluk sokkal gyakoribb *sodor* és *terem* igei tövek alakjaival, ezért célszerűbbnek tartottam ezek kihagyását a vizsgálatból. Úgy véltem, hogy elegendő nyelvi adat birtokában ezek elhagyása nem vezet az eredmények jelentős módosulásához.

A szótárban hangkivetőként megadott töveken túl további 102 tövet választottam ki, amelyek valóban hangkivetők és szerepelnek is a szótárban, de nincsenek hangkivetőként megjelölve. Ezek a szavak a hangkivetőként megjelölt 1093 szótőből létrehozott összetett szavak, amelyekben a hangkivető tő az összetétel jobb oldali záró tagját adja. Az adatbázis mindösszesen 4 tövet ad meg ingadozónak (hangkivetés a megfelelő ragok előtt a Szószablya webkorpusz alapján *bajusz*: 35%, *fókabajusz*: nincs adat, *harcsabajusz*: 57%, *macskabajusz*: 25%), amelyeket jól azonosít.

A korpuszra támaszkodva a morphdb.hu szótára több ponton is javítható lenne, amelynek szóanyaga több korábbi szótár automatikus módszerekkel végrehajtott egyesítésével jött létre. A morphdb.hu más szótári adatbázisokhoz hasonlóan nem kezeli a nyelvre jellemző ingadozást. Szótáraink, mint láthattuk legjobb esetben is csak megjelölik az ingadozást, de annak mértékéről nem tudnak számot adni. Ezzel tulajdonképpen azon szabályalapú nyelvelméletek gyakorlatát követik, amelyek készítésükre a legnagyobb hatást gyakorolták. Elsősorban a hibás besorolások és az ingadozás helyes megadásával lehetne javítani az adatbázisok minőségét.

Az ingadozás mértékének jelölése a morfológiai elemzésnek csak bizonyos, nem túl gyakori eseteinél lehetne szükséges, ahol egy kiegyenlített alak egybeeshet egy másik szó alakjával (pl. *sodort* = *sodor* ige + múlt idő vagy *sodor* főnév + tárgyeset, *kéreget* = *kér* + *gAt* vagy *kéreg* + tárgyeset). Az ingadozás mértékének megadása elsősorban egy olyan adatbázisban nyerne létjogosultságot, amelyet már szóalakok

produkciójára is jól lehet használni, hisz nem mindegy, hogy adott esetben a ritkább hangkivetéses alakot használjuk vagy a már analógiásan kiegyenlített gyakoribb alakot. Vélhetőleg az ingadozás mértéke nyelvi regiszterenként is eltérő lehet, de ennek megállapítására a Szószablya webkorpusz alkalmatlan forrás.

Az egyes tövek ingadozását az összes rag, jel jellegű toldalék előtt megvizsgáltam, amelyek hangkivetést válthatnak ki: tárgyeset, szuperesszívusz, többes szám, birtokos személyragok. A Szószablya webkorpusz alapján megállapítható, hogy a *pityer*, *szlalom*, *vicikvacak* szavak már nem hangkivetők, míg további 114 tő tekinthető ingadozónak, mert ezeknél az esetek legalább 1%-ban a hangkivetést kiváltó toldalékok előtt nem történik meg a hangkivetés. 43 tőnél ez az arány meghaladja a 10%-ot, 15-nél pedig több, mint 50%. Habár célszerű lenne ezeknek az adatoknak az alapján szótárunkat frissíteni, 2009-es Google lekérdezések alapján látható (.hu domain alatt), hogy az analógiás kiegyenlítődés tovább folytatódik (pl. *fátyolt* aránya 2003-ban 67%-ban hangkivető, 2009-ben már 42% a *fátylat* helyett). Természetesen ezen folyamatok pontos leírása és értékelése külön vizsgálatot érdemel az összes alak figyelembevételével.

Adataimat vizsgálataimhoz átkonvertáltam egy olyan írásrendszerbe, ahol egy fonémának egy betű felel meg. Az egyes alakokban itt a szóbelseji zöngésedési folyamatoknak megfelelő fonémát tüntetem fel, amelyeket eredetileg az íráskép nem rögzít, így lesz a *virágcsokor*ból *virákčokor*.

5 Kísérlet és eredmények

A szóhasonlóságot megállapító algoritmusom pontosságát olyan tesztekkel ellenőriztem, amelyek során valós magyar szavakat sorolok be már meglévő szócsoportokba. A besorolások helyessége alapján látható, hogy egy algoritmus mennyire jól ragadja meg azt a feltételezett nyelvi képességet, amely alapján analógiás hasonlítások elvégzésére képesek vagyunk.

Első számításaim a magyar helységnevek lokatívuszaival végeztem, amelynek során azt tapasztaltam, hogy az analógiás keretrendszer jól meg tudja ragadni ezek viselkedését, ellentétben a szabályalapú megközelítésekkel, amelyek általánosításainak a valós adatok nem egyszer ellent mondanak.

A leggyakoribb 100-100 harmónia szempontjából megfelelő alak alapján meghatároztam (100-100 *–bAn*, 100-100 *–(V)n* végű), hogy a következő 40 leggyakoribb alak szuperesszívuszt vagy inesszívuszt vár-e el. A szavak szótári és ragozott alakjai alapján 87,5%-os pontossággal választotta ki algoritmusom a megfelelő szócsoportot. Ritkább alakok esetében már az anyanyelvi beszélők ítéletei is ingadoznak, így ez a 87,5%-os teljesítmény megközelíti az ő eredményeiket, az edit distance teljesítményét pedig messze meghaladja.

Eredményeim megerősítése végett a hangkivető tövekkel is elvégeztem egy a korábbival megegyező vizsgálatot, amelyben az edit distance algoritmus, és a gráfi hasonlóságot figyelembe vevő algoritmus eredményességét hasonlítottam össze saját algoritmusommal.

Az 1205 hangkivető tőből sorrendben az 501. leggyakoribb tőtől a 600. tőig megvizsgáltam, hogy ha egy már meglévő szólistához hasonlítom ezeket a töveket, akkor

milyen pontossággal találnak az egyes algoritmusok a szólistában szereplő hangkivető tövet hasonlósági alapon. A 100 hangkivető tőből összesen 7-nél a hangkivetés elmaradása a releváns toldalékok előtt meghaladta a 10%-ot (*hatökör, ködfátyol, lombsátor, sulyok, tündérfátyol, szalmakazal, zsákvászon*), míg a listán szereplő *pityer*nél a korpusz alapú vizsgálatok alapján láttuk, hogy az analógiás kiegyenlítődés befejeződött vagy befejeződés közeli állapotban van.

A hangkivető tövekhez kontroll csoportként véletlenszerűen kiválasztott, velük azonos gyakoriságú 100 nem hangkivető tövet vettem. Ezek ragozatlan alakjainak a korpusz 93 és 57 közti előfordulást adott meg, azaz ritka, de még használt és valamelyest ismert szavakról van szó. A szavak gyakoriságát besorolásaimhoz minden esetben ragozatlan alakjaik alapján vettem. Ez némileg eltérhet a szó összes alakjai alapján számított gyakoriságától, mégis alkalmazhatjuk ezeket a számokat besorolásukhoz. A hangkivető szavaknak mind ragozatlan alakjairól, mind összes előforduló alakjukról pontos adataim vannak a korpusz alapján, és a kétféle módon számított gyakoriság közt igen magas, 0,758-as korreláció figyelhető meg.

A hangkivető és nem hangkivető tesztszavakat összesen 4 eltérő méretű szólistához hasonlítottam. Ezekben az 50, 100, 200 illetve 500 leggyakoribb hangkivető tő, illetve az ezekkel egyenlő vagy nagyobb gyakoriságú nem hangkivető főnévi tövek szerepeltek, amely listák pontos méretét az 1. táblázat adja meg. Mint látható a hangkivetők aránya a tőszámmal együtt egyenletesen nő, de nem változik olyan radikális mértékben, hogy az egy vizsgálat eredményére jelentős kihatással lehessen.

1. táblázat: Szólisták száma és a hangkivető tövek aránya ezekben.

Hangkivetők száma	Tőszám	hangkivető tövek aránya
50	2828	1,7%
100	5468	1,8%
200	10315	1,9%
500	15333	3,2%
1205 (összes tő)	55762 (összes tő)	1,8%

A teszt során a két 100-100 darabos szócsoportot a nagyobb szólistákhoz hasonlítottam, amelynek eredményét a 2. táblázat mutatja. A százalékok arra utalnak, hogy a 100 tőből hány százalékban választott az adott algoritmus az adott listából azonos típusú tövet. Amennyiben hangkivető tőt kellett választanunk, úgy a találgatás küszöbe 1,7 és 3,2% közt lett volna. Ezt láthatjuk, hogy minden esetben sikerült a vizsgált algoritmusoknak meghaladnia. A nem hangkivető tövek esetében ez a szám jóval magasabb 96,8-98,3%, hisz ezek a tövek jóval nagyobb arányban voltak képviselve a szólistákban, így véletlenszerű kiválasztásukra is nagy esély lett volna. Ezt a szintet egyedül saját algoritmusom haladta meg, azonban csak a legkisebb 50-es szólista esetén, amikor azonban hibátlanul teljesített. Mivel a leggyakoribb tövekhez hasonlítanak algoritmusaim, a gyakorisági szempontok is szerepet kapnak, de csak mérsékelten, hisz a gyakoribb tövek közt a nagyon gyakori és a kevésbé gyakori tő már egyforma súllyal bír.

2. táblázat: Az egyes algoritmusok eredményessége a szavak összehasonlításában.

Szólisták	Edit distance, hangkivető	edit distance, nem hangkivető	saját algoritmus, hangkivető	saját algoritmus, nem hangkivető	gráf alapú, hangkivető
50 hangkivető	39%	98%	51%	100%	7%
100 hangkivető	75%	93%	73%	97%	14%
200 hangkivető	64%	98%	84%	97%	
500 hangkivető	63%	100%	95%	98%	

A gráf alapú algoritmussal a 200 és az 500 hangkivető alakot tartalmazó szólista esetében nem végeztem el az összehasonlításokat, mert az algoritmus jelenlegi implementációja nem teszi lehetővé belátható időn belül ekkora adattömeg összehasonlítását. Kihagytam a táblázatból a nem hangkivető tövekkel való összehasonlítást is ennek az algoritmusnak az alapján, mivel a hangkivetőkkel való összehasonlítás során már megmutatkozott, hogy az algoritmus jelen formájában nem tud megfelelő eredményt hozni.

A táblázat alapján látszik, hogy saját algoritmusom nagy mennyiségű adattal összesen 95%-os, illetve 98%-os eredményt hozott. Eredményeim azt mutatják, hogy algoritmusom megfelelő hatékonysággal tud emberi beavatkozás nélkül is szavakat megfelelően besorolni, amely egybecseng korábbi tapasztalataimmal. A nagy számok természetesen relatívak, hisz a hasonlításhoz felhasznált szavak mennyisége még így is csak a negyede annak, amivel szótárunkban rendelkezünk.

Algoritmusom 5 esetben sorolta be rosszul a következő hangkivető töveket: *pityer, bugyor, orrnyereg, lombsátor, csöbör*. A *pityer* esetében nem beszélhetünk hibázásról, hisz ezt a besorolást a korpusz adatai is támogatják. Ha a *bugyor* (leghasonlóbb *hunyor*) , orrnyereg (leghasonlóbb *hadsereg*) és csöbör (leghasonlóbb *csömör*) esetében a hozzájuk 10 leghasonlóbb szót vesszük, akkor azt figyelhetjük meg, hogy ezek közt már van 3, 2 illetve 4 hangkivető szó. Azaz az algoritmus felfedezi a hangkivető tövekhez a hasonlóságot, csak nem ad ezeknek megfelelő súlyt. Egyedül a *lombsátor*hoz nem talált megfelelő hangkivető szót még az első 10 közt sem, ami jól tükrözi, hogy a *lombsátor* szó ingadozik, de az algoritmus ítélete túlzó. Az algoritmus következetesen, de tévesen az *-átor* végű latin eredetű szavakhoz hasonlítja: *pankrátor, diktátor, organizátor* stb. Még a rontott példákból is látszik, hogy az algoritmus ilyenkor is jól közelíti a hasonlóságot, de teljesítményét célszerűbb lenne nem csak egy választott alak alapján kiértékelni. Az algoritmusnak két tulajdonsága, miszerint hátulról számol, illetve meglehetősen engedékeny egy szekvencián belül kisebb eltérésekre, alkalmassá teszi, hogy hatékonyan hasonlítson.

A 100 nem hangkivető tőhöz való hasonlítás során az algoritmus két hibát követett el: *bikacsök:bütyök* illetve *csucsor:csupor*. Az első esetben a korábbi tesztelések során is tapasztalt hibát figyelhetjük meg, miszerint az algoritmus nem elég érzékeny a hangrendi harmóniára, hisz a második magánhangzó már elég távol van a szó végétől, hogy kis súlyt kapjon. Ezért nem zavarja az algoritmust az *aö* szekvencia hasonlítása az *öö*-höz. A sokkal megfelelőbb jelölt, a *lopótök* csak a 10. leghasonlóbb szónak kerül elő.

Az edit distance algoritmus gyengébb teljesítménye egyértelműen a már leírt hiányosságaira vezethető vissza, a gráf alapú algoritmus pedig jelenlegi implementáció-

jában leginkább az azonos hosszúságú szavakat választja, amely szintén többnyire rossz választáshoz vezet.

Természetesen felmerülhet a kérdés, hogy az ilyen jellegű besorolás mennyire használható a szótárbővítésben, hisz a hangkivető tövek zárt csoportot alkotnak, amely nem bővíthető tovább. Látnunk kell azonban, hogy a szótárbővítés valójában nem új szavak besorolása egy szótári csoportba, hisz ezek a szavak szótárunktól is függetlenül már hangkivetők vagy sem. Esetünkben csak az történik, hogy ezek hangkivető voltát „felismerjük". Igen sok szó van, amelyek besorolása a digitális szótárakba még nem történt meg. Ezek esetében is hasznos lehet az automatikus, de a valós folyamatokhoz közeli besorolás, amely nem alapulhat csak azon, hogy egy új szó esetleg valamely a szótárban már meglévő szóból létrehozott összetett szó-e (lásd *lé:levet*, de *baracklé:baracklét/baracklevet*).

Másrészt ha egy szócsoportot zártnak is veszünk, nem kizárt, hogy a valóságban, ha elég nagy analógiás erővel bír, be tud vonzani új szavakat, mint például a *motrok, bútrok* alakok esetében, amely adatokat gyakran félresöprik, de mégsem hagyhatjuk ezeket figyelmen kívül, mert a nyelvi változás lényegéről beszélnek nekünk. Egy szó besorolása alapvetően hasonló feladat, mint amikor egy szónak egy alakját hozzuk létre beszéd közben, ha már hallottuk ezt az alakot vagy egyenesen nagy gyakoriságú is, akkor jó esélyünk van arra, hogy az „elvárt", hangkivető alakot ejtjük ki. Azaz a gyakoribb hangkivető töveknél nagyobb az esély hangkivető változatok létrehozására, hisz ott több minta mutatja ezt.

6 További kutatási lehetőségek

Habár algoritmusom immár két vizsgálatban is sikeresen bizonyította, hogy elegendő nyelvi minta birtokában hatékonyan tud analógiás párokat találni, számos lehetőség van továbbfejlesztésére úgy, mint nyelvtechnológiai eszköz, és úgy is mint a nyelvi folyamatokat reálisan modelláló algoritmus. Elsősorban a magánhangzó harmónia és a szótagszerkezet iránti érzékenységét lenne érdemes növelni. Erre a célra lehet alkalmas az egyébként nem olyan jól teljesítő gráf alapú algoritmussal való ötvözése. A két nyelvi jelenséggel való vizsgálat már sokat elárult természetéről, de célszerű lenne még további nyelvi jelenségeken is megvizsgálni hatékonyságát (tőbelseji magánhangzó rövidülés, *v*-vel való bővülés).

A rendszer látszólagos hibázásainak felderítése közben korábbi és jelenlegi kutatásaimban is az körvonalazódott, hogy jobb eredményt kaphatnánk, ha az analógiás hasonlításnál nem feltétlenül egy szóhoz, hanem egy valamilyen szempontból konzisztens csoporthoz hasonlítjuk szavainkat, amelyet klaszterezéssel lehetne felderíteni. Ezzel párhuzamosan fel kellene térképezni az egyes szavakra ható analógiás nyomást, amely mentén egy adott szó részt vesz az analógiás folyamatokban. Egy ilyen vizsgálatban a gyakoriságnak már kiemelkedő szerepe lenne, amelynek azonban a jelenlegi irányvonal továbbfejlesztésében is nagyobb szerepet kellene kapnia.

Hivatkozások

1. Bybee, J. L. : Phonology and Language Use, CUP, Cambridge (2001)
2. Goldberg, A.: Constructions. A Construction Grammar approach to argument structure, University of Chicago Press, Chicago (1995)
3. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: Creating Open Language Resources for Hungarian, In: Proceedings of LREC. (2004) 1201–1204
4. Kiefer F.(szerk.): Strukturális Magyar Nyelvtan 2. Fonológia. Akadémiai Kiadó, Budapest (1994)
5. Kraska-Szlenk, I.: Analogy. The relation between Lexicon and Grammar. Lincom, Muenchen (2007)
6. Levenshtein, V.: I. Binary codes capable of correcting deletions, insertions, and reversals, Doklady Akademii Nauk SSSR, 163(4): 845–848 (Russian). English translation in Soviet Physics Doklady, 10(8): (1965-1966) 707–710
7. Rebrus, P.: Morfofonológiai jelenségek. In: Kiefer F. (szerk.) Strukturális magyar nyelvtan 3. Morfológia. Akadémiai Kiadó, Budapest (2000) 763–947
8. Rung, A.: Determining word similarity in the Hungarian language. Papers from the Mókus Conference. Tinta Kiadó, Budapest (2008) 112–118
9. Skousen R.: Analogical Modeling of Language Kluwer Academic Publishers, Dordrecht Boston London (1989)
10. 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. ELRA (2006) 1670–1673

III. Korpusz, ontológia, lexikográfia

A szótárkészítés támogatása párhuzamos korpuszokon végzett szóillesztéssel

Héja Enikő

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
1068, Bp., Benczúr u. 33.
eheja@nytud.hu

Kivonat: Cikkünkben egy kétnyelvű szótárak készítésének gépi támogatására irányuló módszert ismertetünk. A javasolt megközelítés alapja a párhuzamos korpuszokon végzett automatikus szóillesztés. A korpuszvezérelt megközelítés, ezen belül különösen a párhuzamos korpusz használata több okból is hasznosnak bizonyult a lexikográfia számára. Ezek közül a legfontosabb, hogy – megfelelő méretű reprezentatív korpusz használatával – a javasolt megközelítés garantálja, hogy a legrelevánsabb fordítások fognak szerepelni a szótárban. További előnyt jelent, hogy az összes korpuszbeli példamondat könnyedén hozzáférhető, így a poliszém jelentések közül nagy mennyiségű természetes adat alapján választhatjuk ki a legmegfelelőbbet. A két fenti tulajdonság különösen alkalmassá teszi az általunk javasolt módszert aktív szótárak előállítására.

1 Bevezetés

A cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX[1] projekt része. A projekt azt vizsgálta, 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 íródott szótárakra alacsony a kereslet, így az ilyen munkálatok finanszírozása is korlátozott. A bemutatandó munka eredeti célja egy középméretű (kb. 15,000 szócikk) litván-magyar szótár létrehozása volt. A munkafolyamat részeként tesztelési célokra a magyar-szlovén nyelvpárt is vizsgáltuk.

Jelenleg nem létezik olyan módszer, amely lehetővé tenné szótárak *teljesen* automatikus előállítását. Így 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.

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

Az általunk javasolt módszer alapját párhuzamos korpuszokon végzett automatikus szóillesztés képezi. Bár az automatikus szóillesztést széles körben használják szótárfejlesztésre elsősorban a gépi fordítás területén, amennyire tudjuk, mostanáig ezt a megközelítést nem használták lexikográfiai projektekben emberi felhasználásra szánt szótárak készítésének támogatására.

A következő fejezetben röviden bemutatjuk a párhuzamos korpusz használatának előnyeit a szótárkészítésben. A 3. fejezetben áttekintjük a munkafolyamatot, amely három fő részből áll: a párhuzamos korpuszok létrehozásából (3.1), a protoszótárak előállításából (3.2) és a kiértékelésből (3.3). A 4. fejezetben illusztráljuk, hogy az általunk javasolt módszer jól kezeli a poliszémiát. Az utolsó szakasz összefoglalja az eredményeket és a hátralévő feladatokat.

2 A javasolt módszer előnyei és hátrányai

Napjainkban általánosan elfogadott lexikográfus körökben, hogy jó minőségű szótárakat kizárólag korpusz alapján lehet létrehozni (ld. pl. [1]). Ennek az az oka, hogy a korpusz használata jelentősen csökkenti az egyéni intuíció szerepét a szótárkészítési folyamatban.

Azonban még forrásnyelvi és célnyelvi korpusz használata esetén is vannak olyan lépések, amelyek során a szótárkészítőnek elkerülhetetlenül támaszkodnia kell az intuíciójára. Ilyen feladatok például a szótárba felveendő jelentéssel bíró nyelvi egységek (*linguistic unit*, a tovbbiakban LU) meghatározása, ezek célnyelvre való fordítása valamint annak eldöntése, hogy a lefordított LU-k közül melyeket vonják össze célnyelvi oldalon.

A költséghatékonyság mellett a párhuzamos korpusz alapú módszer másik nagy előnye, hogy választ ad arra a kérdésre, hogy hogyan csökkenthető tovább az intuíció szerepe a lexikográfiában. Ebben az esetben az LU-kat nem a lexikográfusok nyerik ki a korpuszból, hanem a statisztikai szóillesztő algoritmus. Így nem az emberi intuíció határozza meg, hogy mi számít egy LU-nak, hanem a szavak kontextusa és célnyelvi fordítása. A megközelítés korpuszvezérelt (*corpus-driven*) jellege biztosítja, hogy a fordításjelöltek megállapítása során a gyakran használt fordítások nagyobb súllyal szerepeljenek. Így – ha rendelkezésünkre áll egy nagyméretű, reprezentatív korpusz – a leggyakrabban használt fordítások biztosan szerepelni fognak szótárunkban.

A megközelítés további előnye, hogy az automatikusan meghatározott fordítási valószínűség alapján a fordítási jelölteket sorba rendezhetjük aszerint, hogy mennyire valószínű fordításai a forrásnyelvi lemmának. Így korpuszadatok alapján megállapíthatjuk, hogy melyik egy lemma leggyakoribb jelentése. A módszer harmadik fontos jellemzője, hogy a fordításpárok természetes kontextusaikkal együtt jeleníthetőek meg. Az alábbi magyar-litván bejegyzésminta azt illusztrálja, hogy a korpuszból származó példamondatok alapján hogyan különíthetjük el egymástól egy szó különböző aljelentéseit.

1. táblázat: Bejegyzésminta a magyar-litván protoszótárból.

MAGYAR LEMMA	LITVÁN LEMMA	FORD. VSZ.	MAGYAR LEMMA-GYAKORISÁG	LITVÁN LEMMA-GYAKORISÁG
Születik	Gimti(-sta,-ė)	0.579	169	174
MAGYAR			**LITVÁN**	
Ő **1870-ben született**.			Jis **gimė 1870 metais**.	
De Fache mintha **erre született** volna.			Bet Fasas, regis, tiesiog **tam gimęs**.	
Úgy látszik, **szerencsétlen csillagzat alatt születtél**.			Turbūt **gimei po nelaiminga žvaigžde**.	
…, mert **ikrei születtek.**			…, nes jai **gimė dvynukai**.	
Maga úriembernek **született.**			Tu **gimei džentlemanu.**	
… hogy Buddha nem **lótuszvirágból született**?			…, kad Buda **gimė** ne **iš lotoso žiedo**?	

A javasolt módszer egyik hátránya, hogy kizárólag lemmák között hozható létre megfeleltetés, így jelenlegi formájában a többszavas kifejezések (nevek, kollokációk, igei szerkezetek) automatikus kezelésére alkalmatlan. A másik fő nehézség, mint azt rövidesen látni is fogjuk, hogy a párhuzamos korpusz összeállítása a kevéssé használt nyelvekre rendkívül időigényes feladat.

A következő részben a magyar-szlovén és magyar-litván protoszótárak elkészítését mutatjuk be.

3 A munkafolyamat

A munkafolyamat három fő szakaszból áll. Először a szükséges erőforrásokat és a szövegfeldolgozáshoz szükséges nyelvspecifikus eszközöket szereztük be (ld. 3.1). Ezt követően a szóillesztés segítségével és különböző szűrők alkalmazásával létrehoztuk a protoszótárakat (ld. 3.2). Az utolsó szakaszban kidolgoztuk a kiértékeléshez szükséges kategóriákat, majd elvégeztük a kiértékelést (3.3).

3.1 A párhuzamos korpuszok létrehozása

Erőforrások és nyelvspecifikus eszközök
Mivel a projekt célja a köznapi szókincset lefedő protoszótárak létrehozása volt, a szövegek gyűjtésekor a regényekre koncentráltunk. A projekt során felmerülő legnagyobb nehézséget a megfelelő mennyiségű, általános szókincsű erőforrás összegyűjtése okozta. Mivel a szlovén-magyar nyelvpár közötti közvetlen fordításokból nagy ráfordítással csak kevés szöveget[2] sikerült szerezni, és a litván-magyar nyelvpárra nem találtunk nagy mennyiségű közvetlen fordítást, úgy döntöttünk, hogy a litván-magyar párhuzamos korpuszt olyan szövegekből állítjuk össze, amelyeket egy harmadik nyelvről fordítottak le mindkét nyelvre. Sajnos azonban sem a szlovén, sem a litván esetében nem állnak rendelkezésre olyan digitális archívumok, mint a Digitális Irodalmi Akadémia[3] és a Magyar Elektronikus Könyvtár[4] a magyar vonatkozásában. Ezért a litván Vytautas Magnus Egyetemen található Számítógépes Nyelvészeti Központ segítségét vettük igénybe. Az intézmény a Litván Nemzeti Korpusz [9] és egy angol-litván párhuzamos korpusz [8] létrehozójaként birtokában van a projekt szempontjából szükséges erőforrásoknak és nyelvspecifikus eszközöknek.

A szótárkészítéshez szükséges szövegfeldolgozó eszközöket (tokenizáló, mondatra bontó, lemmatizáló – egyértelműsítéssel) eszközláncokba beépítve használtuk. A litván elemzést a már említett Számítógépes Nyelvészeti Központ (Vytautas Magnus Egyetem) végezte el. A szlovén szövegeket a Jožef Stefan Intézet honlapján[5] található eszközlánccal elemeztük [4]. A magyar korpusz annotálása pedig a Nyelvtudományi Intézet Nyelvtechnológiai Osztályán kifejlesztett MNSZ egyértelműsítő lánccal történt [7].

A párhuzamos korpuszok létrehozása
A mondatillesztést a *hunalign* mondatillesztővel [10] végeztük. Az illesztés bemeneteként a mondatok lemmatizált változata szerepelt, hogy a gazdag morfológiából fakadó adathiányt a lehető legkisebbre csökkentsük.

Mivel az eredeti feladat a protoszótárak előállítása és hasznosíthatóságuk vizsgálata volt, a rossz mondatillesztés esetleges hatásainak minimalizálására törekedtünk. Ezért először a szövegeket kézileg ellenőriztük, hogy kiszűrjük azokat a szövegrészeket[6], amelyeknek nincsen célnyelvi megfelelőjük. Az illesztés után a szlovén-magyar párhuzamos korpusz egy részkorpuszán a mondatpárokhoz rendelt konfidenciaértékek alapján megállapítottuk, hogy mi az a küszöbérték, amely felett nagy eséllyel már csak jó mondatillesztések vannak. A litván-magyar párhuzamos korpusz esetén is az itt megállapított értéket használtuk. Az 2. táblázatban az eredményül kapott párhuzamos korpuszok mérete szerepel.

---

[2] A szlovén televízió, számos műfordító és kiadó megkeresésével kb. egy 750.000 tokent tartalmazó korpuszt gyűjtöttünk.

[3] http://www.pim.hu/

[4] http://mek.oszk.hu/

[5] http://nl.ijs.si/jos/analyse

[6] A munka elvégzéséért köszönet illeti Mittelholcz Iványt.

2. táblázat: A párhuzamos korpuszok mérete.

LITVÁN	1,765,000 token	147.158 TU[7]
MAGYAR	2,121,000 token	147,158 TU
SZLOVÉN	733,000 token	38,574 TU
MAGYAR	666,000 token	38,574 TU

3.2 A magyar-szlovén és a magyar-litván protoszótárak létrehozása

A protoszótárak generálásának két fő szakasza volt. Első lépésben elvégeztük a szóillesztést. Erre a célra a GIZA++ szóillesztő szoftvert [6] használtuk. A GIZA++ a szóillesztés során fordításjelölteket hoz létre, úgy, hogy a célnyelvi és a forrásnyelvi lemmapárokhoz fordítási valószínűséget rendel. A protoszótárak kiindulási alapját ezek a fordításjelöltek képezték. Ezekből kellett kiszűrnünk a legjobb fordításjelölteket a lehető legtöbb helyes fordításjelölt megtartásával.

A második lépésben tehát ezt a feladatot kívántuk megoldani a magyar-szlovén eredmények egy mintájának kézi kiértékelésével[8]. A kiértékelés során három paramétert vettünk figyelembe: a GIZA++ által meghatározott *fordítási valószínűségnek* az értékét, valamint a *forrásnyelvi* és *célnyelvi jelöltek korpuszgyakoriságát*. Ez az előzetes kiértékelés két konklúzióval szolgált: egyfelől a fordításpár-jelöltben szereplő lemmák mindegyikének legalább 5-ször elő kell fordulnia ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez. Másfelől, a kiértékelt szópárok azt mutatják, hogy a fordítási valószínűségnek legalább 0.5-nek kell lennie. Ez alatt az érték alatt rohamosan csökken a fordításjelöltek pontossága. A fenti paramétereknek megfelelő fordításjelöltek 65%-a volt jó fordítás[9]

3. táblázat: A megfelelő fordításpár-jelöltek a teljes korpuszon.

	A megfelelő fordításjelöltek száma	A várhatóan jó fordításjelöltek száma
Magyar-szlovén	4969	3230
Magyar-litván	4025	2616

[7] *TU* (translation unit) kifejezést használjuk az illesztett egységek jelölésére, mert a *hunalign* engedélyezi a mondatok közötti egy-a többhöz megfeleltetéseket is.

[8] A szlovén-magyar szövegek gyűjtéséért és a magyar-szlovén kiértékelési munkák elvégzéséért köszönettel tartozom Sárossy Bencének.

[9] A magyar-litván esetében egy további korlátozást is bevezettünk: a fordításjelöltek közül kizártuk azokat a párokat, amelyek valamelyik tagjának gyakorisága több, mint 100-szorosa volt a másik tag gyakoriságának.

Mivel célunk nem tökéletes szótárak automatikus előállítása volt, hanem olyan protoszótárak készítése, amelyek a lehető legnagyobb mértékben segítik a lexikográfusok munkáját, jogosnak tűnik egy 65% körüli pontosságot becélozni, mivel könnyebb már meglévő, ám rossz fordításjelölteket kidobni, mint újakat felvenni a szótárba. Így ezeket a paramétereket elfogadva részletesen is kiértékeltük a magyar-litván protoszótárunkat. Ezt mutatja be a következő fejezet.

3.3 A magyar-litván protoszótár részletes kiértékelése

A magyar-litván protoszótár kiértékelését teljesen manuálisan végezték mindkét nyelvet egyaránt beszélő szakértők[10]. Az általánosan elfogadott kiértékelési eljárásokkal szemben itt elsődlegesen nem a *jó* és a *rossz* fordításjelöltek arányára voltunk kíváncsiak, hanem a *lexikográfiailag hasznos* és a *lexikográfiailag nem hasznos* fordításjelöltekére. Ezt a fajta megkülönböztetést egyrészt az olyan jó fordításjelöltek tették szükségessé, amelyek szótárkészítési szempontból irrelevánsak (elsősorban a túl specifikus tulajdonnevek). Másrészt a rossz fordításjelöltek komoly segítséget jelenthetnek a szótárkészítési munkában, elsősorban a kollokációk esetében, hiszen a kontextus alapján könnyű visszafejteni, hogy mi lett volna a helyes megfeleltetés. Az alábbiakban röviden összefoglaljuk azokat a fő kategóriákat (3.3.1), amelyeket a kiértékelés során használtunk, majd ismertetjük a kiértékelés módszertanát és az eredményeket (3.3.2).

A kiértékelés során használt fő kategóriák[11]

(1) Lexikográfiai szempontból hasznos fordításjelöltek:

a. Teljesen jó fordítások [H: ***gyümölcs*** – L: ***vaisius*** – *fruit*]

b. Részlegesen jó fordítások, ebben az esetben utószerkesztés szükséges, elsősorban az alábbiak miatt:
   i. rossz lemmatizáció
   ii. részleges/rossz illeszkedés a több szavas kifejezések esetében.

   Pl:
   1. összetett szavak [H: ***főfelügyelő*** – L: *vyriausiasis* ***inspektorius***],
   2. kollokációk [H: **bíborosi** testület – L: Kardinolų **kolegiją**]

c. Egyéb szemantikai viszony. Pl: hiperonímia [H: **lúdtoll** – L: **plunksna** (toll – madártoll, írótoll)])

[10] A magyar-litván szótár kiértékeléséért köszönet illeti Tölgyesi Beatrixot és Justina Lukaseviciute-t.

[11] A megadott példákban az automatikusan megállapított jelöltpárok félkövérrel vannak szedve.

(2) Lexikográfiai szempontból nem hasznos fordításjelöltek

a. Irreleváns szókincs (pl. gyakran előforduló tulajdonnevek [H: **Abdul** – L: **Abdulas**])
b. Rossz fordítások (általában a túl szabad fordítás miatt)

A kiértékelés eredménye

A fent meghatározott paramétereknek megfelelő[12] 4025 magyar-litván fordításjelölt közül 863 párt értékeltünk ki kézzel. Ebből 520 pár fordítási valószínűsége a [0,5, 0,7) tartományba, 380 pár fordítási valószínűsége pedig a [0,7, 1) tartományba esett. 63 pár fordítási valószínűsége volt 1. A kiértékelés eredményeit a 4. táblázat tartalmazza.

4. táblázat: A magyar-litván szótár kiértékelésének eredményei.

	Hasznos párok		**Nem hasznos párok**	
P(tr)1	OK	Utószerk.	Irreleváns	Rossz
[0.5,	52.1 %	32.9 %	2.3 %	12.7 %
	85 %		15 %	
[0.7, 1)	65.3 %	31.9 %	0.6 %	2.2 %
	97, 2 %		2,8%	
1	38 %	13 %	49 %	0 %

A 4. táblázat alapján a fordításjelöltek 85%-a hasznos a [0.5, 0.7) valószínűségi tartományba eső fordításpárok esetén. Ez az arány még jobb a [0,7, 1) intervallumba eső fordítási valószínűségek esetén (97,2%). Érdekes módon az 1 fordítási valószínűséggel rendelkező párok esetén ez az arány csupán 38%. Az irreleváns párok magas aránya (49%) azt mutatja, hogy ennek elsődleges oka, hogy a nevek hajlamosabbak 1 valószínűséggel együtt előfordulni.

4 A poliszémia kezelése a javasolt módszerrel

Mint már a cikk 2. szakaszában a megközelítés előnyei között említettük, az általunk javasolt módszerrel a korpuszból az összes releváns fordítást kinyerhetjük, ezáltal csökkentve a fordítói intuíció szerepét. Sőt, ezen felül a lehetséges fordítások rende-

[12] Mindkét lemma legalább ötször előfordul, a fordítási valószínűség legalább 0,5 és egyik lemma sem fordul elő százszor többször, mint a másik.

zésével elérhetjük, hogy a szó legvalószínűbb jelentéseit rangsoroljuk előre. Ezek alapján azt várjuk, hogy az általunk javasolt megközelítéssel hatékonyabban kezelhetjük a poliszémiát, mint a hagyományos vagy az egynyelvű korpuszokon alapuló lexikográfia.

Hogy a fenti hipotéziseket közelebbről is megvizsgáljuk, készítettünk egy litván-magyar protoszótárat is, amelyet – a teljes kiértékelés igénye nélkül – összehasonlítottunk a már meglévő litván-magyar szótárral [2].

Abból az előfeltevésből kiindulva, hogy „*erős korreláció figyelhető meg egy szó gyakorisága és szemantikai komplexitása között*” [1], csak azokat a litván lemmákat vettük figyelembe, amelyek legalább százszor előfordultak a korpuszban. Ezzel párhuzamosan a fordítási valószínűséget jelentősen csökkentettük: 0,5-ről 0,02-re. Az így meghatározott paraméterekkel 6550 fordításjelöltet kaptunk, amelyek 1759 litván lemmához tartoztak. Az 5. táblázat jól szemlélteti, hogy a javasolt módszerrel számos különböző fordítást nyerhetünk ki a korpuszból sorba rendezve aszerint, hogy a fordítás mennyire valószínű. Jól látszik továbbá az is, hogy a nagyon gyakori szavak esetében nagyon alacsony fordítási valószínűségű párok is adhatnak jó jelölteket.

5. táblázat: litván *puikus* magyar megfelelői.

LIT	**HUN**	**P(tr)**
puikus	remek	0.071
puikus	tökéletes	0.052
puikus	szép	0.048
puikus	pompás	0.035
puikus	jól	0.035
puikus	nagyszerű	0.035
puikus	finom	0.028
puikus	gyönyörű	0.02

A poliszémia ilyen módon való kezelése különösen alkalmasnak tűnik aktív (a forrásnyelvi beszélő célnyelven való megnyilatkozását segítő) szótárak készítésének támogatására. Szintén az aktív szótárak készítését segítik elő a korpuszból kinyert kontextusok, amelyek segítséget nyújthatnak a legjobb célnyelvi fordítás kiválasztásában. Ezt támasztja alá az alábbi ábra is:

1. ábra. Bejegyzések összehasonlítása.

Míg a hagyományos szótárban egy magyar fordítás található két kontextussal[13], addig az általunk készített szótárban négy magyar fordítás[14] található 75 kontextussal.

5 Konklúziók és további teendők

A cikkben egy párhuzamos korpuszon alapuló korpuszvezérelt megközelítést ismertettünk, amelyet kétnyelvű szótárak készítésének automatikus támogatására használtunk. A javasolt automatikus módszer lexikográfiai célokra számos ok miatt hasznosnak bizonyult. Ezek közül a legfontosabb, hogy – ha egy megfelelő méretű és reprezentatív korpusz rendelkezésre áll – a javasolt megközelítés garantálja, hogy a legrelevánsabb fordítások fognak szerepelni a szótárban. Ezért a javasolt módszer jobban kezeli a poliszémiát, mint akár a hagyományos lexikográfia, akár az egynyelvű korpuszokat felhasználó lexikográfia. Ezenfelül lehetővé válik a fordításjelöltek nyelvhasználaton alapuló rangsorolása: a legvalószínűbb fordításjelöltek szerepelnek először. A megközelítés további előnye, hogy az összes releváns példa könnyedén hozzáférhető, így a poliszém jelentések közül nagy mennyiségű természetes adat alapján választhatjuk ki a legmegfelelőbbet. A fenti tulajdonságok együttese különösen alkalmassá teszi az általunk javasolt módszert aktív szótárak előállítására.

Végül, a javasolt módszerrel könnyen előállíthatjuk a fordított irányú protoszótárat, hiszen csak a szóillesztő algoritmust kell újra alkalmazni.

A módszer hátrányai közé tartozik, hogy a kevéssé használt nyelvekre a megfelelő lefedettséget biztosító korpusz létrehozása rendkívül időigényes. Egyik fő feladatunk a litván-magyar párhuzamos korpusz méretének növelése.

Egy – a szóillesztő algoritmusból fakadó – további nehézség, hogy a módszer jelenlegi formájában nem alkalmas a többszavas kifejezések kezelésére. Egy lehetsé-

[13] A Bojtár-féle szótár valójában két fordítást ad meg, ám a második ezek közül csak a példamondatból derül ki.

[14] Hat javasolt fordításból négy volt jó.

ges megoldás a fordításjelöltekhez tartozó kontextusok alapján a megfelelő fordításokat az utószerkesztési munkálatok során kézzel hozzáadni. Egy további kutatási irányt képez a többszavas kifejezések automatikus kezelése.

Hivatkozások

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press (2008)
2. Bojtár E.: Litván-magyar nagyszótár. Akadémiai Kiadó, Budapest (2007)
3. Digitális Irodalmi Akadémia: http://www.pim.hu/
4. Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R.: Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In: Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland (2005) 32-36
5. Magyar Elektronikus Könyvtár: http://mek.oszk.hu/
6. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, No. 1. (March 2003) 19-51
7. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas (2002) 710-717
8. Rimkutė, E., Daudaravičius, V., Utka, A., Kovalevskaitė, J.: Bilingual Parallel Corpora for English, Czech and Lithuanian. In: The Third Baltic Conference on Human Language Technologies 2007 Conference Proceedings. Kaunas (2008) 319–326
9. Rimkutė, E., Daudaravičius, V., A. Utka.: Morphological Annotation of the Lithuanian Corpus. In: 45th Annual Meeting of the Association for Computational Linguistics; Workshop Balto-Slavonic Natural Language Processing 2007 Conference Proceedings. Praga (2007) 94–99
10. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005. (2005)

A Szeged Treebank függőségi fa formátumban

Vincze Veronika[1], Szauter Dóra[1], Almási Attila[1], Móra György[1],
Alexin Zoltán[2], Csirik János[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{vinczev, szauter, gymora}@inf.u-szeged.hu, vizipal@gmail.com
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
alexin@inf.u-szeged.hu
[3] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
csirik@inf.u-szeged.hu

Kivonat: Az előadásban a Szeged Treebank függőségi fa formátumra történő átalakításának folyamatát mutatjuk be. Az eredetileg frázisstrukturált treebankből automatikus konverzió eredményeképpen létrejött függőségi fákat kézi úton ellenőriztük és javítottuk, létrehozva ezzel az első magyar nyelvű kézzel annotált dependenciakorpuszt. Jelenleg az üzleti híreket, újsághíreket és jogi szövegeket tartalmazó alkorpuszok annotációja fejeződött be, de terveink között szerepel a teljes korpusz átalakítása függőségi fa formátumra. Az elkészült adatbázis hasznosítható többek között az információkinyerésben és a gépi fordításban is.

1 Bevezetés

A Szeged Treebank függőségi fákat tartalmazó szintaktikai annotációjának célja az első, teljes egészében kézzel annotált magyar nyelvű dependenciakorpusz létrehozása. Az adatbázis számítógépes szempontú hasznosíthatósága többrétű, hiszen a gépi fordításban való felhasználás mellett az információkinyerés részterületein is számos alkalmazásban töltheti be a tanító adatbázis szerepét. Az előadásban ismertetjük a korpuszépítési munkafolyamatokat, a konverzió és az annotáció során felmerült problémákat és az azokra született megoldásokat, továbbá a korpusz statisztikai adatait, végül szót ejtünk a korpusz hasznosíthatóságáról is, illetve nemzetközi kontextusban is elhelyezzük a létrehozott adatbázist.

2 Függőségi nyelvtanok

A Szeged Treebank eredetileg frázisstrukturált formában kódolja a mondat összetevői közti szintaktikai viszonyokat. A frázisstrukturált korpuszban a mondatok tagmondatokból felépülő hierarchikus struktúrát alkotnak: a mondat összetevői (konstituensei) konstituensfákká szerveződnek. Maguk a tagmondatok igékre, az igék vonzataira (ezek névszói szerkezetek) és egyéb alkotóelemekre bonthatók, amelyek az egyes szinteken belül azonban nem alkotnak hierarchiát. A mondat szavai a

konstituensfa levelein helyezkednek el, a fa egyéb csomópontjai absztrakt szerveződési egységeket jeleznek (frázisstruktúra-címkékkel ellátva).

A függőségi fa formátum ettől abban tér el, hogy a fában minden egyes csomópont a mondat egy szavának felel meg. A mondatfa csúcsán egy mesterséges gyökérelem található, amelynek alárendeltjei lesznek a mondatban előforduló szavak, vagyis a gyökérelemen kívül nem találhatók absztrakt csomópontok a fában. A mondatban minden egyes szó szigorúan egy másik szó alárendeltségében van: egy szónak csak egy fölérendeltje lehet, egy csomópont alá azonban tartozhat több szó is, például az ige csomópontja alá sorolható be az ige összes bővítménye. A függőségi fában szereplő csomópontok között többféle kapcsolat is lehetséges, ezeket általában különféle címkékkel látják el, amelyek a kapcsolat jellegére utalnak.

Az első függőségi nyelvtannak Tesnière könyve [20] tekinthető, mely lefekteti a rendszer alapjait. Híres hasonlata szerint a mondatnak az ige a központi eleme, mely egész kis drámát fejez ki: a dráma szereplői az ige bővítményei, melyeket Tesnière aktánsoknak nevez. A mondatban így tehát alárendelt elemek és fölérendelt elemek szerveződnek egységbe.

Mel'čuk [17, 18] függőségi nyelvtana az Értelem ↔ Szöveg Elméleten belül jött létre. Nála a függőségi viszony lineáris relációként jelenik meg a szavak között. Mélyszintaktikai szinten 12 viszonytípust feltételez, ebből 6 az ige és különféle bővítményei (aktánsai) között létezik, a többi viszony pedig mellérendelést és különféle módosító szerepet jelez. A Mel'čuk-féle függőségi nyelvtan különlegessége, hogy a mellérendelést is egyfajta alárendelésként fogja fel: a mellérendelés első tagjához kapcsolódik a kötőszó, illetve utóbbihoz a mellérendelés további tagja(i) speciális (COORD) viszonnyal. Egy másik érdekesség, hogy bizonyos esetekben a nyelvtan engedélyezi absztrakt, azaz a mondatban fonetikailag meg nem jelenő nyelvi elemet jelző csomópontok felvételét a függőségi fába: ilyen eset például az egyes szám harmadik személyű jelen idejű létige az oroszban (és a magyarban is), amely fonetikailag nem ölt testet a mondatban, azonban absztrakt szinten mégis jelen van, hiszen múlt és jövő időkben megjelenik testes formában.

A magyar nyelvre alkalmazott függőségi nyelvtanokról [16] és [19] nyújt áttekintést, illetve saját, morféma alapú függőségi nyelvtanuk rövid vázlatát mutatják be a szerzők. Modelljükben a függőségi fák alapelemei a morfémák, mivel agglutináló nyelvekben nem (csak) a szavak, hanem a morfémák képesek a különböző grammatikai viszonyok kifejezésére. Ez a megoldás megkönnyíti a különböző típusú nyelvek függőségi fái közti leképezéseket, mert például az angol *may* segédige csomópontjának a magyar fában a *–hat* morféma csomópontja felel meg. Ezt az eljárást alkalmazva a függőségi fákon alapuló számítógépes fordítórendszerek hatékonysága jelentősen megnövekedhet.

3 Más nyelvű dependenciakorpuszok

A világ számos nyelvére fejlesztettek már ki dependenciakorpuszt. Ezek közül az egyik leghíresebb a cseh nyelvre épített Prague Dependency Treebank [1], mely morfológiai, szintaktikai és tektogrammatikus szintű annotációt is tartalmaz. Ugyanez a műhely angolra és csehre is kifejlesztett egy párhuzamos, dependenciaannotációt

tartalmazó korpuszt [2, 3], illetve arab nyelvű dependenciakorpusz is fűződik a nevükhöz [4]. A fentieken kívül számos európai (többek között svéd [5], görög [6], orosz [7] és szlovén [8]) és Európán kívüli nyelvre (japán [9], kínai [10]) építettek már dependencia treebanket, illetve még holt nyelvekre is: egy latin nyelvű korpusz már létrejött, és alkotói egy ógörög korpuszon dolgoznak jelenleg [11]. Az első magyar nyelvű dependenciakorpusz létrehozásával ehhez a vonulathoz kívánunk csatlakozni.

4 A korpuszépítés folyamata

Ahhoz, hogy az eredetileg frázisstrukturált treebankből dependenciakorpuszt tudjunk készíteni, először is szükség van egy konverziós lépésre, melynek során a konstituensfák függőségi viszonyokká alakulnak át. Mivel az automatikus gépi konverziótól nem várhatunk tökéletes és hiba nélküli eredményt, ezt a munkafázist egy kézi ellenőrzési folyamat követi, melynek során nyelvészek átnézik a fájlokat, és a szükséges esetekben módosítják azokat.

Noha a korábbi szakirodalomban megtalálható a magyarra alkalmazott függőségi nyelvtan rövid vázlata [16, 19], a Szeged Treebank függőségi fa formátumra történő átalakításakor mégsem követjük teljes egészében ezt a modellt. Ennek az a magyarázata, hogy az említett modell morféma alapú, azaz a függőségi fa csomópontjaiban nem szóalakok, hanem morfémák szerepelnek. Ahhoz azonban, hogy a szintaktikai fákat morfémákból építhessük fel, szükség lenne egy jól működő morfológiai elemzőre, mely a Szeged Treebank szóalakjait morfémákra bontaná. Mivel a Szeged Treebank MSD-kódjai a képzéseket nem jelölik, például a műveltető és ható igék képzőit a szótő részeként kezeli a rendszer, vagyis nem lenne képes külön morfémát, azaz külön csomópontot rendelni a képzőkhöz. A morféma alapú függőségi fákra történő konverzió választása még további munkaigényes feladatokkal járna (többek között az MSD-kódrendszer átalakítása úgy, hogy lehessen jelezni a képzéseket, a szóalakok újrakódolása a korpuszon belül, jól működő morfológiai elemző kialakítása a korpuszra stb.). Emiatt csupán a szóalakok közti függőségi viszonyok bejelölésére vállalkoztunk.

A Szeged Treebank 2.0 függőségi fa formátumra való átalakítása első lépésének a 2007-es CoNLL konferencia szervezőbizottsága által kiírt nemzetközi versenyfeladat [12] tekinthető, amikor is a tesztadatbázis elkészítésére való felkérésnek köszönhetően megtörtént a korpusz HVG- és Népszabadság-cikkeket tartalmazó részének konvertálása [13], majd ennek nyomán a teljes korpusz átalakítása.

A Szeged Treebank 2.0-ban az ige és vonzatai közti nyelvtani viszonyok jelölve voltak. Ezeket a viszonyokat kellett függőségi viszonyokká átalakítani. A konverzió során automatikusan, gépi úton történt a viszonyok átcímkézése nyelvészek által előzetesen meghatározott szabályok alapján. A lehetséges függőségi viszonyok az alábbiak:

APPEND – a mondatba szervesen nem illeszkedő mondatrészek
ATT – főnév és jelző, névutó és főnév, főnév(i módosító) és főnév közti viszony
AUX – ige és segédige közti viszony
AUXS – a mondat értékű elem
CONJ – kötőszó
COORD – mellérendelés
DAT – nAk ragos főnévi vonzat
DET – főnév és determináns közti viszony
FROM – honnan? kérdésre válaszoló határozószó, illetve névutós szerkezet
INF – főnévi igenév
LOCY – hol? kérdésre válaszoló határozószó, illetve névutós szerkezet
MODE – egyéb határozószavak, illetve névutós szerkezetek
NEG – tagadószó
OBJ – ige és tárgy közti viszony
OBL – ige és egyéb főnévi bővítménye közti viszony
PRED – ige és névszói állítmány közti viszony
PREVERB – ige és igekötő közti viszony
PUNCT – írásjel
QUE – kérdőszó
ROOT – a mondat fő eleme
SUBJ – ige és alany közti viszony
TFROM – mikortól? kérdésre válaszoló határozószó, illetve névutós szerkezet
TLOCY – mikor? kérdésre válaszoló határozószó, illetve névutós szerkezet
TO – hova? kérdésre válaszoló határozószó, illetve névutós szerkezet
TTO – meddig?, mikorra? kérdésre válaszoló határozószó, illetve névutós szerkezet

A gépi úton előállt fájlokat nyelvészek ellenőrizték, és ha kellett, javították. A javítási munkálatokhoz az erre a célra kifejlesztett, és a magyar nyelv sajátosságainak megfelelően testre szabott TrEd szerkesztőprogramot [14] használtuk.

4.1 Típushibák

A kézi ellenőrzés során elsődleges feladat a gépi konverzió átnézése, szükség esetén javítása volt. A javításra szoruló legtipikusabb hibák két kategóriába estek: (1) a csomópont rossz helyen volt a fában; (2) a csomópont és fölérendeltje nem a megfelelő viszonyban állt.

A hibák nagy része abból fakadt, hogy a frázisstrukturált korpuszban nem minden nyelvtani viszony volt jelölve, például a névelők, számnevek és jelzők a főnévi csoporton belül szerepeltek, és a főnévhez fűződő viszonyuk külön nem volt feltüntetve. A konverzió során automatikusan a főnév alá lettek bekötve ATT viszonnyal mindezen elemek, a mondatban található egyéb elemek pedig az ige alá kerültek be MODE viszonnyal. Ezeket szükség szerint javítani kellett a megfelelő függőségi viszonyra, illetve áthelyezni a megfelelő felettes (anya)csomópont alá.

Az átcímkézést igénylő leggyakoribb esetek a következők voltak:

- **jelzős szerkezeten belüli ragozott főnév**
  A konvertálóprogram a fenti okokból kifolyólag ATT címkével látott el minden főnevet, amely AP (melléknévi csoport) tagja volt, például *a ténylegesnél 1,9_milliárd dollárral magasabb árbevételt* szerkezetben a *ténylegesnél* és a *dollárral* is ATT címkét kapott a helyes OBL helyett, így ezt kézi úton kellett javítani.

- **NE-k kezelése**
  A tulajdonnevek az esetek nagy többségében ATT címkét kaptak a konverzió során, ezeket természetesen javítottuk az adott kontextusnak megfelelő címkére.

- **alárendelő mellékmondatok fő elemének címkéje**
  Az alárendelő mellékmondatokat a Treebankben annak megfelelően címkézték, hogy milyen szerepet tölt be a főmondatban az utalószó (és az utalószó alá is kötötték be, amennyiben volt ilyen a mondatban, l. alább). A dependenciakorpuszban ettől eltérően csak annyit jelölünk, hogy alárendelésről van szó, azaz ATT címkével látjuk el a mellékmondat fő elemét.

- **mellérendelések második, harmadik… tagja**
  A Treebankben a mellérendelések a frázisstruktúra-nyelvtanokban szokásos megoldásnak megfelelően kívülről kaptak egy közös címkét, melynek típusa megegyezett a mellérendelés tagjainak saját címkéjével: tehát két egymás mellé rendelt főnévi csoport (NP) egy külső NP címkével is rendelkezett, mely mindkettőt magában foglalta. Mivel a dependencia-nyelvtanokban nincsenek mesterséges csomópontok, ez az eljárás nem bizonyult követhetőnek, így a Mel'čuk-féle megoldást követtük a mellérendelések elemzésénél, l. lejjebb.

- ***ez/az* mutató névmások**
  A mutató névmások ATT címkét kaptak, ha mutató névmás + névelő + főnév konstrukcióban (*ez a ház*) fordultak elő. Alanyesetű előfordulásukkor DET, azaz determinánsi címke járt nekik, ha pedig esetragot viseltek (pl. *ebben a házban*), akkor az adott esetnek megfelelő címkére kellett javítani (jelen példában OBL-ra).

A csomópontok áthelyezése a fában az alábbi esetekben volt a legszükségesebb:

- ***alárendelő mellékmondatok***
  Amint már fentebb utaltunk rá, a kötőszó nem képezte az alárendelő mellékmondatok részét a Szeged Treebank frázisstrukturált változatában. Ennek eredményeképpen a konverzió után a főmondat fő eleméhez kapcsolódott a kötőszó és a mellékmondat fő eleme is (külön-külön). A kézi ellenőrzés folyamán a nyelvészek a kötőszóhoz kötötték hozzá a mellékmondat fő elemét, így teremtve meg a kapcsolatot a két összetevő között.

- ***birtokos szerkezetek***

  A birtokos szerkezetek két része, a birtokos és a birtok gyakran nem kapcsolódott össze a korpuszban. Különösen érvényes volt ez a *–nAk* ragos birtokosra, főleg, ha nem a birtok melletti pozíciót foglalta el a mondatban. A dependenciakorpuszban a birtokost mindig összekötöttük a birtokkal, még akkor is, ha ezzel keresztező függőségek álltak elő, azaz a fa két éle metszi egymást. (Ez a frázisstruktúra-nyelvtanokban szigorúan tilos, mivel ott lehetségesek a mozgatások, dependencia-nyelvtanokban azonban elfogadott a keresztezések léte.)

- ***mellérendelés***

  Amint már az átcímkézési eseteknél említettük, mellérendelésnél nemcsak a csomópontok címkéit, hanem a helyzetüket is módosítani kellett. A gépi elemzés során általában a kötőszó funkcionált a szerkezet fejeként, és a mellérendelés tagjai vele álltak függőségi viszonyban. A Mel'čuk-féle megoldásnak megfelelően azonban a szerkezet első tagja funkcionál fejként, ez alá kell kötni a kötőszót (amennyiben volt) CONJ viszonnyal, majd a mellérendelés többi tagja következik COORD viszonnyal kapcsolva az előző elemhez.

- ***főnévi igenevek és igekötők***

  Ha a mondatokban szerepelt egy olyan (segéd)ige, amelynek főnévi igenév vonzata volt (*szeret*, *kíván*, *fog*, *kell*…), akkor a gépi elemzés a főnévi igenév esetleges igekötőjét a főigéhez társította. Ezt a hibatípust is kézzel javították a nyelvészek az ellenőrzés során.

4.2 Mellérendelés

A mellérendelés kérdése problémákat vet fel a legtöbb szintaktikai elmélet számára: egyes elméletek hívei azt a megoldást tartják jónak, hogy a kötőszó a koordináció feje, mások pedig amellett érvelnek, hogy a szerkezet feje a mellérendelés egyik tagja. Vizsgáljuk meg ezeket az elképzeléseket külön-külön!

Tegyük fel, hogy a kötőszó a szerkezet feje. Felmerül azonban a kérdés, hogy mit lehet tenni a direkt koordináció eseteiben, amikor nincs az elemek között kötőszó. Ha nincs kötőszó, akkor fel kell tételezni egy virtuális csomópontot, amely képes fejként funkcionálni. Az elképzelésnek azonban más hátulütője is van: ha több mellérendelt elem van, akkor nem tudjuk megkülönböztetni az „A és B és C" típust az „A, B és C" típustól. A problémát meg lehetne úgy kerülni, hogy felveszünk egy absztrakt „és"-t az „A" és „B" fölé, de akkor a „B" egyidejűleg két csomóponthoz (egy virtuális *ÉS* és egy valós *és*) kapcsolódna, ez pedig szigorúan tilos. További hátránya az elgondolásnak, hogy ha például a mellérendelt frázis a mondat alanya, akkor a kötőszó és az ige közt lenne SUBJ viszony, ez pedig igen kevéssé lenne szokványos.

Egy másik elképzelés szerint azonos szinten szerepelnek a koordinált elemek és a kötőszó, de nincsenek összekapcsolva, például a *Jancsi és Juliska mézeskalácsháza* szókapcsolatban a *mézeskalácsháza – Jancsi*, *mézeskalácsháza – és*, valamint *mézeskalácsháza – Juliska* viszonyok állnak fönn. Ez esetben az jelenti a problémát, hogy noha *Jancsi* és *Juliska* összetartozását az azonos címkéjű (ATT) viszony még vala-

hogy tudná jelölni, de eléggé kérdéses, hogy milyen viszonyban állna a *mézeskalács-háza* és az *és*, arról nem is beszélve, hogy eléggé szokatlan, hogy a koordináció két tagját nem kapcsoljuk össze.

A fenti megoldások egyike sem nyújt kielégítő választ a felmerülő problémákra, éppen ezért a korpusz átalakítása során a koordináció esetén a Mel'čuk-féle elképzelést [17, 18] követjük, ahol is a mellérendelés egyfajta „alárendelés". Mindig a koordináció első eleme a fej, mert az tud az egész frázis helyett állni. Vegyük a következő példákat:

Elmentem a boltba Józsival és Kival.
Elmentem a boltba Józsival.
**Elmentem a boltba Józsival és.*
**Elmentem a boltba és Kival.*

A második, illetve a harmadik és negyedik mondat közti különbség mutatja, hogy a koordináció nem bontható fel két egyenrangú részre, hiszen ha a *Józsival* és az *és Kival* elemek egyenértékűek lennének, akkor elfogadhatónak kellene lennie az utolsó mondatnak. A *Józsival* az *és* elemmel sem tartozik szorosan össze, hiszen akkor a harmadik mondat is jó lenne. A megoldás az, hogy három részt feltételezünk a koordinációban: az első elem a fej, ehhez kapcsolódik a kötőszó CONJ viszonnyal, illetve a kötőszót követi a második mellérendelt tag COORD viszonnyal:

Józsival
| CONJ
és
| COORD
Kival

Ez ábrázolás szempontjából igaziból „alárendelés", és így szerkezetben nem lesz különbség mellé- és alárendelés között: csak a viszonyok (ATT, illetve COORD) jelzik, hogy melyikről van szó.

4.3 Predikatív névszók

A magyar nyelv sajátságaiból adódóan a predikatív névszót tartalmazó mondatokban a létige kijelentő mód jelen idő E/3. alakja nem jelenik meg a felszínen, szemben a más módú, idejű vagy számú, illetve személyű formákkal:

*András katona (*van).*
András legyen katona!
András katona lesz.

A mellérendeléshez hasonlóan, jelen problémánál is kétféle megoldási lehetőség létezik. Az első lehetőség szerint a mondat fő elemének a predikatív névszót tekintjük, ez alá csatoljuk az alanyt, és nem veszünk fel virtuális csomópontot. Azonban ennek a megoldásnak az a hátránya, hogy teljesen más szerkezetet tulajdonítunk

ugyanannak a mondatnak jelen és például múlt időben, ami megkérdőjelezhető, mert az egyik esetben a predikatív elem és az alany között közvetlen, másik esetben pedig közvetett kapcsolat van:

A másik megoldás fenntartja az azonos szerkezetet a mondat bármely előfordulása esetén, igaz, ennek az az ára, hogy fel kell tételeznünk egy virtuális csomópontot a létige kijelentő mód jelen idő E/3. alakja számára (VAN). Így a következőképpen alakulnak a függőségi fák:

További érv a virtuális csomópont alkalmazása mellett, hogy szintaktikai szinten mindenképpen jelen van a VAN, hiszen a többi igealak/igeidő/igemód esetében testes morfémaként jelenik meg. Az már másodlagos (morfológiai) kérdés, hogy jelen idő E/3-ban miért zéró morféma az alakja (vö. [18]). Előnyt jelenthet a virtuális csomópont alkalmazása a korpusz nemzetközi felhasználhatóságában is, hiszen például egy függőségi fákra épülő fordítóprogram jóval hatékonyabb működésre képes, ha azonos struktúrájú fát kell leképeznie a másik nyelvre, szemben azzal, ha még ráadásul külön transzformációs lépéseket is be kell iktatnia a fordítás folyamatába.

5 Statisztika

A Szeged Treebank 2.0 állománya 82.000 mondatot, 1,2 millió szövegszót és 250 ezer írásjelet tartalmaz. A szövegek hat különböző témakörből kerültek ki, témakörönként ~200 ezer szó terjedelemben. A témakörök a következők:

- Szépirodalom
- 14-16 éves korú tanulók fogalmazásai
- Újságcikkek (Népszabadság, Népszava, Magyar Hírlap, HVG)
- Számítástechnikai szövegek
- Jogi szövegek
- Gazdasági és pénzügyi rövidhírek

2009 novemberéig a gazdasági és pénzügyi rövidhíreket tartalmazó alkorpusz, az újsághírek és a jogi szövegek dependenciaelemzése készült el teljes egészében, illetve a számítógépes témájú szövegek elemzése zajlik jelenleg. Az eddig elkészült korpusz statisztikai adatai a következő táblázatban foglalhatók össze:

1. táblázat: A korpusz statisztikai adatai.

	newsml	**újsághírek**	**jogi szövegek**	**összesen**
Mondatok	9574	10210	9278	29062
Szavak	186030	182172	220069	588271
Írásjelek	25712	32880	33515	92107

Az annotációs munkálatok várhatóan 2010 elején fejeződnek be.

6 A korpusz hasznosíthatósága

A számítógépes nyelvészet több területén is haszonnal bírhat a függőségi fák alkalmazása: mind a gépi fordításban, mind az információkinyerésben sikeresen felhasználhatók a függőségi fa formátumú korpuszok.

6.1 Gépi fordítás

A szintaktikai transzformáción alapuló gépi fordítási eljárások alapvetően két forrásra építenek: vagy a forrásnyelvi konstituensfákat képezik le a célnyelvi konstituensfára, vagy pedig függőségi fákkal dolgoznak. A konstituensfákat alkalmazó módszer előnyei közé tartozik, hogy rokon nyelvek gépi fordítására jól alkalmazható, hiszen a rokon nyelveknek többnyire hasonló a szintaxisa, továbbá az eltérő szórendből adódó problémákat is elfogadható mértékben oldja meg. A módszer hátránya viszont, hogy rendkívül bonyolult és költséges transzformációs szabályokat kell bevezetni a rendszerbe, ráadásul ha a mondatnak teljesen eltérő szintaktikai szerkezete van a forrás-, illetve a célnyelvben, a fordítás teljesen elfogadhatatlanná válik.

Gyakori hiba továbbá a konstituensfákat használó fordítórendszerekben, hogy az elemző gyakran hibás szerkezetet tulajdonít a fának, felesleges címkéket szúr be vagy rossz csomópontokat feleltet meg egymásnak. A mesterséges csomópontokból adódó hibák kiküszöbölését sikeresen oldják meg a függőségi fákra alapuló fordítórendszerek, hiszen a függőségi fákban nincsenek absztrakt (mesterséges) csomópontok. A fa minden csomópontja így egy természetes nyelvi elemnek feleltethető meg a mondatban, a fa nem tartalmaz szintaktikai csomópontokat, a nyelvek közti szintaktikai különbségek így eltűnnek. A gépi fordítási eljárás során minden csomópont lefordítódik, és ha szükséges, akkor a csomópontok újrarendeződnek bizonyos előre megadott valószínűségek mentén. A függőségi fákat alkalmazó gépi fordítási eljárás különösen a nem rokon vagy eltérő szintaxisú nyelvpárok esetén lehet gyümölcsöző.

6.2 Információkinyerés

A számítógépes nyelvészet egy más területén, az információkinyerésben is hasznosíthatók a függőségi fák. A szintaktikailag annotált korpuszok igen fontos szereppel bírnak az automatikus információkinyerés területén, ugyanis nem elégséges csak azt tudni, hogy milyen szavak, illetve kifejezések szerepelnek az adott szövegben, annak is nagy jelentősége van, hogy ezek egymással milyen viszonyban állnak. Például gazdasági jellegű szövegekben a különböző tranzakciókról szóló információk között szerepelnie kell annak is, hogy ha A és B cég vett részt egy adásvételi folyamatban, akkor melyik cég vásárolta fel a másikat (azaz melyik a *felvásárol* ige alanya és tárgya). Ahhoz azonban, hogy ezt nagy biztonsággal meg lehessen állapítani, szintaktikai viszonyokat is tudni kell elemeznie az információkinyerő rendszernek. A szintaktikai viszonyok tanításában rendkívüli szereppel bírnak a szintaktikailag annotált korpuszok.

A kötött szórenddel rendelkező nyelvek esetén jó alternatíva lehet a konstituensfákat használó, szintaktikailag annotált korpusz: ezekben ugyanis adott szintaktikai szerkezethez adott szintaktikai viszony társul. A függőségi nyelvtanokra épülő korpuszok azonban inkább a szabad szórendű nyelvek esetén nyújtanak nagy segítséget az információkinyerésben, hiszen esetükben a szintaktikai viszonyokat illetően nem lehet a szórendet segítségül hívni: a függőségi nyelvtanok lényege, hogy a szórendtől függetlenül képes meghatározni a mondat szintaktikai szerkezetét.

Jelen korpuszban jelölve vannak az ige és bővítményei közti alapvető viszonyok, azaz a bővítmények közül az alany, tárgy és részeshatározó szerepű argumentumok könnyen azonosíthatók (SUBJ, OBJ és DAT címkével vannak ellátva), a további bővítmények pedig OBL címkével rendelkeznek. Így az információkinyerő program is sikeresen meg tudja állapítani a következő példában rejlő szintaktikai viszonyokat:

Az E.ON_Hungária_Energetikai_Rt. 87,713 százalékra növelte részesedését a Tiszántúli_Áramszolgáltató_Rt-ben.

A kinyerhető releváns szintaktikai viszonyok a következők:

növelte - Az E.ON_Hungária_Energetikai_Rt. (alany)
növelte – részesedését (tárgy)
növelte – a Tiszántúli_Áramszolgáltató_Rt-ben (bővítmény)

A szintaktikai viszonyokból a számítógép számára is kiderül, hogy a mondatban szereplő két Named Entity viszonya milyen, azaz az E.ON rendelkezik tulajdonrészszel a Titászban, és nem fordítva, ezáltal a szintaktikai viszonyokat is felhasználó információkinyerés pontossága igencsak megjavul az azokat nem hasznosító modellekhez képest.

6.3 Többnyelvűség

A magyar nyelvű dependeciakorpusz létrehozásával lehetőség nyílik a többnyelvűséget szem előtt tartó alkalmazások fejlesztésére is. A Szeged Treebank alkorpuszai

közül a kapcsolódási pontot a többnyelvű (párhuzamos) korpuszokhoz az *1984* és a *Windows2000* szövegállományok jelenthetik, hiszen ezeknek a szövegeknek bizonyosan létezik idegen nyelvű megfelelője is. Amennyiben az idegen nyelvű verziók tartalmaznak függőségi viszonyokra alapuló szintaktikai annotációt, könnyen létre lehet hozni egy magyar-adott nyelvű párhuzamos dependenciakorpuszt. Ez nagyban elősegítené egyrészt a többnyelvű információkinyerést támogató rendszerek fejlesztését, másrészt pedig a függőségi fákon alapuló, szintaktikai módszerekre építő gépi fordítóprogramok létrehozását. A korpusz létrehozása tehát mind elméleti, mind gyakorlati szempontok alapján jelentőségteljesnek és haszonnal kecsegtetőnek nevezhető.

7 Összegzés

A tanulmányban a Szeged Treebank függőségi fa formátumra történő átalakításának folyamatát mutattuk be: ismertettük a munkafolyamatokat, a felmerült problémákat és az azokra nyújtott megoldásokat. Szót ejtettünk a korpusz gépi fordításban, illetve információkinyerésben való hasznosíthatóságáról, továbbá a kontrasztív nyelvészet és a dependeciaszintaxis kutatói is számára haszonnal bírhat az adatbázis. A későbbiekben szeretnénk továbbá kifejleszteni egy magyar nyelvű dependenciaparsert is (vagy egy már rendelkezésre álló korábbi (például a MaltParser [15]) testreszabásával, vagy pedig önálló kutatás-fejlesztés eredményeként), melyhez az elkészült korpusz tanító adatbázisként szolgálhat.

Köszönetnyilvánítás

A kutatást – részben – a TUDORKA és a MASZEKER projekt (Jedlik Ányos programok) keretében az NKTH támogatta.

Hivatkozások

1. Hajič, J., Böhmová, A., Hajičová, E., Vidová Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: A. Abeillé (ed.): Treebanks: Building and Using Parsed Corpora, Amsterdam:Kluwer (2000) 103-127
2. Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kuboň, V.: Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In: 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
3. Čmejrek, M., Cuřín, J., Havelka, J.: Prague Czech-English Dependecy Treebank: Any Hopes for a Common Annotation Scheme? In: HLT/NAACL 2004 Workshop: Frontiers in Corpus Annotation, Boston, Massachusetts (2004) 47-54
4. Hajič, J., Smrž, O., Zemánek, P:, Šnaidauf, J., Beška, E.: Prague Arabic Dependency Treebank: Development in Data and Tools. In: Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools. Cairo, Egypt, September 2004. (2004) 110-117

5. Nivre, J.: Theory-Supporting Treebanks. In: Nivre, J. and Hinrichs, E. (eds.) Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö University Press (2003) 117-128
6. Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., Piperidis, S.: Theoretical and practical issues in the Construction of a Greek Dependency Corpus. In: Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-2005), Barcelona (2005)
7. Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., Frid, N.: Dependency Treebank for Russian: Concept, Tools, Types of Information. In: Proceedings of the 18th conference on Computational linguistics. Saarbrücken, Germany (2000) 987–991
8. Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., Žele, A.: Towards a Slovene Dependency Treebank. In: Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06, 24-26 May 2006. Genoa, Italy (2006)
9. Lepage, Y., Shin-Ichi, A., Susumu, A., Hitoshi, I.: An annotated corpus in Japanese using Tesnière's structural syntax. In: Proceedings of COLING-ACL'98 Workshop on the Processing of Dependency-based Grammars, Montreal (1998)
10. Liu, H.: Building and Using a Chinese Dependency Treebank. GrKG/Humankybernetik No. 48 Vol. 1 (2007) 3–14
11. Bamman, D., Crane, G.: The Design and Use of a Latin Dependency Treebank. In: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006) (Prague) (2006) 67-78
12. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, Prague (2007) 915–932
13. Alexin, Z.: A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra. In: Tanács, A., Csendes, D. (szerk.): V. Magyar Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). Szegedi Tudományegyetem, Szeged (2007) 263-266
14. http://ufal.mff.cuni.cz/~pajas/tred/
15. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, No. 13, Vol. 2. (2007) 95-135.
16. Koutny I., Wacha B.: Magyar nyelvtan függőségi alapon. Magyar Nyelv Vol. 87 No. 4. (1991) 393–404.
17. Mel'čuk, I. A.: Dependency Syntax: theory and practice. State University of New York Press, Albany, NY (1988)
18. Mel'čuk, I. A.: Levels of Dependency in Linguistic Description: Concepts and Problems. In Agel, V., Eichinnger, L., Eroms, H.-W., Hellwig, P., Herringer, H. J., Lobin, H. (eds.): Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, Berlin-New York, W. de Gruyter (2003) 188-229
19. Prószéky, G., Koutny, I., Wacha, B.: Dependency Syntax of Hungarian. In: Maxwell, Dan; Klaus Schubert (eds.) Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation), Foris, Dordrecht, The Netherlands (1989) 151–181
20. Tesnière, L.: Éléments de syntaxe structurale. Paris, Klincksieck (1959)

Fokozó értelmű szókapcsolatok detektálása magyar szövegkorpuszokban

Kiss Márton

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
mkiss@inf.u-szeged.hu

Kivonat: A cikkben ismertetem azt az általam kidolgozott módszert, amely egyrészről alkalmas fokozó értelmű szókapcsolatok (mint pl.: *borzasztóan fázik*) relevanciájának vizsgálatára, másrészről pedig ilyen szókapcsolatokat automatikusan felismer, detektál. A létrehozott algoritmusok segítségével megvizsgáltam, hogy egy nyomtatott formában már megjelent fokozó szótár [1] szókapcsolatai közül a korpuszbeli előfordulások alapján melyek relevánsak és melyeket célszerű a szótárból törlésre javasolni. Ugyanakkor az eljárás eredményeképpen javaslat születik, hogy a szótárat mely korpuszbeli fokozó szókapcsolatokkal célszerű kiegészíteni. A kidolgozott eljárás statisztikai módszerek és gyakorisági mutatók segítségével képes "ellenőrizni" egy szókapcsolat relevanciáját, illetve adott kulcsszóhoz egy korpuszból képes kigyűjteni a releváns fokozó szókat. Az eredmények a továbbiakban az olyan lexikográfusok számára nyújtanak segítséget, akik fokozó értelmű szókapcsolatokat vizsgálnak, illetve fokozó szótárat írnak, tágabb értelembe véve pedig jól szemléltetik, hogy korpusznyelvészeti eszközökkel hatékony segítség nyújtható a szótáríróknak

1 Bevezetés

Cikkemben a kollokációk egy speciális részhalmazát, a fokozó értelmű szókapcsolatokat vizsgáltam. Kiindulásként egy nyomtatott szótár, a Fokozó szótárban meghatározott szókapcsolatokat vizsgáltam, majd a szótárban felsorolt kulcsszavakból és fokozó szavakból indultam ki és kerestem új szókapcsolatokat.

Először arra kerestem a választ, hogy a Fokozó értelmű szókapcsolatok szótárában [1] található szókapcsolatok, hogyan fordulnak elő korpuszokban, illetve internetes keresőkben. A vizsgált korpuszok az MNSZ és a Webcorpus, az internetes keresők pedig a Google, a Live Search, valamint a Yahoo. A jobb szemléltetés és a tesztelés elősegítése érdekében létrehoztam a Fokozó szótár [1] egy kicsinyített változatát, a mini szótárat, mely 7 szócikket tartalmaz. Vizsgálódásaimat mindig először e mini szótáron végeztem, majd, ha az eredmények biztatóak voltak, elvégeztem az adott kísérletet az egész művön vagy korpuszon. E lépés eredményeképpen a Fokozó szótárban lévő címszavak alá rendelt szókapcsolatok egy részét törlésre javasoltam, törölhetőnek ítéltem.

Második lépésben automatikusan kerestem fokozó értelmű szókapcsolatokat az MNSZ-ben és a Hungarian Webcorpusban. Az így talált új fokozó szókapcsolatokkal kiegészítettem a szótárat, valamint gyakorisági számok bevezetésével csoportokba osztottam a szókapcsolatokat, előfordulási gyakoriságuk alapján. Így a fokozó szótár egy kiegészített, gyakorisági számokkal ellátott változata jött létre.

A kidolgozott algoritmust az *iszik* címszó élettörténetén mutatom be. Itt szemléltetem, hogy az *iszik* címszó, hogyan módosult, változott a munka előrehaladtával.

2 A fokozó szótár vizsgálata

Első lépésként, kiindulásként felhasználtam Székely Gábor szótárát [1]. A szótárban felsorolt körülbelül 700, kulcsszó szerint rendezett címszavakból választott 7 címszó (*alázatos* mn, *boldogság* fn, *iszik* i, *küzd* i, *lelkesedés* fn, *logika* fn, *vizsga* fn) (mini szótár), összesen 270 kapcsolódását vizsgáltam a következő korpuszokban:

- Magyar Nemzeti Szövegtár, 187,6 millió szó, (http://corpus.nytud.hu/mnsz/)
- Hungarian Webcorpus, 1,5 milliárd szó (http://mokk.bme.hu/resources/)
- Google (www.google.com)
- Live Search (http://search.msn.com)
- Yahoo (http://search.yahoo.com)

Azt a közel 270 szókapcsolatot, amelyekben a kiválasztott 7 kulcsszó szerepel egy erre alkalmas, saját fejlesztésű programmal kerestem meg a felsorolt internetes keresőkben, valamint korpuszokban.

Listát készítettem a korpuszokban előforduló fokozó szókapcsolatokból, gyakorisági rendbe állítottam őket. Az eredményt összevetettem a szótár anyagával, és megjelöltem azokat a szótározott fokozó szókapcsolatokat, melyek egy adott gyakorisági szintnél kevesebbszer fordulnak elő a korpuszokban vagy egyáltalán nem fordultak elő. Ezzel a lépéssel kiszűrtem a nyomtatott szótárban fellelhető fokozó szókapcsolatok közül azokat, melyeket nem ítéltem létező szókapcsolatnak. Ezek a szókapcsolatok ugyan használható, nyelvileg jónak tűnő szókapcsolatok voltak, de a kutatásaink azt mutatták, hogy nem használjuk őket.

2.1 Címszavak vizsgálata az MNSZ-ben

1. táblázat: A Fokozó szótár *iszik* címszavában található szókapcsolatok előfordulása az MNSZ-ben; összesen 56 szókapcsolat szerepel a szótárban, de csak a táblázatban szereplő 11 szerepelt az MNSZ-ben.

szókapcsolat	db	szókapcsolat	db	szókapcsolat	db
iszik:	1870	**eléggé iszik:**	**1**	**módfelett iszik:**	**1**
mohón iszik:	**22**	**erősen iszik:**	**4**	**nagyon iszik:**	**26**
nagyon iszik:	**26**	**keményen iszik:**	**11**	**rettenetesen iszik:**	**1**
nagyot iszik:	**14**	**komolyan iszik:**	**1**	**rettentően iszik:**	**1**

Az első érdekes eredményt a mini szótár címszavainak vizsgálatakor tapasztaltam: a szótárban található szókapcsolatoknak, csak nagyon kevés hányada fordult elő az MNSZ-ben, és az előfordultak közül is nagyon sok volt a kicsi gyakorisági számmal, 5 vagy az alattival, rendelkező. Pedig azt vártam volna, hogy a szótárban szereplő szókapcsolatok sokkal nagyobb hányada fordul elő az MNSZ-ben. Az eredményeket a 2. táblázatban gyűjtöttem össze.

Ezt az eredményt, nyelvészekkel folytatott beszélgetéseim során két dologgal tudtuk magyarázni:

- Intuitív nehezen ragadhatók meg és hiányosan sorolhatók fel a fokozó értelmű szókapcsolatok.
- Az MNSZ mérete (187,6 millió szövegszó) kicsi a fokozó értelmű szókapcsolatok kereséséhez.

Az első problémával később fogok foglalkozni, a második ellenőrzésére keressünk nagyobb korpuszt!

A nemzetközi szakirodalom áttanulmányozása során [4] találtam, egy kísérletet, ahol a New York Times korpuszban nem talált kollokációkat az AltaVista keresőben keresték, mondván az egy nagyobb korpusz. Én három internetes keresőben kerestem a mini szótáram tartalmát. Az eredményekkel még mindig nem voltam megelégedve, így a Hungarian Webcorpus-t is installáltam.

2. táblázat: A mini szótár szókapcsolatainak előfordulása az MNSZ-ben.

hányszor fordult elő egy szókapcsolat az MNSZ-ben	**db**
0-szor fordult elő	155
1-5-szor fordult elő	55
6-19-szer fordult elő	29
20-771-szor fordult elő	25
összesen:	264

2.2 A mini szótárban található címszavak vizsgálata internetes keresőkben

Mivel a keresők belső működéséről nagyon keveset árulnak el a keresőüzemeltetők, így kiválasztásukkor csak ismertségük volt segítségemre. Azon hármat választottam ki, melyek a legnagyobbak. Vizsgálódásaimhoz a következő internetes keresőket választottam:

- Google, http://www.google.com.
- Live Search, http://search.msn.com.
- Yahoo, http://www.yahoo.com.

A mini szótár 270 kollokációját kérdeztem le az említett három keresőben. Szemléltetésképpen összegyűjtöttem az *iszik* kulcsszóhoz tartozó kollokációkat a 3. táblázatban. A 3. táblázatban csak azon szókapcsolatokat sorolom föl, melyek legalább egy keresőben előfordultak. A táblázat első oszlopa tartalmazza a szókapcsolatokat, a második az MNSZ adatait, a harmadik oszlop a Webcorpus adatait és a többi oszlop

pedig a három kereső adatait tartalmazza. A Webcorpus adatainak elemzésére később kerül sor. Minden adatoszlop után megtalálható a relatív gyakoriság is, melyet az adott korpuszban vagy keresőben előforduló címszóhoz viszonyítva számolok ki. Például, az *iszik* szó a Google-ben 42 900-szer fordul elő és a *mohón iszik* 356-szor, így a *mohón iszik* címszóhoz viszonyított relatív gyakorisága: 0,08%.

3. táblázat: A Fokozó szótár *iszik* címszavában található szókapcsolatok előfordulása az MNSZ-ben, a Webcorpusban és internetes keresőkben.

	MNSZ	%	**Web-corpus**	%	**Google**	%	**MSN**	%	**Yahoo**	%
iszik	**17877**		37452		**429000**		**57934**		**251000**	
intenzíven iszik			5	0,02	9				2	
mohón iszik	22	0.12	67	0,18	356	0.08	28	0.05	74	0.03
nagyon iszik	28	0.16	22	0,06	1270	0.3	30	0.05	468	0.19
nagyot iszik			48	0,13	220	0.05	68	0.12	81	0.03
vadul iszik					9		4	0.01	7	
vég nélkül iszik			1							
borzasztóan iszik			1							
eléggé iszik			4	0,01	35	0.01	2		5	
erősen iszik	4	0.02	12	0,03	402	0.09	11	0.02		
halálian iszik			32	0,09						
igazán iszik	1	0.01	4	0,01	226	0.05	3	0.01	45	0.02
intenzíven iszik			5	0,01	9				2	
istentelenül iszik			4	0,01						
keményen iszik	11	0.06	18	0,05	227	0.05	32	0.06	108	0.04
komolyan iszik	1	0.01	3	0,01	35	0.01	3	0.01	10	0
marhára iszik			5	0,01	1					
módfelett iszik	1	0.01								
nagyon iszik	28	0.16	22	0,06	12709	0.3	30	0.05	468	0.19
piszkosul iszik										
rettenetesen iszik	1	0.01			4				1	
rettentően iszik	1	0.01			1		1			
szerfelett iszik			3	0,01						
szerfölött iszik			1	0,01						
ugyancsak iszik			9	0,02						
vadul iszik					9		4	0.01	7	
veszettül iszik					62	0.01	1		5	

A táblázatban jól látható, hogy azok a szókapcsolatok, melyek az MNSZ-ben előfordultak, az összes keresőben megtalálhatóak. És új szókapcsolatok is megjelentek, mint használt formák. Tehát az a gondolatunk, hogy ezek a keresők nyelvileg gazdagabbak és több nyelvileg releváns adatot tartalmaznak, mint az MNSZ, igaz. Azt figyelembe véve, hogy e keresőkben csak szótövesítés nélkül tudtunk keresni, az MNSZ-ben pedig szótövesítve, még jobban "felértékelődik" minden egyes megtalált szókapcsolat.

A másik érdekes dolog, hogy a keresők egységes eredményhalmazt mutatnak. Azon szókapcsolatok, melyek egy adott keresőben előfordultak, nagy valószínűséggel mindegyikben előfordulnak, és az előfordulási számuk is nagyon hasonló. A Google általában nagyobb előfordulást mond, mint a többi, de arányaiban nézve az előfordulásokat, egymáshoz képest hasonló számok jelennek meg. A Google nagy előfordulási számai mögött az is megbújik, hogy kereséskor, úgy kértem le az eredményoldalt, hogy ne csoportosítsa a találatokat, így azok a weboldalak, melyek ugyanazt a szöveget tartalmazták, külön weboldalon jelennek meg, míg ezen opció kikapcsolására a többi keresőben nincs lehetőség.

Az eredmények tanulmányozása során megállapítottam, hogy több kollokációt találtam, mint az MNSZ-ben, de ez a növekedés nem számottevő. Az a tény, hogy a keresőkben nem tudtam szótövesítve keresni, vagyishogy csak azon találatok jelentek meg egy adott keresés eredményeként, melyekben az adott kollokáció szótári alakban fordult elő, azt sugallta, hogy sokkal többször fordulhatnak elő a keresett szókapcsolatok, csak "rejtve" maradnak. Ezért döntöttem úgy, hogy megpróbálom kereshető formába hozni a Webcorpust, hogy tudjak a magyar web egy offline változatán keresni. A kereshető formába hozásra azért volt szükség, mert ez a korpusz csak mint összegyűjtött weblapokból kinyert szövegek XML-be konvertált gyűjteménye érhető el az interneten.

2.3 Találati számok a Webcorpusban

A 3. táblázat tartalmazz a Webcorpus eredményeit is. A táblázatban vastagon szedtem azon szavakat, melyek legalább egy korpuszban kétszer fordultak elő. Azon kulcsszókat, címszavakat, melyek legalább három korpuszban előfordultak dőlttel szedtem, ezekről a szókapcsolatokról valóban úgy érezzük, hogy gyakrabban használjuk.

A 3. táblázatban látható tendencia igaz a mini szótárban található összes címszóra. Vannak olyan szókapcsolatok, melyeket sokszor használunk, és első "ránézésre" is úgy tűnik, úgy érezzük, hogy használatuk gyakori. Viszont van jó néhány, a szókapcsolatok körülbelül fele, melyeket nem találhatunk meg, vagy csak nagyon kevésszer korpuszokban. Ezen szókapcsolatokat ugyan használhatnánk, de ránézésre is úgy érezzük, hogy erőltetettnek hatna. Ezen címszavak bevétele egy újabb szótárba meggondolandó.

Az eddigi vizsgálódásaim legfontosabb eredménye, hogy egy fokozó szótár ellenőrzésére adnak lehetőséget azt itt alkalmazott módszerek. Egy már kész szótárban megjelölhetőek azon szókapcsolatok, melyeket nem gyakran egy bizonyos gyakorisági szint alatt használunk, és ezután ezen szócikkeket kézzel ellenőrizve tökéletesíthetjük a szótárat azon címszavak kihúzásával, melyekre nem tartunk igényt.

3 Fokozó értelmű szókapcsolatok automatikus detektálása

Második lépésben eljárást dolgoztam ki fokozó értelmű szókapcsolatok felismerésére és detektálására magyar nyelvű szövegekben, korpuszokban. Ezen a ponton azzal az egyszerűsítéssel éltem, hogy a kulcs- és a fokozószavak halmazát is a Fokozó szótár-

ban előforduló kulcs- és fokozószavak alkotják. A halmazok bővítése egy későbbi kutatás témája lehet. A halmazok fokozatos, gyakoriság szerinti bővítése lehetőséget adhat a kollokációk csoportosítására. A fokozó értelmű szókapcsolatok detektálásának lépései:

- A kulcs- és a fokozószavak halmazának előállítása
- Az összes lehetséges kulcsszó és fokozószó pár gyakoriságának kigyűjtése
- Azon gyakorisági szint meghatározása, mely felett egy kollokációt fokozó értelmű szókapcsolatnak tekintünk

Az eljárás segítségével vizsgáltam azon kollokációkat, melyek a vizsgált korpuszokban adott gyakorisági szintet elérnek, de a szótárban nincsenek címszavak alá rendelve. Az így megtalált fokozó értelmű szókapcsolatokat címszavak alá rendezve, kézi ellenőrzés után egy új fokozó értelmű szótárat kapunk.

3.1 Kulcs- és fokozószavak halmazának előállítása

Kulcsszavak kigyűjtése, a kulcsszóhalmaz előállítása

A fokozó értelmű szókapcsolatok szótárának [1] magyar részében kulcsszavak és fokozó lexémák szerint is címszavak alá vannak rendelve a fokozó értelmű szókapcsolatok. A kulcsszavak kigyűjtésére két módszer kínálkozott:

- A címszavak kigyűjtése abból a részből, mely a kulcsszók szerint van rendezve.
- A kulcsszók kigyűjtése abból a részből, mely a fokozó lexémák szerint van rendezve.

Érdekességképpen mind a két módszerrel kigyűjtöttem a kulcsszókat. Az előbbi módszerrel 899 db kulcsszót, az utóbbival 902-t találtam.

Fokozó lexémák kigyűjtése, a lexémahalmaz előállítása

A Fokozó szótárnak [1] magyar részében a fokozó lexémák alá vannak rendelve a fokozó értelmű szókapcsolatok, itt mint címszavak jelennek meg a fokozó lexémák. A fokozó lexémákat a címszavak kinyerésével kaptam meg. Fokozó lexémából 783 db van a szótárban. Az algoritmus ugyanaz volt, mint kulcsszók kigyűjtésekor. A fokozó lexémák halmazának automatikus bővítése könnyen megvalósítható, ha a korpuszból leválogatjuk a kulcsszavak előtt, adott ablakmérettel, előforduló szavakat. A létrejött halmaz adott gyakorisági szintet elérő szavait kézi ellenőrzés után felvehetjük a fokozó lexémák közé.

3.2 Új fokozó értelmű szókapcsolatok keresése

A kigyűjtött fokozó lexémák és kulcsszavak összes lehetséges kombinációját kikerestem. A keresés megtörtént az MNSZ-ben és a Webcorpuszban is. Ezek után megvizsgáltam a kapott eredményt és megállapítottam a gyakorisági szintet.

A gyakorisági szint megállapítása

Mind a Webcorpusz, mind az MNSZ esetében a gyakorisági szintet 2-nél húztam meg, vagyis, ha egy fokozó értelmű szókapcsolat legalább 2-szer fordul elő, akkor felvettem. Az alábbi két táblázatban szerepelnek az *iszik* kulcsszóhoz detektált fokozó lexémák.

4. táblázat: Az *iszik* szó előtt előforduló fokozó lexémák az *MNSZ-ben*
(csak azon szavak szerepelnek a listában, melyek legalább 2-szer fordultak elő).

fokozószó	előford.	fokozószó	előford.	fokozószó	előford.
sok(at)	261	**mértéktelen**	6	eredeti	2
jó	41	***erős(en)***	*6*	**hosszas(an)**	**2**
mohó(n)	**29**	kér	5	öreg	2
nagyon	**28**	**sűrű(n)**	**4**	ugyan	2
nagy(ot)	***23***	sötét	4	forró	2
ritka(n)	**22**	igen	4	**csúnya(n)**	**2**
gyors(an)	**16**	**szörny(en)**	**4**	vég	2
halál(ra)	**13**	**makacs(ul)**	**3**	hisz	2
elég(et)	13	**finom(an)**	**3**	***rettentő(en)***	***2***
kemény(en)	**12**	meleg(en)	3	**derekas(an)**	**2**
vér(t)	11	**jól**	**3**	váratlan(ul)	2
biztos(an)	11	**eszméletlen(ül)**	**3**	***isten(telenül)***	***2***
állandó(an)	**10**	disznó	3	**bőséges(en)**	**2**
keserű(en)	**8**	**rendes(en)**	**3**	**bősz(en)**	**2**
halálos(an)	**7**	néma(n)	3	**kutya(ul)**	**2**
jócskán	**7**	kivételes(en)	3		

5. táblázat: Az *iszik* szó előtt előforduló fokozó lexémák a *Webcorpus*-ban
(csak azon szavak szerepelnek a listában, melyek legalább 2-szer fordultak elő).

fokozószó	előf.	fokozószó	előf.	fokozószó	előf.
sok(at)	472	**mértéktelen(ül)**	**10**	***isten(telenül)***	***4***
jó	61	**eszmélelen(ül)**	**10**	***eléggé***	***4***
mohó(n)	***61***	**jól**	**9**	***komoly(an)***	***3***
elég(et)	41	***ugyancsak***	***9***	**hosszas(an)**	**3**
ritka(n)	**40**	**állandó(an)**	**9**	***igazán***	***3***
halál(ian)	**27**	hideg	8	***szerfelett***	***3***
nagy(ot)	***24***	ugyan	8	**lázas(an)**	**3**
mély(et/en)	20	**rendes(en)**	**8**	**bátor(an)**	**3**
meleg	19	**szorgalmas(an)**	**7**	**kifejezett(en)**	**3**
kemény(en)	***18***	**néma(n)**	**7**	**sír(va)**	**3**
halálos(an)	**17**	forró	6	**kétségtelen(ül)**	**2**
nagyon	***15***	por	5	**öreg(esen)**	**2**
ennivaló	15	***intenzív(en)***	***5***	való	2
jócskán	**15**	**keserű(en)**	**5**	**teljes(en)**	**2**
gyors(an)	**14**	vég	5	**kiadós(an)**	**2**
fontos	13	**biztos(an)**	**5**	szerető	2

finom(an/at)	**12**	***marha(ra)***	***5***	**szenvedélyes(en)**	**2**
alapos(an)	11	**sűrű(n)**	**4**	**kiváló(an)**	**2**
igen	**11**	**feltétlen(ül)**	**4**	**korlátlan(ul)**	**2**
erős(en)	***11***	közel	4	**különös(en)**	**2**
bőséges(en)	**11**	**szépen**	**4**		

4 Az eljárás korlátai avagy miért szükséges a kézi ellenőrzés?

Az eljárás nem tud különbséget tenni ugyanazon szó több jelentése között. Az *iszik* szónak a Fokozó szótárban két jelentése van megkülönböztetve:

1. (ember, állat folyadékot, italt) kortyolva nyeléssel a gyomrába juttat.
2. szeszes italt az alkohol kedvéért fogyaszt.

Ezt a két jelentést együtt vizsgálja a 3. táblázat. Egy kézi szétválogatás után a további vizsgálatok az egyes szókapcsolatokra már mint külön jelentésekre is elvégezhetőek.

A kézi ellenőrzést nem lehet kikerülni, mert sok olyan fokozó lexéma van, melyek bizonyos szavaknál nem lehetnek fokozó lexémák. Az *iszik* címszónál ilyen a *vér* szó, ez a szó azért jelent meg a listában, mert sokszor fordult elő, hogy *"vért iszik",* (11-szer). Ugyanakkor a *vér* szó a fokozó lexémák közé úgy kerülhetett be, hogy szerepel az a szókapcsolat a szótárban, hogy "vérig sért". És a fokozó lexémák szótövesítését sem tudjuk kikerülni, mert különben nem tudnánk a szótövesített korpuszban keresni.

Egy másik nagy probléma, hogy az itt alkalmazott algoritmusok csak kéttagú fokozó szókapcsolatokat tudnak vizsgálni, mert a három vagy többtagú szókapcsolatok olyan kis számban fordultak elő a fokozó szótárban, hogy nem akartam a többtagúsággal bíbelődni, addig, amíg nem voltam kész a kéttagúakkal.

Azonban a legnagyobb hiányossága az eljárásnak, hogy a szótövesítés miatt a szavak szótövesítve jelennek meg a végső listában. Nyilván azért, hogy a gyakorisági szám helyesen szerepeljen. Ezért a legtöbb szókapcsolatnál kézzel kell kiírni azt a ragot, mely a kulcsszóhoz kapcsolja a fokozó lexémát. Ezek a ragot a 3. táblázatban, zárójelekben szerepelnek, ezeket kézzel írtam be a táblázatba.

5 A gyakorisági számok bevezetése

A kollokációk gyakoriságát a Webcorpuszban vizsgáltam. A szópárok gyakorisági intervallumát [1-61] felosztottam három részre, így kaptam három csoportot. Minden szókapcsolatot besoroltam egy csoportba az előfordulási száma szerint. Ezek elnevezése a következő:

- gyakran használt (3) 10-szer vagy többször fordulnak elő
- átlagosan használt (2) 4 és 9 között fordulnak elő
- ritkán használt (1) 9-nél kevesebbszer fordulnak elő

Azon szavak, melyek nem szerepeltek a Webcorpus-ban, hanem a más korpuszokból kerültek be, mint új szavak, gyakoriságukat a korpusz méretével arányosan korrigálni kellett. A korrigált gyakoriságokat mutatja az alábbi 6. táblázat.

6. táblázat: Az *iszik* címszó kollokációinak korrigált gyakorisága, a Webcorpus-hoz viszonyítva.

címszó	előf.	mely korpuszból	szorzó	korrigált gyakoriság
bőszen	2	MNSZ	2,09	4
csúnyán	2	MNSZ	2,09	4
derekasan	2	MNSZ	2,09	4
kutyául	2	MNSZ	2,09	4
makacsul	3	MNSZ	2,09	6
szörnyen	4	MNSZ	2,09	8
piszkos	9	Google	0,09	~1
rettenetesen	4	Google	0,09	~1
vadul	9	Google	0,09	~1
veszettül	62	Google	0,09	6

7. táblázat: Az *iszik* szó előtt előforduló fokozó lexémák a Webcorpus-ban (csak azon szavak szerepelnek a listában, melyek legalább 2-szer fordultak elő).

fokozószó	előf.	fokozószó	előf.	fokozószó	előf.
sok(at)	472	**mértéktelen(ül)**	**10**	***isten(telenül)***	***4***
jó	61	**eszméletlen(ül)**	**10**	***eléggé***	***4***
mohó(n)	***61***	**jól**	**9**	***komoly(an)***	***3***
elég(et)	41	***ugyancsak***	***9***	**hosszas(an)**	**3**
ritka(n)	**40**	**állandó(an)**	**9**	***igazán***	***3***
halál(ian)	**27**	hideg	8	***szerfelett***	***3***
nagy(ot)	***24***	ugyan	8	**lázas(an)**	**3**
mély(et/en)	20	**rendes(en)**	**8**	**bátor(an)**	**3**
meleg	19	**szorgalmas(an)**	**7**	**kifejezett(en)**	**3**
kemény(en)	***18***	**néma(n)**	**7**	**sír(va)**	**3**
halálos(an)	**17**	forró	6	**kétségtelen(ül)**	**2**
nagyon	***15***	por	5	**öreg(esen)**	**2**
ennivaló	15	***intenzív(en)***	***5***	való	2
jócskán	**15**	**keserű(en)**	**5**	**teljes(en)**	**2**
gyors(an)	**14**	vég	5	**kiadós(an)**	**2**
fontos	13	**biztos(an)**	**5**	szerető	2
finom(an/at)	**12**	***marha(ra)***	***5***	**szenvedélyes(en)**	**2**
alapos(an)	11	**sűrű(n)**	**4**	**kiváló(an)**	**2**
igen	**11**	**feltétlen(ül)**	**4**	**korlátlan(ul)**	**2**
erős(en)	***11***	közel	4	**különös(en)**	**2**
bőséges(en)	**11**	**szépen**	**4**		

A 7. táblázat jelölései megegyeznek a 4. táblázat jelöléseivel.

6 Az *iszik* címszó élettörténete

Összefoglalásképpen bemutatom, hogy hogyan változott az *iszik* szó munkálataim során. Ebben a részben a kihúzott szókapcsolatok áthúzva, míg a felvettek vastagon szedve jelennek meg a szócikkekben.

6.1 Az eredeti szócikk a Fokozó szótárban

ISZIK i trinken h.
1. '(ember, állat folyadékot, italt) kortyolva nyeléssel a gyomrába juttat'
♦ csillapíthatatlanul, intenzíven, mohón, nagyon, nagyot, teljes erővel, vadul, vég nélkül
◊ vedel
2. 'szeszes italt az alkohol kedvéért fogyaszt'
♦ állatian/állatira durva, baromian durva, borzalmasan biz, borzasztóan, bődületesen biz, eléggé, elképesztően biz, erősen, felettébb/fölöttébb vál, feltűnően, fenemód(on) biz, fokozott mértékben, halálian szleng, hallatlanul, határozottan, igazán, igencsak biz, intenzíven, irtózatosan biz, istentelenül biz, iszonyatosan vál, kegyetlenül, keményen, komolyan, marhára durva, meglehetősen, módfelett/módfölött vál, nagyon, nem mindennapi mértékben, oltárian szleng, őrületesen biz, piszkosul szleng, pokolian, rendkívüli módon, rettenetesen/rettentően biz, roppant mód(on), szédületesen biz, szerfelett/szerfölött vál, túlságosan, túlzottan, ugyancsak, vadul, veszettül
❀ ~, mint a kefekötő/gödény

6.2 A ritkán használt szókapcsolatok törlése

ISZIK
~~csillapíthatatlanul~~, intenzíven, mohón, nagyon, nagyot, ~~teljes erővel~~, vadul, ~~vég nélkül~~, ~~állatian~~, ~~állatira~~, ~~baromian~~, ~~borzalmasan~~, ~~borzasztóan~~, ~~bődületesen~~, eléggé, ~~elképesztően~~, erősen, ~~felettébb~~, ~~fölöttébb~~, ~~feltűnően~~, ~~fenemódon~~, ~~fokozott mértékben~~, halálian, ~~hallatlanul~~, ~~határozottan~~, igazán, ~~igencsak~~, intenzíven, ~~irtózatosan~~, istentelenül, ~~iszonyatosan~~, ~~kegyetlenül~~, keményen, komolyan, marhára, ~~meglehetősen~~, ~~módfelett~~, ~~módfölött~~, nagyon, ~~nem mindennapi mértékben~~, ~~oltárian~~, ~~őrületesen~~, piszkosul, ~~pokolian~~, ~~rendkívüli módon~~, rettenetesen, ~~rettentően~~, ~~roppant módon~~, ~~szédületesen~~, szerfelett, ~~szerfölött~~, ~~túlságosan~~, ~~túlzottan~~, ugyancsak, vadul, veszettül

6.3 A felderített új szókapcsolatok

ISZIK
állandóan, ~~állatian~~, ~~állatira~~, ~~baromian~~, **bátran**, **biztosan**, ~~borzalmasan~~, ~~borzasztóan~~, ~~bődületesen~~, **bőségesen**, **bőszen**, ~~csillapíthatatlanul~~, **csúnyán**, **derekasan**, eléggé, ~~elképesztően~~, erősen, **eszméletlenül**, ~~felettébb~~, **feltétlenül**, ~~feltűnően~~, ~~fenemódon~~, **finoman**, ~~fokozott mértékben~~, ~~fölöttébb~~, **gyorsan**, halálian, **halálosan**, ~~hallatlanul~~, ~~határozottan~~, **hosszasan**, igazán, **igen**, ~~igencsak~~, intenzíven, ~~irtózatosan~~, istentelenül,

~~iszonyatosan~~, **jócskán**, **jól**, ~~kegyetlenül~~, keményen, **keserűen**, **kétségtelenül**, **kiadósan**, **kifejezetten**, **kiválóan**, komolyan, **korlátlanul**, **kutyául**, **különösen**, **lázasan**, **makacsul**, marhára, ~~meglehetősen~~, **mértéktelenül**, ~~módfelett~~, ~~módfölött~~, mohón, nagyon, nagyot, ~~nem mindennapi mértékben~~, **némán**, ~~oltárian~~, **öregesen**, ~~őrületesen~~, piszkosul, ~~pokolian~~, **rendesen**, **ritkán**, ~~rendkívüli módon~~, rettenetesen, ~~rettentően~~, ~~roppant módon~~, **sírva**, **sűrűn**, ~~szédületesen~~, **szenvedélyesen**, **szépen**, **szörnyen**, szerfelett, ~~szerfölött~~, **szorgalmasan**, ~~teljes erővel~~, **teljesen**, ~~túlságosan~~, ~~túlzottan~~, ugyancsak, vadul, ~~vég nélkül~~, veszettül

A fenti szócikkben az eredeti szócikket egészítettem ki azokkal a szópárokkal, melyeket a 4.1. táblázatban vagy a 4.2. táblázatban újként jelöltem meg.

6.4 A jelentések szétválasztása

ISZIK i
1.'(ember, állat folyadékot, italt) kortyolva nyeléssel a gyomrába juttat'
♦ állandóan (2), bőségesen (3), finoman (3), gyorsan (3), hosszasan (1), intenzíven (2), kiadósan (1), korlátlanul (1), mértéktelenül (3), mohón (3), nagyon (3), nagyot (3), ritkán (3), sírva (1), sűrűn (2), szerfelett (1), teljesen (1), ugyancsak (2)
2. 'szeszes italt az alkohol kedvéért fogyaszt'
♦ állandóan (2), bátran (1), biztosan (2), bőségesen (3), bőszen (2), csúnyán (2), derekasan (2), eléggé (2), erősen (3), eszméletlenül (3), feltétlenül (2), gyorsan (3), halálian (3), halálosan (3), hosszasan (1), igazán (1), igen (3), intenzíven (2), istentelenül (2), jócskán (3), jól (2), keményen (3), keserűen (2), kétségtelenül (1), kiadósan (1), kifejezetten (1), kiválóan (1), komolyan (1), korlátlanul (1), kutyául (2), különösen (1), lázasan (1), makacsul (2), marhára (2), mértéktelenül (3), mohón (3), nagyon (3), nagyot (3), némán (2), öregesen (1), piszkosul (1), rendesen (2), rettenetesen (1), ritkán (3), sírva (1), sűrűn (2), szenvedélyesen (1), szépen (2), szerfelett (1), szorgalmasan (2), szörnyen (2), teljesen (1), ugyancsak (2), vadul (1), veszettül (2)

7 Eredmények: a létrejött új szótár

A kutatás végén a kiindulásként használt Fokozó szótár [1] teljes kibővített változata csak az eredmények kézi ellenőrzése után lehetséges, mert a szótárban szereplő címszavak jelentés szerinti csoportosítását nem tudtam automatikusan megoldani. A mini szótáramban található 7 címszóra elvégeztem a kézi ellenőrzést. A mini szótár 7 címszavában található 270 szókapcsolatából 92-t ítéltem meg úgy, hogy törölhető és mintegy 287 új szókapcsolatot detektáltam.

Hivatkozások

1. Székely G.: A fokozó értelmű szókapcsolatok magyar és német szótára. Tinta Könyvkiadó, Budapest (2003)

2. Székely G.: Egy sajátos nyelvi jelenség, a fokozás; Tinta Könyvkiadó. Budapest (2007)
3. Székely G.: A lexikai fokozás. Scholastica kiadó. Budapest (2001)
4. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
5. Oakes, M. P.: Statistics for Corpus Linguistics. Edinburg University Press (1998)

Adó- és jövedéki jogi wordnet (TaXWN)

Almási Attila[1], Vincze Veronika[1], Sulyok Márton[2], Csirik János[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
vizipal@gmail.com, vinczev@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Alkotmányjogi Tanszék
Szeged, Tisza Lajos krt. 54.
msulyok@juris.u-szeged.hu
[3] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
csirik@inf.u-szeged.hu

Kivonat: A magyar jogi fogalomháló eddig elkészült része (TaXWN) adó és jövedéki tárgykörből tartalmaz mintegy 650 synsetet. Első lépésben a wordnet építéséhez szükséges számítógépes előfeldolgozási munkára került sor. Ezt követte a TaXWN mint adatbázis létrehozása, majd létrehoztuk a synsetjelöltek definícióit és hozzárendeltük a jogi szempontból nélkülözhetetlen megjegyzéseket. Ezután következett a fogalmi háló felépítése. Alapszintű kapcsolódás jött létre a LOIS nemzetközi jogi wordnettel. A rendszer továbbfejleszthető egyrészt a HuWN irányába, másrészt pedig egyéb jogterületek fogalmaival is bővíthető.

1 Bevezetés

A modern információs társadalomban kulcsfontosságú szerepet tölt be az információ gyors megszerzése és hatékony kezelése. A felhasználók széles körének szerteágazó igényei mellett a legkülönfélébb területeken dolgozó szakemberek számára is elengedhetetlenné vált, hogy megfelelő gyorsasággal tudjanak tájékozódni az adott szakirodalomban, az adott probléma megoldásához hatékonyan tudják előkeresni a kapcsolódó háttérismereteket, és egyeztetni tudják egymással a különböző – esetlegesen akár többnyelvű – forrásokból származó információkat. A hatékonyság jellemzői közül a visszakeresett információ terjedelmét, minőségét, relevanciáját, a megszerzéshez szükséges időt és költségeket emeljük ki.

A jogi szövegek számítógéppel támogatott feldolgozásához és kutatásához szükséges a jogi szakszövegek digitalizálása, az elektronikus formában való nyilvánossá tétele és az elektronikus dokumentumkezelés biztosítása. Ez utóbbi előfeltétele egy, az adott szakterület fogalomkészletét lefedő ontológia [8].

A magyar jogi wordnet kezdeményezés a magyar jogrendszer nemzetközi (közelebbről európai uniós) jogrendbe történő integrációjában tölthet be fontos szerepet, hiszen egy olyan jogi tudásbázist alakít ki, amely az uniós jogközelítés, jogharmonizáció lexikális hátterét adja, ezáltal megkönnyíti a szükséges változtatások elvégzését. A cél tehát az EuroWordNet [2] eredményeire és formalizmusára építő, szemantikai-

lag strukturált, jogi vonatkozású fogalomtár létrehozása a magyar nyelvre, melyet a jog, az informatika és nyelvészet tudományágai együttműködve alkotnak meg. Ennek első lépéseként az adó- és jövedéki jogi wordnet (TaXWN) készült el.

2 A magyar jogi wordnettel kapcsolatos informatikai feladatok

Első lépésben a wordnet építéséhez szükséges számítógépes előfeldolgozási munkálatokra került sor. A rendelkezésre álló jogszabályok, illetve egyéb szakirodalmi anyagok feldolgozásával terminológiajelölt listák kerültek előállításra, majd a továbbiakban ezek felhasználásával, lényegében emberi erővel történt meg a tényleges adatbázis-fejlesztés.

Az informatikai feladatok a wordnet kézi építése során döntően különböző formai (szólisták konverziója az adatbázis formátumára, automatikusan kitölthető adatbázis mezők felvitele, stb.), illetve validációs (az adatbázis formalizmusa által támasztott követelményeknek meg nem felelő fogalmak kiszűrése, szintaktikailag hibás fogalmak szűrése, stb.) munkákra korlátozódtak. A felmerült feladatok elvégzéséhez, valamint az adatbázis automatikus minőségbiztosítási mechanizmusának kifejlesztéséhez elkészült egy, az adatbázis formátumát felolvasni képes osztálykönyvtár, melyben az egyes – a nyelvész kollégák által megfogalmazott, illetve strukturális (XML-validáció) – szűrések gyorsan és egyszerűen megvalósíthatók voltak.

3 A magyar jogi wordnettel kapcsolatos jogi feladatok

A projektnek az SZTE Állam- és Jogtudományi Kar Alkotmányjogi Tanszék részéről koordinált részének célkitűzései a következők voltak:

a) az adózás rendjéről szóló, 2003. évi XCII, illetve az adózás és jövedék témakörökhöz kapcsolódó egyéb jogszabályok szakterminológiájának kivonatolása;
b) egy – elsősorban eljárási, jogalkalmazói felhasználási célhoz igazodó – értelmező szótár készítése a fenti kivonat alapján.

A kivonatolandó fogalmi kört először úgy próbáltuk meghatározni, hogy vettük az adott témához kapcsolódó irodalmi terminológiát (a továbbiakban: lit-források), illetve a kapcsolódó jogszabályok terminológiáját (a továbbiakban: lex-források). Miután nyilvánvalóvá vált, hogy a két forrásanyag között olyan eltérések lehetnek, amelyek a fentebb leírt alapcélokat ellehetetlenítik, így a későbbiekben el kellett tekintenünk a lit-források használatától[1]. Ennek oka az volt, hogy az egyes lit-források gyakorta nem azonos definíciót használtak az egyes lex-forrásokban fellelhető fogalmakra, ami egyrészt a szerzői szabadság és szubjektivitás, másrészt pedig az oktatási célra fel-

[1] A későbbiekben a lex-forrásokból nyert fogalmak értelmetlensége és esetleges hiányossága esetén pontosításokra, egyértelműsítésre használtuk fel a lit-források anyagait számos esetben, a jogalkalmazó munkájának megkönnyítése céljából.

használandó szellemi alkotásoktól elvárt közérthetőség követelményéből adódik, s ez gyakran a szakkifejezések egyszerűsítését, átfogalmazását igényli. A jogalkalmazó azonban elsősorban a jogforrásokban, jogszabályokban fellelhető fogalmakra támaszkodik munkája során, tehát ezek elsőbbsége egy jogi wordnet kialakításában indokolt.

Mégis felhasznált irodalmi forrás vagy saját elképzelések alapján átalakított definíciók esetében kiegészítő információként (*Megjegyzés / Note*) jelöltük a hivatkozott forrást, illetve „egyéni" címkével látjuk el az adott meghatározást (vö. 4.1 és 4.2).

A jogi wordnet alapjául szolgáló szakterminológia-kivonat XLS formátumú, ún. *LEXtract* (jogszabályi kivonat) listákba rendezve az alábbi elemeket tartalmazta: a kifejezéseket, azok definícióit és a többletinformációt tartalmazó megjegyzéseket.

1. táblázat: TaXWN LEXtract

TaXWN_LEXtract		
Kifejezés	*Definíció*	*Megjegyzés*
adatgyűjtésre irányuló ellenőrzés	Olyan ellenőrzési eljárás, amelynek célja az adóhatóság nyilvántartásában és az adózó nyilvántartásában, bevallásában szereplő adatok, tények, körülmények valóságtartalmának, illetőleg ezek hitelességének megállapítása.	Art. 119 § (1); Az adatgyűjtésre irányuló ellenőrzés során az adóhatóság a bevallási időszak lezárását megelőzően is adatokat gyűjthet.
adatkérés	A kapcsolattartó közigazgatási szervhez az Európai Közösség tagállami illetékes hatósága által küldött olyan kérelem, amelyben ez utóbbi a tartozás behajtásához szükséges adatok átadását kéri.	Art. 61 § (1)
adóbevallás	Az adózó azonosításához, az adóalap, a mentességek, a kedvezmények, az adó, a költségvetési támogatás alapja és összege megállapításához szükséges adatokat tartalmazó nyilatkozat.	Art. 31. § (1)-(14); Jöt. 48./B§ (2) Az adóalany adóbevallási kötelezettségét elektronikus úton a külön jogszabályban foglalt módon és technikai feltételekkel teljesíti.

4 A magyar jogi wordnettel kapcsolatos nyelvészeti feladatok

A szűrés, listázás és átválogatás után megmaradt fogalomjelölteket synsetekbe rendeztük. Ezt követően került sor a definíciók, megjegyzések, valamint a hierarchia létrehozására. A szakontológia építésével kapcsolatban itt kell megemlítenünk, hogy mivel a jelen jogi wordnet adó- és jövedéki ontológiája túlzottan specifikus, annak synsetjei gyakran egyeleműek, ami az általános wordnetek (például HuWN, PWN) esetében ritkaságnak számít.

4.1 A definíciók létrehozása

A nyelvészet és a jog által támasztott követelmények gyakran ellentmondásba kerültek, ezért ki kellett mondanunk, hogy a rendszer építésénél csak az egyik tudományterület (jelen esetben a jog) igényeinek felelhetünk meg teljesen, de amennyiben lehetséges, a nyelvészet követelményeit is megpróbáljuk figyelembe venni.

A feladat kezdetén tehát rögzítettük, hogy a rendszer egy jogi alapon álló fogalmi háló (wordnet) lesz. Ennek következtében például módosult az a wordnetépítésben megszokott nyelvi szabály is, miszerint egy fogalom definíciójának tartalmaznia kell a fogalom egy hipernimáját (egy általánosabb fogalmat) vagy annak valamely szinonimáját [1]. Ez az esetek nagy részében nem így történt, mivel a definíciók – amelyeket jogász szakértők állítottak össze törvények szövegeire támaszkodva – gyakran csupán fölsorolások, melyeknek egyes elemei egy nyelvészeti szempontból kielégítő hálóban csak meronimák lehettek volna.

Így rendszerünk olyan definíciókat tartalmaz, amelyek teljes mértékben kielégítik a jogtudomány támasztotta igényeket, és esetenként a nyelvészeti elvárásoknak is megfelelnek.

4.2 A megjegyzések kialakítása

A magyar wordnet (HuWN) [4, 5] létrehozásakor a *Megjegyzés* a synseten belül egy olyan egység volt, amely a megállapodás szerint rövid kiegészítő megjegyzések felvételét tette lehetővé. Itt elsősorban a Princeton WordNetben (PWN) [1] és a HuWN-ben eltérő szófajúként megjelenő synseteket jelöltük, valamint ide kerültek a gazdasági szakontológia synsetjeit jelölő „*szak*” megjegyzések. Ezeken kívül, a javítási fázist megelőzően lehetőség volt arra, hogy a synset létrehozását végző nyelvész saját megjegyzéseit is felvehesse, amely a későbbiekben támpontot nyújtott a javítást végző munkatársaknak.

A jogi szakontológiában azonban a *Megjegyzés* egy teljesen eltérő funkcióval rendelkezik. Ide kerültek azok az információk, amelyek magába a definícióba nem fértek be, de a meghatározandó fogalommal kapcsolatban olyan adatokat tartalmaznak, amelyek szerepeltetése nélkül a kívánt jogi tartalom nem lenne teljes. Ezenkívül itt találhatók azok a kiegészítések is, amelyek arról nyújtanak információt, hogy az adott fogalom mely törvényben lett szabályozva, hogy milyen számszerű adatok lehetnek lényegesek a jövendő felhasználó számára a fogalommal kapcsolatban (pl. alkoholfok, importálható árucikk mennyisége stb.).

4.3 A hierarchia létrehozása

A TaXWN megalkotásának lényegi mozzanata volt a fogalmi háló felépítése. Ennek létrehozásakor nem támaszkodhattunk egy már elkészült rendszerre, mint pl. a HuWN esetében, s ez egyszerre jelentett könnyebbséget és nehézséget is. Könnyebbséget azért jelentett, mert a hierarchia létrehozásánál nem kellett azzal törődnünk, hogy az egy másik rendszerrel összevethető, esetleg összekapcsolható legyen. Nehezebb azért volt, mert ebben az esetben saját kútfőre támaszkodva kellett egy használható, értelmes hierarchiát fölállítanunk.

Újabb csomópontok
A hierarchia kialakítása során a *bottom-up* módszert követtük, mert a törvényi forrásokból származó anyag igen specifikus kifejezések, s ezáltal általában csak alapsynsetek létrehozását tette lehetővé. Ezzel egyébként a munka egyszerűbbé is vált, mivel a jogi domént elhagyva a hipernimákat legtöbbször már a HuWN synsetjeinek és hierarchiájára támaszkodva tudtuk kiválasztani.

Az úgynevezett *unique beginner*
Rendszerünkben kilenc *unique beginner* synset található, amelyek a hierarchia legáltalánosabb synsetjei. Ezeknek a synseteknek a megtalálása általában magától értetődő volt, máskor viszont hosszadalmas utánajárást igényelt. Ennek oka először is a hierarchiaépítés megegyezés szerinti első szabályában keresendő, miszerint a jogi wordnet egy, a jogi szakszókincs által körülhatárolt háló lesz és a rendszer kialakítása során eltekintünk attól, hogy minden esetben a specifikusabb nyelvészeti szempontokat vegyük figyelembe. Így fordulhatott többször elő, hogy egy, az alapsynset szintjén még tárgyként azonosított elem magasabb szinten egy nem tárgyként azonosítható hipernima alá, a legfölső szinten pedig akár az *elvont fogalom* vagy *állapot* alá került bekötésre. Azonban jogi wordnetünkben nem lehetséges az összes ilyen, nyelvészeti szempontból úgymond lehetetlen állapot megszüntetése. A jog nyelvezetéből és a fentebb említett megállapodásból eredően ezek a látszólagos „következetlenségek" meg kell, hogy maradjanak.

5 Kapcsolódás a LOIS-hoz

A LOIS Project, vagyis a *Lexical Ontologies for Legal Information Sharing* rendszere hatékony, európai szintű, információs-kommunikációs, fejlesztési együttműködést céloz meg. Ez a program a EuroWN-en keresztül kapcsol össze 6 különböző tagállami wordnetet (cseh, angol, német, holland, portugál és olasz) [6, 7].

Az SZTE Informatikai Tanszékcsoport és a LOIS konzorcium vezető intézménye, az *Institute of Legal Information Theory and Techniques* között létrejött megállapodás alapján vállaltuk, hogy megvizsgáljuk a jogi wordnet LOIS-hoz való kapcsolásának lehetőségét, és a kutatás eredményeit megosztjuk a LOIS projekt felelőseivel.

A LOIS megközelítőleg 7000 kifejezést tartalmazó angol nyelvű általános jogi terminológiájának XML-fájljából kivontuk a LEMMA="kifejezés" sorokat (ebben szerepeltek a magyar terminológiával összehasonlítandó szavak), majd azokat össze-

vetettük a magyar jogi wordnet mintegy 650 kifejezésből álló adó- és jövedéki terminológiájával.

Maga az összekapcsolás a következő módon történt: az egyező synsetek, fogalmak LOIS-ban található azonosítóját fölvettük – a TaXWN építésénél használt VisDic szerkesztő [3] segítségével – a megfelelő synset *Megjegyzés* ablakába, mégpedig a következő formában: LOIS ID=”xxx”.

A kapcsolódási pontok megtalálása nem volt egyszerű feladat, mert az a magyar jogi terminológia, amely a céltartomány (adó, jövedék) kiegészítéseként kellett belekerüljön a TaXWN rendszerébe – a fogalmi pontosság és a megfelelő synset kapcsolatok kialakítása végett – nem minden esetben volt pontosan megfeleltethető az uniós (ez esetben angol nyelvű) terminológiának. Például:

A LOIS-ban szereplő kifejezések: *company, business, undertaking, entreprise, firm, corporation, concern, business corporation*
A TaXWNkifejezései: *társaság, cég, vállalkozás, vállalat, vállalkozási tevékenység*

1. ábra. A LOIS és a TaXWN kifejezéseinek megfeleltetése.

A nyolc angol kifejezéseknek öt magyar felel meg valamilyen mértékben, de az egy az egyhez való megfeleltetés jogi szempontból nem lehetséges.

A kapcsolódás gyakran azért sem lehetséges, mert a LOIS definícióiban sok helyütt úgy utal az egyes kifejezésekre, mint pl. „ezen Egyezmény értelmében” vagy „a 25. cikkben írtakkal összhangban”. Ezek a kontextusok a mi terminológiánkra vonatkoztatva elvesztik értelmüket, mivel anyagunk nem tartalmazza a közösségi terminológia által használt kifejezéseket, illetve forrásanyagokat.

Nyelvész és jogász munkatársaink tehát megvizsgálták az összes kifejezés angol és magyar nyelvű definícióját, hogy a lehető legmegfelelőbb kapcsolatokat alakíthassák ki.

Nehézséget okozott még, hogy a LOIS nem minden esetben ad meg definíciót és/vagy példát az adott synsethez, ezért ezeknek a synseteknek a minden kétséget kizáró azonosítására nem volt lehetőségünk, s így nem is vettük föl őket a TaXWN-be.

Előfordult, hogy a TaXWN-ben szereplő synset egyszerre több LOIS synsethez is kapcsolható volt. Ez arra hívja föl a figyelmet, hogy ilyen specifikus, magyar nyelvű szakszóanyagot nem lehet minden esetben egyértelműen megfeleltetni például egy általánosabb, ebben az esetben angol nyelvű jogi kifejezésnek.

Az összekapcsolás nehézségeinek bemutatására áll itt a *minor/juvenile* és a *kiskorú/fiatalkorú* kifejezések példája.

A LOIS-ban:

```
  <WORD_MEANING        ID="1429"        PART_OF_SPEECH="N"
STATUS="FINISHED">
  <SOURCEBASE>LEXDB</SOURCEBASE>
  <NOTE/>
```

```
  <GLOSS>a person who has not reached full legal
age</GLOSS>
  <CONCEPTS/>
  <VARIANTS>
  <LITERAL LEMMA="minor" SENSE="1">
  <EXAMPLES>not of legal age; &amp;quot;minor
children&amp;quot;</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="minor" SENSE="1">
   <EXAMPLES>a person who has not reached full legal age;
a child or juvenile</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="juvenile" SENSE="1">
   <EXAMPLES>a person who has not reached the age
(usually 18) at which one should be treated as an adult
by the criminal justice system</EXAMPLES>
   </LITERAL>
   </VARIANTS>
   <INTERNAL_LINKS>
   <RELATION ID="3089" TYPE="involved_patient">
   <TARGET_WM ID="379" PART_OF_SPEECH="N"/>
   </RELATION>
   </INTERNAL_LINKS>
   <EQ_LINKS>
   <RELATION ID="14290000" TYPE="eq_has_hyperonym">
   <TARGET_WM ID="2858" PART_OF_SPEECH="N"/>
   </RELATION>
   </EQ_LINKS>
   </WORD_MEANING>
```

A TaXWN-ben:

<SYNSET>
<ID>HuWN-1665382021</ID>
<SYNONYM>
<LITERAL>**kiskorú**
<SENSE>0</SENSE>
</LITERAL>
</SYNONYM>
<DEF>A. 18. életévét be nem töltött személy, kivéve, ha házasságot kötött.</DEF>
<SNOTE></SNOTE>
<SNOTE>Art.35.§ (2) i) A szülői felügyelet alatt álló kiskorú adótartozásáért a szülői felügyeletet gyakorló szülő az általa kezelt va</SNOTE>
<SNOTE>Art.5.§ (2) b</SNOTE>
<SNOTE>LOIS ID="1429"; a magyar jogrendben kis- és fiatalkorú megkülönböztetés l</SNOTE>
<SNOTE>jog</SNOTE>
<POS>n</POS>

```
    <ILR>HuWN-148541600
      <TYPE>hypernym</TYPE>
    </ILR>
    <STAMP>almasi 2008/12/02</STAMP>
   </SYNSET>

   <SYNSET>
    <ID>HuWN-911671085</ID>
    <SYNONYM>
     <LITERAL>fiatalkorú
      <SENSE>0</SENSE>
     </LITERAL>
    </SYNONYM>
    <DEF>Fiatalkorú az, aki a bűncselekmény elkövetésekor tizennegyedik életévét betöltötte, de a tizennyolcadikat még nem.</DEF>
    <SNOTE>1978. évi IV. tv. Btk. 107.§. (</SNOTE>
    <SNOTE>LOIS ID="1429"; a magyar jogrendben kis- és fiatalkorú megkülönböztetés l</SNOTE>
    <SNOTE>jog</SNOTE>
    <POS>n</POS>
    <ILR>HuWN-148541600
      <TYPE>hypernym</TYPE>
    </ILR>
    <STAMP>almasi 2008/12/02</STAMP>
   </SYNSET>
```


A TaXWN-ben a *kiskorú*: **A 18. életévét be nem töltött személy, kivéve, ha házasságot kötött;** míg a *fiatalkorú*: **Az, aki a bűncselekmény elkövetésekor tizennegyedik életévét betöltötte, de a tizennyolcadikat még nem.** Tehát, amíg a LOIS-ban a *minor* (kiskorú) kifejezésbe jogilag beletartozik a *juvenile* (fiatalkorú) kifejezés is, addig a TaXWN azokat külön kezeli, saját definíciókkal. Más kérdés, hogy a LOIS-ban a *juvenile* kifejezésnek van saját definíciója, noha a *minor*-é eleve magában foglalja a *juvenile*-t is. A magyar büntetőjogi terminológia viszont nem a kiskorúak részeként kezeli a fiatalkorúakat, hanem jogilag külön kategóriába sorolja őket. Ezt a két fogalmat végül mi is külön vettük föl és a *Megjegyzés*-ben jelöltük, hogy jogi értelemben a magyar terminológia hogyan különbözteti meg a *kiskorú* és *fiatalkorú* kifejezéseket.

6 Statisztika

2. táblázat: A TaXWN synsetjeinek megoszlása.

	LOIS-hoz kapcsolható	LOIS-hoz nem kapcsolható	TaXWN
általános	81	116	197
jövedéki	113	337	450
összesen	194	453	647

Amint látható, a TaXWN 647 synsetje közül biztonsággal 194-et lehetett a LOIS nemzetközi jogi ontológia synsetjeihez kapcsolni. Ebből 113 szorosan kapcsolódik a jövedéki terminológiához, 81 pedig általánosabb tartalmú kifejezés.

A TaXWN-en belül az adó és jövedéki területhez szorosan kapcsolódó, törvényi anyagokkal megtámogatott synsetek száma 450, az általánosabb kifejezéseket tartalmazó synseteké pedig 197. A 647 synset egy kilenc fából álló fogalomhálót képez a rendszerben meglévő *unique beginner* synsetek (*állapot, cselekmény, cselekvés, együttes/összesség, elvont fogalom, entitás, hely, jelenség, tulajdonság*) alapján. Azonban gyakran előfordul, hogy már a jogi tartalmú synset hipernimája elhagyja a szűken értelmezett jövedéki domént és egy általános fogalmi hálóba illő synsetnek tekinthető.

7 A rendszer bővítésének, frissítésének lehetőségei

Először is a magyar jogi wordnet eddig elkészült részét tovább bővíthetjük más törvényi anyagok (pl. büntetőjog, polgári jog, stb.) kifejezéseivel és egy összetettebb hierarchiába szervezhetjük őket.

Másodszor, a magyar jogi wordnet és HuWN összekapcsolása kétféleképpen is elképzelhető:

1. A HuWN-ben már meglevő jogi tartalmú synsetek összehangolása a magyar jogi wordnettel

Itt a következő probléma adódik. A HuWN-be korábban fölvett, jogi tartalmú synsetek rendszerint nem ütik meg azt a mércét, amivel egy jogi fogalomtárnak rendelkeznie kell. Ennek egyik oka az, hogy a HuWN elkészítésében nem vett részt jogász szakértő, másrészt azt is érdemes újra megjegyezni, hogy a HuWN synsetjeinek döntő többsége a PWN-ből került átvételre, lett lefordítva. A fordítási nehézségeken túl gondot okoz még az is, hogy a PWN-be az angolszász jogrend szakkifejezései kerültek be. Ezeknek nem mindig van megfelelője a magyar (és/vagy uniós) jogrendben, vagy pedig nem egészen azonos a jogi tartalmuk, ezért ezen problémák kiküszöbölésére jogász szakértő segítségét kell igénybe venni.

2. A magyar jogi wordnet általános (nem jövedéki jogi) synsetjeinek összekapcsolása a HuWN-nel

Mint fentebb említettük, a jogi ontológia egy igen szűk tartományban mozog és azt elhagyva, magasabb szinteken már átválthat az általános magyar nyelvi ontológiába, így a magyar jogi wordnet és a HuWN összekapcsolása könnyen megvalósítható.

Harmadszor, a LOIS teljes anyagát át lehetne ültetni magyar nyelvre, ezáltal Magyarország is teljes mértékben kapcsolódhat a nemzetközi jogi adatbázishoz.

Negyedszer, a rendszer frissítése bizonyos időközönként szükségesnek bizonyul, ami a jogalkalmazó munkája szempontjából mindenképpen elengedhetetlen. Az egyes jogszabályok esetleges megváltozása így könnyen nyomon követhetővé válik a hazánkban ismert nagyobb elektronikus jogszabálygyűjteményeknek, illetve online adatbázisoknak a rendszerbe történő automatikus integrálása által.

8 A rendszer felhasználhatóságának lehetőségei

A magyar jogi wordnet az elektronikus dokumentumkezelés ideális kiegészítő programja lehet, amely, többek között, például a LOIS Project adatbázisain keresztül megteremtheti a szükséges többnyelvű tudásbázist ahhoz, hogy az adott tagállamban eljáró szerv az általa nem ismert nyelven kitöltött dokumentumokat is könnyen értelmezhesse, kezelhesse, és ezzel az ügyintézési időt lerövidítse. A jogi korpusz egyéb felhasználási területei lehetnek még például az emberi ellenőrzés mellett futó, de alapvetően automatikus szakfordító programként történő alkalmazás, vagy éppen egy webalapú, bárki számára hozzáférhető, jogi szakszótári funkció is.

9 Összefoglalás

A TaXWN létrehozásával egy általános magyar jogi wordnet első lépcsője, az adó- és jövedéki alapterminológiát tartalmazó rész valósult meg. A fogalmi háló létrehozása után következő munkálatok magukban foglalták a LOIS jogi wordnethez való csatlakozás lehetőségének megteremtését, ami először is a TaXWN és a LOIS angol nyelvű változatában szereplő fogalmak, synsetek összevetését jelentette. Ezt követte azután a LOIS-ban egyezést mutató synsetek azonosítóinak fölvétele – a nyelvközi index (ILI) segítségével – a TaXWN-be. Jelen állapotban az egyirányú indexelés miatt csak a magyar jogi wordnet felől érhetők el a LOIS synsetjei. Az elkészült rendszer nagyságrendjében és minőségében elérte a kezdeti elvárásokat és jó elméleti és gyakorlati alapot nyújthat egy következő lépésben létrehozandó egyéb jogterületet lefedő fogalmi hálónak. A HuWN jogi synsetjeinek javítása és a LOIS jogi wordnethez történő, magasabb szintű, átfogóbb csatlakozás pedig későbbi projektek témája lehet.

Köszönetnyilvánítás

A kutatást – részben – a TUDORKA és a MASZEKER projekt (Jedlik Ányos programok) keretében az NKTH támogatta.

Hivatkozások

1. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an On-line Lexical Database. International Journal of Lexicography, Vol. 3, No. 4 (1990) 235–244
2. Alonge, A., Bloksma, L., Calzolari, N., Castellon, I., Marti, T., Peters, W., Vossen P.: The Linguistic Design of the EuroWordNet Database, Computers and the Humanities. Special Issue on EuroWordNet, Vol. 32, No. 2–3 (1998) 91–115
3. Horák, A., Smrž, P.: New Features of Wordnet Editor VisDic. Romanian Journal of Information Science and Technology. Special Issue on BalkaNet, Vol. 7, No. 1–2 (2004) 201–213
4. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy.: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. Project report. In: Proceedings of the Third International WordNet Conference (GWC2006), January 22–26, South Jeju Island, Korea (2006) 291–292
5. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. GWC 2008, University of Szeged, Department of Informatics (2008) 311–320
6. Sagri, T., Tiscornia, D.: Semantic Lexicons for Accessing Legal Information. In: Traunmüller, R. (ed.): Electronic Government, Third International Conference, EGOV 2004 Proceedings (Zaragoza, Spain, 2004 30 August - 4 September) (2004) 72-81
7. Peters, W.: The LOIS Project. In: Sojka, P., Choi, K.-S., Fellbaum, C., Vossen, P. (eds.): GWC 2006, Proceedings (2006) 331–332
8. Sulyok M., Gyenge B.: Jog és nyelv kapcsolata egy nem mindennapi vállalkozásban. Közjogi Szemle, 2. évfolyam, 2009. szeptember, (2009) 49-60

A jól szerkesztett mérnöki ontológiákról

Szőts Miklós, Simonyi András

Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7.
e-mail:{szots,simonyi}@all.hu

Kivonat Az ImportNET projekt keretein belül folytatott munkánk során azzal a problémával szembesültünk, hogy nem léteznek a mérnöki tervezést hatékonyan segítő ontológiák. Cikkünkben olyan általános ontológiatervezési elveket és mintákat mutatunk be, melyek segítségével jól strukturált, a mérnöki szemlélethez közelálló csúcsontológiák hozhatók létre.

Kulcsszavak: mérnöki ontológia, az ontológiatervezés módszertana, ontológiamodularizáció, ontológaszegmentálás

1. Bevezetés

Az ImportNET projekt[1] [8] egy kollaboratív mechatronikai tervezést segítő, ontológiaalapú szoftver létrehozását tűzte ki célul. A megvalósult rendszer a kollaboratív tervezési folyamat megkezdésekor egy átfogó mechatronikai doménontológiából választja ki azt az ontológiaszegmenst (az ún. kollaborációs ontológiát), amely az adott kollaboráció szempontjából releváns mechatronikai tudást tartalmazza. Az ontológia szegmentálása félautomatikusan történik: a domént jól ismerő, de a formális ontológiák területén járatlan szakértő egy grafikus felhasználói felületen kiválaszt néhány, a tervezés során várhatóan releváns, illetve bizonyosan irreleváns fogalmat és relációt, és a rendszer ennek alapján automatikusan generál egy kollaborációs ontológiát, amelyet a felhasználó tovább finomíthat.

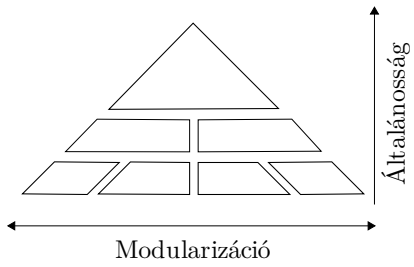
A projekt keretében végzett munkánk során azzal a problémával szembesültünk, hogy a doménontológia naív felhasználók általi szerkesztésének támogatása, illetve a szegmentálás csak megfelelően strukturált, jól szerkesztett ontológián végezhető el hatékonyan. A jól szerkesztettség általunk talált kritériumainak jelentős része ontológiafüggetlennek bizonyult – cikkünkben ezeknek az ontológiafüggetlen strukturális követelményeknek, illetve elveknek az összefoglalására teszünk kísérletet, az ImportNET projekthez kapcsolódó példákon mutat be gyakorlati alkalmazásukat.

[1] Az ImportNET projekt az Európai Bizottság támogatásával, a 6. Keretprogramom belül valósult meg, az IST-2006-033610 számú szerződés alapján.

2. Rétegzés és modularizáció

Az általunk talált egyik legfontosabb ontológiaszerkesztési elv a *rétegzés* elve: a reprezentálandó tudást célszerű az általánosság foka szerint rétegekre osztani. Az ontológia minden osztálya és relációja eleme egy és csak egy rétegnek, és a kevésbé általános rétegekhez tartozó osztályok részosztályai az általánosabb rétegek osztályainak. Mivel a specifikusabb rétegek többnyire komplexebbek, és több információt tartalmaznak az általánosabbaknál, ezért egy bizonyos általánossági szint alatt a rétegeket koordinált *modulokra* célszerű osztani. A programmodulokhoz hasonlóan az ontológiamodulok olyan ontológiarészek, melyek elemei között sok kapcsolat található, és melyeknek viszonylag kevés kapcsolata van a modulon kívüli elemekkel.

Az ontológiák szokásos kétdimenziós ábrázolására (az egyes osztályok és relációk részosztályaik, illetve részrelációik fölött helyezkednek el) támaszkodva azt mondhatjuk, hogy a rétegzés az ontológia vertikális, a modularizáció pedig horizontális felosztásának felel meg (lásd az 1. ábrát).



1. ábra. Egy komplex ontológia rétegzése és modularizációja

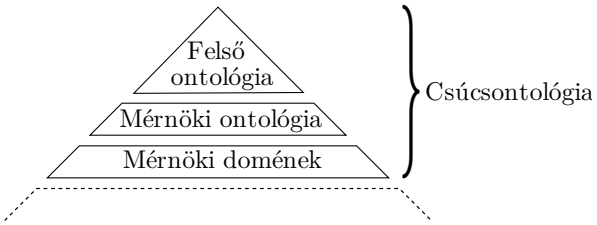
Egy átfogó ontológiában a legfontosabb vertikális tagolás a csúcsontológia (top ontology) elkülönítése. Bár a ,,csúcsontológia" terminust gyakran használják a legáltalánosabb, doménfüggetlen fogalmakat tartalmazó *felső ontológia* (upper ontology) értelemben, a mi szóhasználatunkban egy átfogó ontológia csúcsontológiai rétege az a szegmens, amely meghatározza a teljes ontológia alapszerkezetét azáltal, hogy rögzíti a *relációk* modellezésének módját. Ebből adódóan a csúcsréteg az ontológia összes relációját tartalmazza: az ontológia megmaradó része új elemként kizárólag osztályokat és individuumokat vezethet be.

Mivel a modellezési kérdések már a csúcsontológiai rétegben eldőlnek, ezért ontológiaszak-értői munkát csak ennek a rétegnek a kidolgozása igényel. Az ontológia további része többé-kevésbé mechanikus ,,T-box benépesítéssel" tölthető fel, pl. létező taxonómiák importálásával, vagy a doménszakértők által könnyen kezelhető, a feltöltést segítő felhasználói felületen keresztül.

3. Komplex ontológiák tagolása

A rétegzést egy több tudásterületet (domént) átfogó, komplex ontológiára alkalmazva olyan vertikálisan tagolt ontológiához jutunk, melynek csúcsontológiai része a következő rétegekből áll:

- a legátfogóbb, doménfüggetlen osztályokat és relációkat tartalmazó felső ontológia,
- a leírt doménokra együttesen alkalmazható, de nem doménfüggetlen osztályokat és relációkat tartalmazó réteg, végül pedig
- egy réteg, mely doménspecifikus tudást tartalmaz (az ImportNET ontológiában ez a réteg többek között mechanikai és elektronikai tudást fed le).



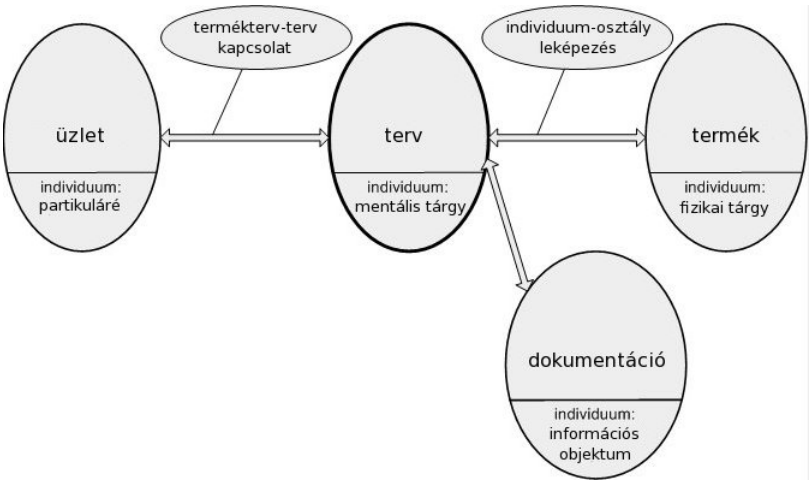
2. ábra. Egy vertikálisan tagolt, komplex ontológia felső rétegei

3.1. A felső ontológia

Mivel a közismert felső ontológiák, pl. a DOLCE [10], a SUMO [11] és a BFO [5] erősen különböző módon reprezentálják a legalapvetőbb relációkat (pl. a téridő viszonyokat), a felső ontológia megválasztása meghatározza a csúcsontológia további rétegeinek szerkezetét is. Ennek ellenére, az ImportNET projekt mérnöki ontológiájának fejlesztése során néhány olyan modellezési problémával is szembesültünk, melyek függetlenek ezektől a különbségektől.

A legfontosabb ilyen kérdések egyike az volt, hogy miként célszerű reprezentálni a mérnöki terveket, a tervek alapján készülő konkrét termékeket, valamint a köztük fennálló viszonyt. Mivel egy felső ontológia elképzelhetetlen egy, a konkrét fizikai tárgyakat tartalmazó osztály nélkül, ezért a konkrét termékek kategorizációja viszonylag könnyű feladat: pl. a DOLCE felső ontológiában ezek a PHYSICAL-OBJECT osztály példányainak tekinthetők. A *tervek* kategóriájának meghatározása már jóval nehezebb feladat. Habár a mérnökök rajzok és írott (papíron vagy elektronikus formában tárolt) dokumentumok segítségével reprezentálják terveiket, azok nem azonosak konkrét fizikai reprezentációikkal — pontosan azért, mert az utóbbiak csupán reprezentálják őket. A tervek kategorizációs problémájának két legfontosabb megközelítését ,,realista" és ,,konstruktivista" megközelítésnek nevezhetjük.

Az első megközelítés a terveket téren és időn kívüli absztrakt objektumoknak tekinti, melyek ontológiai státusza hasonló ahhoz, melyet a matematikai platonisták tulajdonítanak a matematika tárgyainak: az ember nem létrehozza, csupán felfedez(het)i őket. A DOLCE és a SUMO esetében ez a megoldás a terveket az ABSTRACT osztály példányainak tekintené. A realista megközelítéssel ellentétben a konstruktivista felfogás a terveket *mentális objektum*okként kezeli, melyek ez emberi elme tevékenységének eredményei. Ennek megfelelően a konstruktivista felfogás szerint minden terv csak egy adott időponttól kezdve létezik. A DOLCE tartalmaz egy MENTAL-OBJECT osztályt, más felső ontológiák azonban csak közvetett eszközökkel rendelkeznek a mentális objektumok leírásához. A SUMO-ban pl. található egy INTENTIONAL-PROCESS osztály, melyhez egy műszaki cikk megtervezésének *folyamata* tartozik, és a tervek maguk olyan dolgokként jellemezhetőek, melyek résztvevői egy tervezési folyamatnak (vö. [7]).



3. ábra. Egy modularizált mérnöki ontológia

3.2. Modularizáció

A 3. ábra egy mérnöki ontológia egy természetesnek tűnő modularizációját mutatja, mely a következő egységekre bontja a műszaki réteget:

– **Üzleti ontológia.** Ez a modul a kollaborációkkal kapcsolatos üzleti-gazdasági tudást fedi le, így pl. tartalmazza azokat az osztályokat és relációkat, melyek az együttműködésben résztvevő vállalatokra, illetve dolgozóikra vonatkozó információ reprezentációjához szükségesek, különös tekintettel az együttműködésben betöltött szerepükre (pl. ki a kollaborációs projekt vezetője stb.). A termékmenedzsmenttel kapcsolatos tudást szintén ez a modul reprezentálja.

- **Tervezési ontológia.** A tervezési ontológia indiviuumtartománya kizárólag mérnöki tervekből áll, vagyis olyan objektumokból, amelyek a jövőben gyártásra kerülő konkrét termékek tulajdonságait reprezentálják. A modul osztályhierarchiájának jelentős része izomorf a termékek osztályainak hierarchiájával (lásd a következő részt).
- **Dokumentációs ontológia.** A dokumentációs ontológia azokat az információs objektumokat reprezentálja, melyek a kollaboráció során jönnek létre, pl. írott terveket, tervrajzokat, műszaki dokumentációt stb. Ezek az információs objektumok nem keverendőek össze konkrét fizikai megvalósulásaikkal: egy tervrajznak (ami egy információs objektum) sok különböző fizikai példánya, másolata létezhet.
- **Gyártási ontológia.** Ha a kollaboráció sikeres volt, akkor az elkészült tervek alapján legyárthatóak a konkrét termékek. A gyártási ontológia indivuumai fizikai tárgyak: az elkészülő műszaki cikkek és részeik.

A modularizációval szemben támasztott követelményünkkel összhangban a különböző modulok elemei között viszonylag kevés a kapcsolódás.

Az üzleti és a tervezési modul között egyetlen fontos kapcsolat áll fent: bizonyos tervek *terméktervvé* válnak, vagyis az általuk leírt tárgyakat gyártják és forgalomba hozzák. A *termékterv* fogalom tipikus szerepfogalom (abban az értelemben, ahogyan ezt a metatulajdonságot az OntoClean [6] metodológia használja), mivel akkor alkalmazható egy individuumra, ha az részt vesz egy *kontingens* üzleti folyamatban (v.ö. [6, 16]). Ebből adódóan a terméktervek osztályát nem célszerű egyszerűen a TERV osztály részének tekinteni — előnyösebb megoldás a tervek üzleti szerepeinek reifikációja, mely esetben a szerepeket kizárólag az üzleti ontológia individuumtartományában szükséges szerepeltetni.

Terjedelmi okok miatt nem térhetünk ki az üzleti és a dokumentációs modul között fennálló, igen komplex kapcsolatrendszerre, de a tervek és termékek közötti viszony olyan kiemelkedő fontosságú, hogy mindenképpen szólnunk kell róla röviden.

3.3. Tervek és termékek

A tervek és termékek viszonyával kapcsolatos reprezentációs nehézségek a következő feszültségből adódnak: Egyfelől, a tervek lényegileg különböznek a szerintük legyártott termékektől, mivel tulajdonságaik túlnyomó része különbözik (pl. egy számítógép terve maga nem számítógép). Természetesen van kapcsolat egy számítógépterv és a ,,számítógépnek lenni" tulajdonság között: a terv valamiképpen reprezentálja, illetve *kódolja* a tulajdonságot, és minden, a tervet megvalósító tárgy exemplifikálja azt. Másfelől, a tervezőmérnökök gyakran kezelik úgy a terveiket, mintha azok rendelkeznének az általuk kódolt tulajdonságokkal — ez a gyakorlat különösen hasznos akkor, amikor tervekkel kapcsolatos következtetéseket kell végezni. Pl. természetesnek tűnik az a következtetés, hogy ha minden számítógép tartalmaz egy processzort, akkor hiányosak azok a számítógéptervek, melyekből ,,hiányzik a processzor."

A tervek és az őket megvalósító termékek közti viszony most vázolt két oldala két egymást kiegészítő követelményhez vezet a viszony formális reprezentációjára nézve:

- A reprezentáció nem feltételezheti, hogy a tervek és az őket megvalósító fizikai tárgyak általában ugyanazon osztályok példányai.
- Ennek ellenére, tükröznie kell azt a tényt, hogy szoros kapcsolat áll fent a tervek és a termékek tulajdonságai között, amely a következőképpen jellemezhető:
  - Minden t tervre van olyan φ osztály, melynek példányai azok a fizikai tárgyak, melyek *megvalósítják* a tervet:

$$\forall t \exists \varphi \forall a(\varphi(a) \equiv \mathcal{M}(a,t)) \tag{1}$$

  - Létezik egy $\mathcal{K}$ *kódolás* reláció a tervek és a termékosztályok között, amely a következő tulajdonságokkal bír:
    - Ha egy terv kódol egy tulajdonságot, akkor minden, a szóbanforgó terv szerint gyártott termék rendelkezik az adott tulajdonsággal:

$$\forall t \forall \varphi(\mathcal{K}(t,\varphi) \rightarrow \forall a(\mathcal{M}(a,t) \rightarrow \varphi(a))). \tag{2}$$

    - Ha a φ-t kódoló tervek osztálya részosztálya a ψ-t kódoló tervek osztályának, akkor φ részosztálya ψ-nek:

$$\forall \varphi \forall \psi(\forall t(\mathcal{K}(t,\varphi) \rightarrow \mathcal{K}(t,\psi)) \rightarrow \forall a(\varphi(a) \rightarrow \psi(a))). \tag{3}$$

Sajnos a fenti formális jellemzés nem fejezhető ki közvetlenül deskripciós logikai (DL) nyelveken a standard DL-szemantika segítségével, mivel osztályok fölött kvantifikál, és szerepel benne a másodrendű $\mathcal{K}$ reláció. Ebből adódóan, ha a mérnöki ontológiát egy DL-formalizmusra támaszkodva kívánjuk reprezentálni, akkor a tervek és termékek közti viszonyt vagy egy nemstandard DL-szemantika használatával, vagy a használt DL nyelven kívüleső eszközökkel fejezhetjük ki. A következőkben a probléma három lehetséges megközelítését tárgyaljuk röviden: külön termék- és tervontológia használatát egy köztük megadott leképezéssel (ontology mapping), egyetlen ontológia használatát metaszabályokkal, és végül a DOLCE Descriptions and Situations kiterjesztésének [3,4] alkalmazását.

Ontológialeképezés. A leképezésalapú megközelítés külön tervontológia és gyártási ontológia létrehozását igényli, melyek terméktulajdonságokra utaló közös osztályneveket tartalmaznak, pl. ‘CPU’, ‘32BIT_CPU’ stb. A közös nevek szemantikája különböző, de közel álló: Ha egy F osztálynév az $\{x : \varphi(x)\}$ termékosztályra referál a gyártási ontológiában, akkor az $\{x : \mathcal{K}(x,\varphi)\}$ tervosztályra referál a tervontológiában, vagyis azon tervek osztályára, amelyek *kódolják* az F által kifejezett tulajdonságot. Például míg a ‘CPU(a_{17})’ formula interpretációja a gyártási ontológiában az lehet, hogy ‘a_{17}’ referenciája egy konkrét központi processzor, addig a tervontológiában ugyanez a formula azt jelenheti, hogy ‘a_{17}’ referenciája egy központi processzor *terve*. A két ontológia közti $\mathcal{L}$ leképezésnek a következő tulajdonságokkal kell rendelkeznie:

– Összhangban az (1) megkötéssel a tervontológia individuumait (vagyis a terveket) a gyártási ontológia azon osztályaira képezi le, melyek az adott terv alapján legyártott termékeket tartalmazzák. Például a tervontológia 'a_{17}' nevű terve a gyártási ontológia 'INTEL80486DX' nevű osztályára lehet leképezve. Habár az $\mathcal{M}$ megvalósítás reláció nem fejezhető ki a két ontológiában, a leképezés segítségével könnyen definiálható: egy a termék pontosan akkor a megvalósítása egy t tervnek, ha a példánya a $\mathcal{L}(t)$ osztálynak.
– A (2) megkötést követve a tervontológia A-box formuláit a gyártási ontológia bizonyos T-box formuláira képezi le: az '$F(t)$' alakú állításokhoz, ahol t egy tervre utal, F pedig a $\{x : \mathcal{K}(x, \varphi)\}$ osztályra, az '$F \sqsupseteq \mathcal{L}(t)$' állításra képezi le. Például a tervontológia '32BIT_CPU(a_{17})' állításának képe a gyártási ontológia '32BIT_CPU $\sqsupseteq$ INTEL80486DX' állítása lesz.
– Végezetül, a (3) megkötésnek megfelelően ha F az $\{x : \mathcal{K}(x, \varphi)\}$ osztályt jelöli a tervontológiában, és az $\{x : \varphi(x)\}$ osztályt a gyártási ontológiában, és egyúttal G az $\{x : \mathcal{K}(x, \psi)\}$ osztályt jelöli a tervontológiában és az $\{x : \psi(x)\}$ osztályt gyártási ontológiában, akkor a tervontológia '$F \sqsubseteq G$'formulájának a leképezés szerinti képe sajátmaga.

A leképezésen alapuló megoldás legfontosabb előnye az, hogy (a lehetőségekhez mérten) megfelel a tervezőmérnöki szemléletnek, és követi a tervezők nyelvi gyakorlatát. E szerint a megközelítés szerint a CPU-hoz hasonló predikátumok tervekre és termékekre egyaránt alkalmazhatóak, de ,,szisztematikusan többértelműek": míg termékekre alkalmazva azt állítják, hogy a kérdéses termék rendelkezik egy tulajdonsággal, addig egy tervről azt mondják, hogy *kódolja* a szóban forgó tulajdonságot.

Egyetlen ontológia metaszabályokkal. A második megközelítés egyetlen ontológiában reprezentálja mind a terveket, mind a termékeket, és az (1), (2) és (3) megkötéseket részben az ontológia metanyelvén fejezi ki. Fontos előnye ennek a megoldásnak, hogy az $\mathcal{M}$ megvalósítási reláció az ontológia nyelvén reprezentálható, és ebből adódóan egy t tervet megvalósító termékek osztálya egyszerűen definiálható a [MEGVALÓSÍTJA : t] osztályként. A leképezésen alapuló megoldással szemben egy tulajdonság példányainak osztálya és az ugyanezen tulajdonságot kódoló tervek osztálya nem kaphat azonos nevet, de az összetartozó nevek összekapcsolhatók egy megfelelően választott elnevezési séma segítségével, pl. kiköthető, hogy ha F az $\{x : \varphi(x)\}$ osztályt jelöli, akkor az 'F_KÓDOLÓJA' osztálynév az $\{x : \mathcal{K}(x, \varphi)\}$ osztályt jelölje. Ezt az elnevezési sémát használva a szükséges metaszabályok a következőképpen fogalmazhatók meg:

– Ha az ontológia tartalmaz egy 'F_KÓDOLÓJA' alakú osztálynevet, akkor tartalmaz egy F nevű osztályt is.
– Ha az ontológia tartalmaz egy 'F_KÓDOLÓJA(t)' alakú állítást, akkor tartalmaz egy '[MEGVALÓSÍTJA : t] $\sqsubseteq F$' alakú állítást is.
– Ha az ontológia tartalmaz egy 'F_KÓDOLÓJA $\sqsubseteq G$_KÓDOLÓJA' alakú állítást, akkor tartalmazza az '$F \sqsubseteq G$' állítást is.

Descriptions and Situations. Az utolsó megközelítés, amelyet röviden meg szeretnénk említeni, a DOLCE felső ontológia Desriptions and Situations (röviden DnS) kiterjesztésének segítségével reprezentálja a tervek és termékek kapcsolatát. A DnS-t sikeresen használtuk az ImportNET projektben a cselekvési tervek formális reprezentációjára [1], és olyan osztályhierarchiával rendelkezik, amelybe a tervek egyszerűen elhelyezhetőek, mivel a SYSTEM-DESIGN osztály példányainak tekinthetők. Ennek ellenére ez a megoldás meglehetősen problematikus.

Az egyik nehézség az, hogy a megvalósítás reláció egyetlen szóba jöhető reprezentánsa a DnS rendszerben a SATISFIES, amelynek az értelmezési tartománya a SITUATION osztály, amely viszont részosztálya a NON-PHYSICAL-OBJECT osztálynak. Ebből adódóan a DnS keretei között a termékek csak nemfizikai individuumoknak tekinthetők. Bár fontos filozófiai érvek szólnak amellett, hogy a termékeket szociális konstrukciónak, és ne fizikai tárgynak tekintsük, (lásd pl. [12]), ez a megkülönböztetés idegen a mérnöki szemlélettől, és nyilvánvaló előnyök nélkül növeli a reprezentáció bonyolultságát. A módszer egy másik hiányossága, hogy nem teszi lehetővé a tervkomponensek egymáshoz való viszonyának a mérnöki szemléletnek megfelelő reprezentációját. Végezetül megjegyzendő, hogy az előző szakaszban tárgyalt megoldáshoz hasonlóan a DnS-alapú megközelítés egyetlen ontológiában reprezentálja a terveket és a termékeket, és ezért alkalmazása esetén a megvalósítás reláció fontos jellemzői csak metaszabályokkal vagy egyéb, az ontológia nyelvén kívül eső eszközökkel fejezhető ki.

3.4. Reprezentációs mélység

A ‘mérnöki ontológia’ kifejezés többértelmű. Még ha rögzítjük is a domént (pl. az elektronika területét), a reprezentáció *mélysége* nyitott kérdés marad: nem lesz tisztázott, hogy a tervezési folyamat mely fázisait támogatja az ontológia. A következő reprezentációs szinteket különböztethetjük meg:

1. PDM (termékadat-kezelés) szint: a tervezett tárgyak komponenseinek tulajdonságait és mereológiai viszonyait ábrázolja az ontológia.
2. Topológiai szint: a komponensek topológiai kapcsolatait szintén reprezentálja az ontológia, de a pontos geometriai részletek nélkül (pl. csak áramkör diagramokat ad meg).
3. Geometriai szint: a komponensek elhelyezkedése és mérete is ábrázolásra kerül.
4. Működési szint: az ontológia a tervezett tárgy működését is reprezentálja, esetleg a működés helyessége is ellenőrizhető a segítségével (helyesek-e az elvégzett számítások, a méretezés stb.).

A szintek fenti sorrendje egyúttal bonyolultsági, összetettségi sorrend is: a lejjebb elhelyezkedő szintek a magasabban lévők fogalmi eszközeit is felhasználják. A négy szint közül a tervek PDM-szintű reprezentációja jól ismert, és viszonylag egyszerű eszközökkel elvégezhető, mivel csak

– a tervek atomi és nematomi komponenseinek megkülönböztetését, valamint
– a RÉSZE reláció és a
– a tervkomponensek tulajdonságainak reprezentációját igényli (ez utóbbi megtehető pl. a DOLCE minőségeket repreztentáló mechanizmusának OWL-DL implementációjára támaszkodva [10, 16]).

A PDM-szinttel szemben a többi szint reprezentációja komoly kihívást jelentő feladat. A topológiai szint speciális relációinak reprezentációja minden bizonnyal e kapcsolatok reifikációjával oldható csak meg, a geometriai tulajdonságok leírását pedig nagyon megnehezíti a hely fogalmának relativitása [2]. Végezetül, egy a működési szintet is reprezentáló ontológia kifejlesztése valószínűleg egy kvalitatív fizikai elmélet formalizációját is szükségessé teszi, ami (az esetleges kvantitatív adatokat is figyelembe véve) igen komoly nehézségekbe ütközhet egy DL-alapú környezetben.

4. Összefoglalás

Cikkünkben olyan ontológiafejlesztési elveket javasoltunk, melyek segítségével jól strukturált mérnöki ontológiák hozhatók létre. Véleményünk szerint kifejleszthetőek olyan mérnöki ontológiák, melyek eleget tesznek ezeknek a módszertani elveknek, és egyúttal jól modellezik a gyakorló mérnökök szemléletmódját, fogalomrendszerét.

Két, az átfogó ontológiák szerkesztése során alkalmazható szerkezeti alapelvet ismertettünk: a rétegzés, vagyis az általánosság foka szerinti vertikális szegmentáció elvét és a modularizáció, vagyis a viszonylag kevés külső kapcsolattal rendelkező horizontális szegmensekre, modulokra bontás elvét.

Egy mérnöki ontológia fontos további dimenziója, hogy milyen mélységig képes reprezentálni műszaki terveket, illetve tervezési folyamatokat. Amellett érveltünk, hogy a tervezett tárgyak mereológiai szerkezetét és komponenseik tulajdonságait viszonylag egyszerű reprezentálni, ugyanez azonban távolról sem mondható el a komponensek topológiai, geometriai és működési viszonyairól, mivel az utóbbi három terület formális reprezentációja komoly kihívást jelentő feladat, különösen DL-alapú ontológiai nyelvek használata esetén.

Hivatkozások

1. Damjanovic, V., Behrendt, W., Plössnig, M., Holzapfel, M.: Developing Ontologies for Collaborative Engineering in Mechatronics. In: Proceedings of the 4th European Semantic Web Conference, Innsbruck (2007)
2. Donnelly, M.: Relative Places. Applied Ontology **1** (2005) 55–75
3. Gangemi, A., Mika, P.: Understanding the Semantic Web through Descriptions and Situations. In: Meersman, R. (ed.): Proceedings of ODBASE'03 Conference, Springer (2003)
4. Gangemi, A., Borgo, S., Catenacci, C., Lehmann, J.: Task Taxonomies for Knowledge Content. Deliverable D07 of the METOKIS Project (2005)

5. Grenon, P.: BFO in a Nutshell: A Bi-categorial Axiomatization of BFO and Comparison with DOLCE. Technical report. Ifomis (2003)
6. Guarino, N., Welty, C.: An Overview of OntoClean. In: Handbook on Ontologies. Springer (2004) 151–159
7. Hung, L.C., Beng, L.H., Wah, N.G., Yin, H.K.: Plan Ontology and its Applications. In: 7th Int. Conference on Information Fusion (2004)
8. Mahl, A., Semenenko, A., Ovtcharova, J.: Virtual Organisation In Cross Domain Engineering. In: Establishing The Foundation Of Collaborative Networks. Springer (2007) 601-608
9. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: WonderWeb Deliverable D18: Ontology Library. Technical report. Laboratory for Applied Ontology (2003)
10. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: WonderWeb Deliverable D18: Ontology Library. Technical report. Laboratory for Applied Ontology (2003)
11. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: FOIS '01: Proceedings of the International Conference on Formal Ontology in Information Systems. New York, ACM (2001) 2–9
12. Vieu, L., Borgo, S., Masolo, C.: Artefacts and Roles: Modelling Strategies in a Multiplicative Ontology. In: Proceedings of FOIS 2008 (2008)

Online helyesírási szótár és megvalósítási nehézségei

Pintér Tibor[1], Mártonfi Attila[1], Oravecz Csaba[1]

[1] MTA Nyelvtudományi Intézet, Benczúr utca 33.,
1068 Budapest, Magyarország
{tpinter, martonfi.attila, oravecz}@nytud.hu

Kivonat: A magyar társadalom helyesírás és nyelvhelyesség iránti igénye mármár szakmai közhelynek számít. A helyesírás számítógépes modellezésének eddigi gyakorlata azt mutatja, hogy egy online helyesírási szótár, nyelvi tanácsadó szolgáltatás triviálisan nem oldható meg csupán gépi erőforrással, például egy nyelvtan mögött álló szótárral. A helyes alak felismeréséhez mindenképpen szükség van morfológiai elemzőre, illetve az elemzés kimeneteként keletkező homonimák egyértelműsítésekor bizonyos mértékben a kérdező interaktivitására is. A morfológiai elemzést segíti a főként szemantikai szempontok alapján szerkesztett szótár, amelyben az egyes lexikai tételek több szempontból annotálva vannak (ehhez a szótárat különféle szemantikai kategóriák alapján egyértelműsítettük, valamint az interakciót elősegítendő, egyszerű mondatokkal rávezetjük a kérdezőt az adódó lehetőségek közti választásra). Sok esetben a morfológiai elemző és a szótár önmagában nem elegendő a helyes alak kiválasztásához, így némely esetben a lokális szintaktikai környezet elemzését is fel kell vállalnunk. Az online helyesírási tanácsadó rendszer erősen formális felépítésű. Hatékony működése érdekében teljesen új – formális rendszert követő – alapokon kell leírnunk a helyesírás számos részrendszerét.

1 Bevezetés

A magyar nyelvre alkalmazott nyelvtechnológiai kutatások mostohán kezelik a helyesírási relevanciájú internetes segédeszközöket. Bár a hibátlan, „helyes" írás megmozgatja a művelt magyar társadalmat, ezekben a kérdésekben leginkább az e-mailes és telefonos segítség, illetve a különféle fórumok által közvetített ember-ember interakció az, amit a nyelvhasználók leginkább igénybe vesznek. Ennek oka nem elsősorban a megfelelő nyelvtechnológiai eszköz hiánya (általában morfológiai elemzővel kiegészített, szótári keresésen alapuló eszközök vannak forgalomban; MorphoLogic: Helyes-e?; Németh László: Hunspell, Szabad magyar szótár), hanem a magyar helyesírásnak az a tulajdonsága, hogy bizonyos pontokon a szabályalkalmazók anyanyelvi kompetenciájára és szövegértelmezésére hivatkozik, illetve számos, a szabályrendszernek ellentmondó íráshagyományt is továbbörökít. E miatt az összetett függés miatt valószínűtlennek tartjuk egy olyan program kifejlesztését, amely emberi segítség (felhasználói interaktivitás) nélkül képes lenne hatékonyan kezelni a magyar helyesírás minden pontját (vö. [1, 2]).

Az MTA Nyelvtudományi Intézete éppen ezért olyan portál elkészítésén dolgozik, amely megszüntetné a fent említett űrt: egy pontos és gyors, mindenki által elérhető, azonnal segítséget nyújtó internetes nyelvi tanácsadó portál, a helyesiras.hu megalkotásán. A rendszer működőképessége három alappilléren, 1. egy robusztus, többrétegű, annotált szótáron, 2. pontos, formális nyelvtanon és 3. a kérdező interaktivitásán alapszik (ez utóbbira a helyesírás egyes részeinek erőteljes szemantikai beágyazottsága, az ún. értelemtükröztetés miatt van szükség). A már működő internetes helyesírási segédletekhez képest a most készülő rendszer nagyobb fedésű és remélhetőleg jóval megbízhatóbb és pontosabb lesz, nem pusztán egy helyesírási szótár szolgai számítógépes másolata. A pontosság mellett egyéb olyan tulajdonságai is lesznek, amelyek reményeink szerint nem csak a helyesírási alapismeretekkel rendelkezőket és nem csak a magyarországi nyelvhasználókat ösztönzik majd a portál használatára. A helyesiras.hu számos újítása miatt új felhasználói irányban is nyit.

2 A nyelvtan

2.1 Milyen nyelvtanra van szükség?

Az előmunkálatok folyamán nyilvánvalóvá vált, hogy a helyesírási problémák nagy része lefedhető szótárral, vagy megoldható egyszerű grammatikával. A valódi kihívást ezért csupán a magyar helyesírás bizonyos pontjai jelentik (ám önmagukban ezek megoldása jelentős munkával jár). A magyar helyesírás létező számítógépes modelljei azt mutatják, hogy hatékony helyesírási tanácsadás nem valósítható meg csupán gépi erőforrással és a nyelvtan mögött álló szótárral (még több százezer szavas háttérkorpusz esetén sem). Az egyszerű szójegyzéken alapuló tanácsadás (ezt csinálják az interneten jelenleg elérhető helyesírási tanácsadók) csak akkor ad kielégítő eredményt, ha a beírt (lekérdezett) szó eleve helyesen van írva, valamint megtalálható a rendszer mögött álló szótárban (illetve jobb esetben a mögöttes nyelvtan össze tudja rakni). A helyesen írt, ugyanakkor nem ismert szavakat az ilyen elemzők hibás írásmódúként adják vissza, vagyis nem nyújtanak többet egy átlagos, szabályzattal nem rendelkező papírszótárnál. Pontosabban lényegesen kevesebbet nyújtanak, ugyanis egy papírszótár készítője az anyag elrendezésével (tehát a keresett elem betűrendi és szócikkbeli környezetével) tekintélyes mértékű információt tud adni a szótárt lapozgató felhasználónak, hiszen ezen a módon interakcióba tud lépni a szótárhasználó anyanyelvi intuíciójával, egyéb ismereteivel és kognitív működésével.

Az általunk fejlesztett rendszerben a kérdező által beírt szót vagy többtagú kifejezést a webfelület mögött működtetett elemző értelmezi, megpróbálja azonosítani a lehetséges helyesírási problémakört, majd megválaszolni, illetve jóváhagyni a helyes alakot. A keresett alak felismeréséhez mindenképpen szükség van morfológiai elemzésre (pl. a különféle, különösen az *-ó/-ő* képzős igenevek felismerése, az alkotó tagokban szereplő tőmorfémák számlálása). A nyelvtan és a szótár együttes használata sem jelent azonban minden esetben megoldást, hiszen például a keresés kimenetén megjelenő homonimák egyértelműsítése bizonyos mértékben már a kérdező interak-

tivitását igényli. A helyesírásukban eltérő, kiejtésükben (vagy legalábbis a szegmentális hangszerkezetben) azonos, tehát homofón alakpárok, -többesek esetében számos alakváltozat helyes lehet (pl. *klónozottkukorica-termesztő* 'klónozott kukoricát termesztő személy' – *klónozott kukoricatermesztő* 'olyan kukoricatermesztő, akit klónoztak', *adalékanyag* 'az adalék anyaga' – *adalék anyag* 'adalékul használt anyag', *csuklósbusz-vezető* 'csuklós busz vezetésére alkalmazott gépkocsivezető' – *csuklós buszvezető* 'csuklásra hajlamos autóbusz-vezető'), mivel azonban az éppen keresett alak azonosítása magas szintű, tág szövegkörnyezetre támaszkodó nyelvi elemzést igényelne, és a tanácsot kérő csak egy szót vagy szókapcsolatot ad meg, a tanácsadó a megfelelő alak kiválasztása érdekében ilyen esetekben kénytelen az elemzési folyamatba bevonni a kérdezőt is.

Milyen morfológiai elemzésekre is van a helyesírás szempontjából szükség? A bemeneti karaktersorozaton végrehajtandó elsődleges elemzés a tőmorfémákra bontás (mivel a helyesírásban használt ÖSSZETÉTELI TAG fogalom valójában ennek a nyelvtani kategóriának felel meg) – nem mindegy például, hogy az elemző hogyan szegmentálja például a következő szavakat: *rendszer (= rend+szer), valószínűség (= va[ló]+szín[űség]); szemöldök* (képzett alak, nem összetétel), hiszen a helyes szegmentálás képezi a magyar helyesírás különírás-egybeírási részrendszerében a szótagszámlálás szabályának egyik bemenetét (*valószínűség-számítás* és nem **valószínűségszámítás*, mivel a *valószínűség* összetett szóalak, így megvan a 3 tőmorféma és a 7 szótag). Ugyancsak a különírás és egybeírás kategóriájához tartozik a toldalékmorfémák pontos szegmentálása és típusok szerinti elkülönítése (a fenti szótagszámba beleszámítanak a képzők, de a jelek, ragok nem), ez azonban teljes mértékben gépesíthető.

A program kezeli továbbá többek között a különféle, hagyományokon alapuló külön- és egybeírást. Rendszerszerű hagyomány szerinti írásúnak tekinthetők például az anyagnevek, a színnévi jelzős összetételek vagy a számnévi jelzős, *-s, -i, -ú/-ű/ -jú/-jű, -nyi, -nként, -nta* toldalékra végződő alakulatok. Ha a jelzői szerepű szó és az alaptag egyszerű szó, akkor egybe kell őket írni (1+1=1), s ezt a program követi is. Ha valamelyik tag önmagában is összetett szó, akkor már különírandók (2+1|1+2=2): *selyemköntös ~ nyersselyem köntös*, *ötéves ~ öt hónapos*, *kétévnyi ~ tizenkét évnyi*, *kéthavonta ~ tizenkét havonta*. Hasonló algoritmus mozgatja az anyagnévi mozgószabályt is, ahol a különírt szószerkezet anyagnévi jelzőként szerepel: *valódi bőr,* de: *valódibőr kabát*; *fehér márvány,* de: *fehérmárvány vízcsap; tömör arany,* de: *tömörarany nyaklánc*. A fenti helyes írásmódok kialakításához arra is szükség van, hogy a program meghatározza az egyes alkotótagok közötti szintaktikai függéseket, valamint felismerje az ANYAGNÉV szemantikai kategóriát. Ez utóbbiban kapnak szerepet az annotált szótárak.

A magyar szavak külön- és egybeírása a felhasználó számára is meglehetősen bonyolult, egy helyesírási tanácsadó számára is szinte megoldhatatlan, bár részlegesen nyelvtannal és szótárral jól kezelhető. (A gépi választ nem eredményező esetekben, illetve azokban, amelyek során a kérdező nem elégedett a válasszal, a rendszer felkínálja a humán tanácsadói segítség igénybevételének lehetőségét.) A morfológiai elemzőknek általában alapvető problémájuk, hogy az elemzést két szóköz között hajtják végre, így csak a hibás egybeírást képesek észrevenni, a különírást viszont nem, vagy csak korlátozott mértékben (l. pl. a Helyesek „zöld aláhúzása"). A helyesiras.hu a részletesen annotált szótárak segítségével hatékonyan (bár nem teljes körű-

en) kezeli a magyar külön és egybeírás szemantikai jellegű komponenseit is. A szótárakkal és visszakérdező modullal kiegészített rendszer képes szemantikailag is különbséget tenni (és így a kérdezett alakot helyesen visszaadni) például az *-ó/-ő* képzős melléknévi igeneves szerkezetek vagy az összetett főnevek külön- és egybeírásának kérdésében (*csomagoló papír* 'olyan papír, amely éppen csomagol' – *csomagolópapír* 'csomagolásra készített papír', *napra forgó* 'a nap hatására meg-megforduló' – *napraforgó* 'magjáért, olajáért tartott haszonnövény', *járólapos* 'járólappal rendelkező, azzal felszerelt' – *járó lapos* 'gyalogló kismellű', *vendégfogadó* 'vendégül látó személy, ill. panzió' – *vendég fogadó* 'vendégségbe jött bukméker', *tanulószoba* 'tanulás tevékenységére rendszeresített helyiség' ~ *tanuló szoba* 'olyan szoba, amely tanul').

A tőmorfémák számának megállapítására irányuló szegmentálás mellett a morfoszintaktikai komponensnek kezelnie kell a szófajokat is. Erre is elsősorban a külön- és egybeírás miatt van szükség, hiszen például a színnévi jelzős összetételek, bizonyos fokozó szerkezetek vagy akár az anyagnévi mozgószabály helyes kezeléséhez ez elengedhetetlen. Lássunk erre is pár példát: a fokozó szerepű melléknévi vagy főnévi etimonú szó (azaz fokozópartikula) mindig külön áll a rákövetkező melléknévtől, például: *borzasztó rossz, böszme nagy, csoda jó, jó nagy, kutya hideg, marha erős, szép kövér, tök hangos.* Ettől eltér a hasonlítást kifejező jelentéssűrítő összetételek írásmódja, például: *csodaszép* 'a csodához hasonlatosan szép', *hófehér* 'a hó színéhez hasonlóan fehér', *hollófekete* 'a holló színéhez hasonlóan fekete'.

A magyar helyesírás, illetve a mögötte álló grammatikai modell összetett volta miatt a nyelvtani modulnak ki kell egészülnie kivételszótárral. Ez az MTA Nyelvtudományi Intézetében évtizedek óta működő helyesírási tanácsadói munkatapasztalat, az ezeket rögzítő jegyzőkönyvek, illetve a helyesírási szabályzatok szerkesztésekor felhalmozott tudás alapján készült.

2.2 Morfológia mellett lokális szintaxis

Mint erre korábban utaltunk, sok esetben a morfológiai elemző és a szótár önmagában nem elegendő a helyes alak kiválasztásához, így némely esetben a lokális szintaktikai környezet elemzését is fel kell vállalnunk (pl. bizonyos bővítmények megléte kulcsként szolgálhat annak eldöntésében, hogy egy alakulat szókapcsolat vagy összetétel-e, pl. *takarítónő* 'foglalkozásszerűen helyiségeket tisztává tevő nő' – *takarító nő* 'olyan nő, aki helyiségeket éppen most tesz tisztává' – *sokat takarító nő* 'olyan nő, aki sokat takarít'). Elsősorban a homofon alakok egyértelműsítése érdekében ennek a kérdező segítségét igénybe kell vennie – rávezető kérdéseken keresztül.

3 A szótár

A legtöbb helyesírás-segítő szolgáltatás szótár alapján működik: ez elkerülhetetlen alap, önmagában azonban nem megoldás, mivel a végeredmény így számos hiányt, kívánnivalót hagy maga után. A pusztán szótáron alapuló megoldás hátránya, hogy a keresés kimenete csak azt adja meg, hogy a beírt szó (karaktersorozat) megvan-e az

adatbázisban: akkor sem fogunk pozitív eredményt kapni, ha olyan szót keresünk, amely helyesen van ugyan írva, de az adatbázis nem tartalmazza. A fentiek ismeretében a morfológiai elemző sem elég hatékony megoldás önmagában, gazdag és részletesen annotált szótárak nélkül nem képzelhető el jól működő helyesírás-elemző és tanácsadó rendszer. A helyesiras.hu morfológiai elemzőjét főként szemantikai szempontok alapján annotált részszótárak gyűjteménye segíti, amelyben az egyes lexikai tételek több szempontból is kódolva vannak (ehhez a szótárat különféle szemantikai kategóriák alapján egyértelműsítettük). A kiejtésben az írásképtől jelentősen eltérő szavak, nevek, mozaikszók esetében szükség van a szótárban kiejtésjelölésre is az elválasztás, a toldalékolás, illetve a névelőzés helyes meghatározásához.

3.1 Szótári erőforrások

A portál alapvető lexikális erőforrásait egyrészt a Magyar Nemzeti Szövegtár 187 millió szavas, kontextuális stílusok szerint tagolt korpusza, másrészt egy külön erre a célra összeállított több mint 400 millió szavas, címkézett gyűjtemény adja. Ez utóbbi több mint 4 millió elemzett szóalakot, közel 2 millió szótövet tartalmazó, műfaji kategóriákba sorolt gyakorisági adatbázis. Az adatbázishoz kapcsolódó lekérdező felület már működik, ezzel a szótárnak a kritikus helyesírási problémákat tartalmazó, jellemző szóalakok feletti fedése vizsgálható közvetlenül (1. ábra). Ezek mellett az alapvető források mellett a rendszert a felhasználói kérdésre adott pontos válasz megtalálásában egy több tízezer többtagú kifejezést tartalmazó szótár, valamint több, specifikus szemantikai jegyek alapján összeállított szólista támogatja (pl. csak kis- és nagybetűben vagy különírás-egybeírásban eltérő stb. minimális párok, anyagnevek, számnevek, jelzők, állatnevek, növénynevek, településnevek, magyar családnevek és kiejtésük, különböző szókapcsolatok listája [*-ó/-ő* képzős melléknévi igeneves szerkezetek, fn+fn, mn+fn], *a* végű szavak listája). Az aktuális problémának a számítógép számára érthető formális meghatározásában további segítséget nyújt egy mintegy 6000 rekordos adatbázis, amely a közönségszolgálati jegyzőkönyvekben található kérdés-válaszokat rendszerezi és osztályozza.

Az annotált részszótárak közül külön érdemes foglalkozni a minimális párokat, anyagneveket, melléknévi igeneves szerkezeteket stb. feldolgozó szótárakkal. A minimális párok szótára 1040 olyan párt tartalmaz, amelyek között egykarakternyi eltérés található (ez lehet akár kis- és nagybetű, illetve szóköz is).

1. táblázat: Mutatvány a minimális párok szótárából.

abba *(nm.)*	**abba-** *(ik.)*
Ábrahámhegy *(település)*	**Ábrahám-hegy** *(hegy)*
adalékanyag 'az adalék anyaga'	**adalék anyag** 'adalékul használt anyag'
adóvevő *(fn.)*	**adó-vevő**
afelé *(hsz.)*	**a felé** *(nm.)*
afelett *(hsz.)* ~ afölött	**a felett** *(nm.)* ~ a fölött
afelől *(hsz.)*	**a felől** *(nm.)*
Ag <ezüst>	**AG**

ági	**Ági**
ágrólszakadt ’nyomorult’	**ágról szakadt** ’olyan, ami leszakadt egy ágról’
ahelyett *(hsz.)*	**a helyett** *(nm.)*
akadémia ’főiskola’	**Akadémia** ’Magyar Tudományos Akadémia’
akár	**akár-**
akárcsak ’mint’ *(ksz.)*	**akár csak** ’akár csupán’
akárhogy ’bármilyen módon’	**akár hogy** *(kihagyásos szerkezetben)*

A minimális párok megfelelő kezelése elsősorban a visszakérdezés során oldható meg, mivel a két elem közti eltérések főként szemantikaiak, így a pontos alak kiválasztásában legfőként a kérdező tud segíteni interaktív kérdéseken keresztül (hiszen a kérdező szándékát közvetlenül nem ismerhetjük). A kérdező a helyesírás fogalmi rendszerében gyakran nem tudja artikulálni teljes pontossággal a kérdését (ha tudná, nem kérdezne), így a rávezető kérdéseknek olyan releváns és főképpen egyszerűen közölt információkat kell tartalmaznia, amelyek nyelvtani-helyesírási ismeretekre nem építenek, csupán a kérdező anyanyelvi kompetenciájára, és amelyekből a kérdező számára kiderül, pontosan melyik alakváltozatra is van szüksége (pl. *tanítónő – tanító nő*).

tanítónő	» éppen a cselekvést, tevékenységet végzi, esetleg folyamatot átéli, elszenvedi (nő, aki éppen most tanít)	» *tanító nő*
	» valamire rendeltetett, valamit általában, foglalkozásszerűen űz, nem vagy nem pusztán pillanatnyi cselekvést, tevékenységet végez, illetve folyamatot átél, elszenved (tanításra való nő)	*» tanítónő*
kávéscsésze	» valamit tartalmazó, valamivel szennyezett edény (kávét tartalmazó, kávéval szennyezett csésze)	» *kávés csésze*
	» valaminek a felszolgálására, fogyasztására használt, szokásosan meghatározott méretű és formájú edény (kávé felszolgálására, fogyasztására szolgáló csésze)	» *kávéscsésze*

Bár tudjuk, hogy a szemantikai információ megfelelő minőségű kezelésétől még távol vagyunk, nem kerülhetjük meg a szavak bizonyos jelentésjegyeinek beépítését. Erre alakítottuk ki az annotált szótárakat, amelyek a megfelelő nyelvtani szabályokkal kiegészítve hatékonyan kezelik a helyesírás azon pontjait, ahol a morfológiai-szintaktikai elveket kiegészítik a szemantikai kategóriák.

3.2 Feldolgozó modulok

A rendszer működését a helyesírás részrendszerei köré szervezett modulok vezérlik, amelyeket az alábbi attribútumok jellemeznek:

1. a modul feladata: a modul által kezelt jelenség leírása;

2. a modul működéséhez szükséges erőforrások és jellemzőik specifikációja (pl. milyen speciális szólista szükséges a kérdéses jelenség kezeléséhez);

3. a modulhoz rendelhető felhasználói kérdés géppel azonosítható jegyei, illetve ezek hiányában a felhasználótól bekérendő további információ meghatározása;

4. a modul működésének forgatókönyve: a modulok működését forgatókönyvek írják elő, amelyek megadják, hogy amennyiben az adott felhasználói lekérdezés a modulhoz rendelődik, milyen processzáló lépések szükségesek a válasz megadásához (pl. a lekérdezett alak szerepel-e a modulhoz rendelt lexikális erőforrásokban → igen → rendben; → nem → felhasználótól további információ, ennek alapján válasz generálása).

4 A további, speciálisabb részrendszerek kezelése

A szavak, egyszerűbb szókapcsolatok szótár és nyelvtan egységén alapuló kezelésének vázlatát mutattuk be az eddigiekben. Szükséges azonban szólni azokról a részrendszerekről, amelyeknek a működtetéséhez ezek a műveleti elemek nem nyújtanak elégendő támpontot. Ezek többnyire diffúzabb problematikát mutatnak, így a számítógépes kezelésük is nehezebben körülhatárolható, ugyanakkor alapvető jelentőséggel bír, hogy az MTA Nyelvtudományi Intézet közönségszolgálati jegyzőkönyveinek tanúsága szerint a felvetett kérdések túlnyomó többsége a különírás és egybeírás kérdéskörét érinti elsősorban. Mindazonáltal nem maradhatnak megválaszolatlanul az alábbi részrendszereket érintő kérdések sem.

4.1 Tulajdonnevek

A legnagyobb összetartozó problémakört a különféle tulajdonnevek jelentik. Noha ezt a kategóriát szófaji megnevezésként is szokás használni, számítógépes nyelvészeti értelemben nem érdemes szófajnak tekinteni – túlnyomó többségük ugyanis többszónyi terjedelmű (azaz a tulajdonnévi egységet adó karakterláncok rendszerint tartalmaznak szóközt). Ezen a ponton természetesen érintkezik a tulajdonnevek írásának kérdésköre a különírás és egybeírás területével, ez kiegészül azonban a kis- és nagybetűk használatának problematikájával is. Itt talán még fokozottabb szerepe van a szemantikának, hiszen a denotátum tulajdonnévi osztályai is tükröződhetnek az írásképben, például: *Magyar Nyelv* (folyóiratcím) – *Magyar nyelv* (könyvcím), *Tátrai-vonósnégyes* 'Tátrai Vilmos által alapított, általa vezetett kvartett, illetve őáltala komponált, ilyen összeállítású hangszeregyüttesre írt ciklikus mű' – *Tátrai vonósnégyes* 'Tátrai Vilmos emlékére, tiszteletére elnevezett kvartett' – *Tátrai Vonósnégyes* 'ez utóbbi mint jogilag is intézménnyé alakult társaság', *Gellért-hegy* 'domb Budán a Duna jobb partján az Erzsébet hídnál' – *Gellérthegy* 'ez mint városrész', *Tisza híd* 'Tisza Kálmánról elnevezett híd' – *Tisza-híd* 'a Tiszán átívelő híd', *magyar állam* (közszói megnevezés) – *Ohio állam* (országrésznév, vö. *Csongrád megye*) – *Izrael(i) Állam* (államnév, vö. *Magyar Köztársaság*), *Szent István* 'a magyar államot megalapító király' – *Szentistván* (település), *Madách Színház – Madách mozi, Béke Szálló –*

Béke étterem; Békás patak (a patak neve önmagában a *Békás*) – *Gombás-patak* (a patak nevének része a *patak* földrajzi köznévi utótag is).

A kategoriális különbségek megjelennek az *-i, -s, -beli* képzős alakokban is. Itt külön szerepe van az egyes alkotótagok tulajdonnévi vagy közszói voltának is: *kossuthi – shakespeare-i – rippl-rónais – Csokonai Vitéz-i, nemzeti színházi – Madách színházi, Békás pataki – Békás-szorosi* (mert az előtag a *Békás* patak tulajdonneve) – *gombás-pataki, országos Széchényi könyvtári – holt-Tisza-bereki, móriczi – Móricz-féle; kosztolányis – Népszabadság-os – nyugatos* (egyszeri kivétel) stb.

További problémát jelent bizonyos tulajdonnévi kategóriák esetében a kodifikáció és az úzus között feszülő oly mértékű diszkrepancia, amelyről valamilyen formában már a tanácsadásnak is tudomást kell vennie (pl. események, rendezvények elnevezésének, illetve intézmények alegységeinek szabálytalan, de általánosan elterjedt nagybetűs írása), valamint azok a tulajdonnévtípusok, amelyeket nem vagy csak nagyvonalakban kodifikált az 1984-ben megjelent, ma is hatályos helyesírási szabályzat (pl. a címadási szokások megváltozása; a címmel ellátható műfajok sokaságának megjelenése; a programok, akciók, pályázatok korábban elképzelhetetlen változatosságban való használata; a márkanevek jogi kérdéseket is felvető íráspróblémái; a legkülönbözőbb fajtájú alapítványnevek; a díjak, kitüntetések elnevezésének alapjaiban új típusai).

A földrajzi nevek bonyolult szaknyelvi szabályozásáról vagy a kémiai elnevezések helyesírásáról, az állat- és növényneveknek a taxonómiát tükröző írásmódjaival csak a távolabbi jövőben lesz mód foglalkozni.

4.2 A magyar nyelvbe bekerülő idegen elemek

Az idegen szavak, nevek, illetve kifejezések részrendszere alapvetően két lényegi kérdést vet fel.

Az első és általánosabb annak problémája, hogy egy újonnan a magyar nyelvbe kerülő szó, kifejezés idegenes vagy magyaros írásmóddal írattassék-e. Az ennek meghatározásához szükséges, formális és kategoriális szempontokon alapuló döntési fa a szükséges kommentárokkal együtt megtalálható az Osiris Helyesírásban [3]. Ezt egészíti ki az egyszavas köznevek kezelésére vonatkozó eljárás. Ennek lényege, hogy azon idegen eredetű szavak esetében, amelyek korábban nem szerepeltek normatívnak tekinthető szótárban, 40%-os vagy a feletti magyar írásmódú korpusz-előfordulás esetén (ha egyéb, releváns szempont nem merül fel), a magyaros írásmód támogatandó. Korlátozottan, de ugyanez követendő, ha szerepel az adott szó normatív szótárban, de idegenes írásmóddal (ekkor ugyanis nyelvhasználati változás tehető fel).

A második és speciálisabb probléma az idegen írásrendszerből való átírás kérdéskörre. Mivel az átírási szabályzatok jól formalizálhatók (akár az eredetiből, akár más átírásból indulunk ki), ennek számítógépes támogatása igen sikeres lehet.

4.3 Írásjelhasználat

Az írásjelhasználat szabályozása sok tekintetben fakultatív, alapjául azonban mégis csak a szintaktikai szerkezet elemzése szolgál. Ebben a tekintetben – igaz korlátozot-

tan – használhatók parciális szintaktikai szabályok (pl. két azonos esetű főnév általában nem követheti közvetlenül, írásjel nélkül egymást, de: *a városban decemberben;* két véges igealak között általában kell lennie egy írásjelnek, de problémát jelentenek a befejezett melléknévi, illetve az igei igenevek mint a véges igealakokkal homonim formák: *ettem az anyám sütötte kenyérből, ettem az anyám által sütött kenyérből*). A felvethető kérdéseknek ezek azonban csak szűkebb körére adnak választ. Szükséges tehát a mélyebb szintaktikai elemzés kialakításán túlmenően bizonyos szövegtani, stilisztikai, pragmatikai szempontok figyelembevétele. Hogy ezekből mennyi formalizálható, illetve milyen módon lehet ezeknek az esetében az interaktív felületet felhasználni, további megfontolásokat igényel. Ezek kifejlesztése csak a távolabbi időben lehetséges.

4.4 Rövidítések, mozaikszók

A rövidítésekre és mozaikszókra különféle helyesírási szabályok sokasága vonatkozik, a tény azonban mégiscsak az, hogy a szabályos írásmódú formák kisebbségben vannak a különféle hagyományos esetekkel szemben. Így ebben a körben a szabályismertetésen és a szótári keresésen túlmutató megoldást tervezni jelen ismereteink szerint nem lehetséges.

4.5 Keltezés, a számok írása

A keltezéssel, illetve a számok írásával kapcsolatos helyesírási tudnivalók igen egyszerűek és eleve formálisak, tehát számítógépes támogatásuk nem okoz komolyabb nehézséget.

5 További feladatok – kiejtéskövető írás vs. helyesírás, hibás szavak gyűjteménye, illetve a magyar nyelv határon túli változataiban használatos szavak gyűjteménye

A helyesírási segédletek (legyen az könyv vagy számítógép) elsősorban azok számára jelentenek támogatást, akik tisztában vannak a helyesírás alapvető kategóriáival (pl. a hangjelölés alapelveivel [kiejtés szerinti, szóelemző, hagyományos, egyszerűsítő írásmód], a helyesírás alapfogalmaival [pl. értelemtükröztetés, tulajdonnévosztályok], illetve a helyesírási kodifikáció mögött álló nyelvtani modell felépítésével és fogalomhasználatával). A szélesebb felhasználói kör kiszolgálásának érdekében a tervek közt szerepel egy olyan modul beiktatása is, amely hatékonyan kezeli a kiejtéskövető írásmódot is. A magyarországi helyesírási segédeszközök között újítás lenne, hogy a szoftver nemcsak a helyesírási vétség(ek), illetve az elütés(ek) miatt hibásan leírt szavakat ismerné fel és tudná javítani, hanem a köz- vagy tájnyelvi kiejtést tükrözve leírtat is. A hibásan beírt szavak esetében egyrészt a szokásos eljárás szerint felkínálja a lehetséges jó változatokat (ez elsősorban elgépelésnél lehet hasznos), másrészt egy

speciális elemző modul segítségével felismeri a kiejtés alapján a mögöttes morfémaszerkezetet, s végül felkínálja a helyesírás szerinti alakot.

Ez azért meghatározó újdonság, mivel azok, akik nincsenek tisztában a helyesírás alapvető szabályaival sem, a kiejtést tükröző alakot hallás után leírva eleve nem férnek hozzá a helyesírási szótárakban elérhető ismeretanyaghoz. A magyar nyelv szavainak, kifejezésének írott és beszélt formája között feszülő eltérés alapvető szabályait felhasználva lehetőség nyílik a kiejtést tükrözve leírt szavak írott alakúra történő változtatására (illetve a kétszintű morfológiához hasonlatos módon az ellenirányú átalakítás is megoldható szükség esetén). Hangtani szabályok ismerete alapján a rendszer felismeri a kérdező szándékát, és ez alapján generálja a szóelemzés elvét is figyelembe vevő alakot, például:

szimpad » szinpad [mp↔np], színpad [i↔í]
teccik » tetszik [cc↔tsz]
aggyá » adjá [ggy↔dj], adjál [szó végi l↔Ø]
kiszcica » kiscica [szc↔sc]
egésség » egészség [ss↔szs]
pallament » parlament [ll↔rl]
tejjes » teljes [jj↔lj]
báttya » bátyja [tty↔tyj]

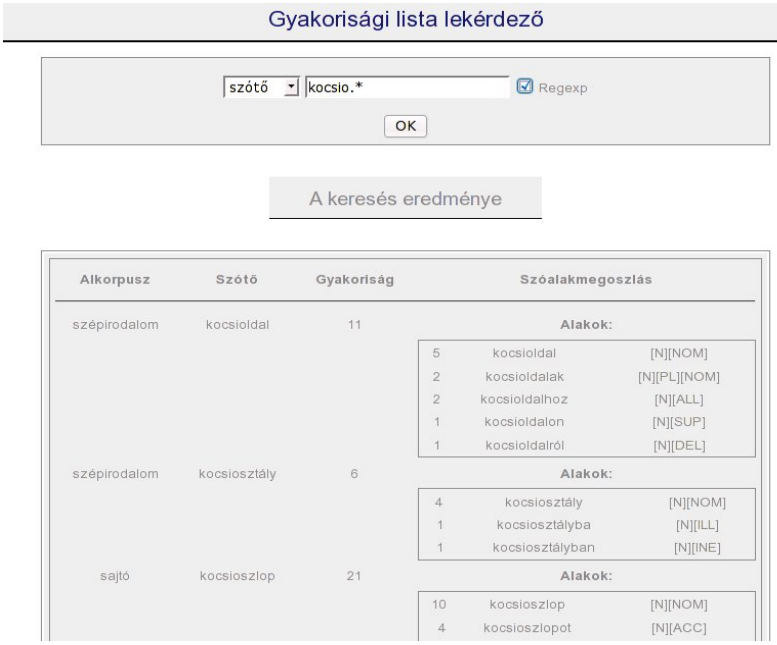
A hibásan írt szavak kezelésének további erőforrása a leggyakrabban hibásan írt szavak gyűjteménye (mintegy 120 ezer tétel), amely javarészt az MTA Nyelvtudományi Intézetében zajló helyesírási tanácsadás gyakorlatából származik, a gyakran előforduló, tipikus hibák gyűjteményén alapszik.

Amint látható, a hibás alakban keresett szót több szűrőn keresztül ellenőrizve jutunk el a helyesen leírt alakig, amely még korántsem a végső alak, mivel több lehetséges megoldás esetén itt is szükség lehet még a kérdező általi egyértelműsítésre.

A helyesiras.hu célközönségeként nemcsak a magyarországi nyelvhasználókra, hanem a legtágabb értelemben vett magyar nyelvközösségre is gondolunk. Éppen ezért a szótár nemcsak a magyarországi magyar nyelvváltozatok szókészletét tartalmazza majd. (Természetesen a magyarországi magyar nyelvváltozatok közül a kizárólag beszélt nyelvi formában élő területi, illetve csoport- és rétegnyelvi változatok problémáival, tehát azon lexikai tételekkel, amelyeknek nincs és esetleg nem is lehet kodifikált helyesírásuk, nem foglalkozunk.) Már az alapvető erőforrásnak számító MNSz. is tartalmaz mintegy 23 millió szövegszónyi határon túli korpuszt, amely mellé bekerül egy közvetlen kölcsönszavakból álló, annotált, ún. ht-szólista (http://ht.nytud.hu). Ez még kiegészül az MTA határon túli kutatóállomásai által gyűjtött magyar etimonú földrajzi nevek, intézménynevek, díjak és címek megnevezéseit tartalmazó szóanyaggal. (A földrajzi neveknek a Földrajzinév-bizottsággal való egyeztetése ehhez elkerülhetetlen.) Ez utóbbiak országra utaló megkülönböztető jelzéssel lesznek ellátva, így lehet ugyanis kezelni a nyelvváltozatok helyesírási vetületének esetleges ütközéseit is, bár az ilyen esetek számát a minimálisra kell szorítani a helyesírás egysége érdekében.

Hasonló módon kezelhetők a jövőben egyes szaknyelvi részszótárak is. Ezek, illetve általában a szaknyelvi helyesírás kérdései további bővítési-fejlesztési lehetőséget

kínálnak a helyesiras.hu portál számára. Ezek megoldásához az egyes szakmák művelőivel is ki kell építeni a megfelelően szoros munkakapcsolatot.



1. ábra. Az adatbázis már működő lekérdezőfelülete.

Hivatkozások

1. Kis Ádám: Gépszerű helyesírás. Az akadémiai helyesírási szabályzat és a számítógép. http://mek.iif.hu/porta/szint/tarsad/nyelvtud/gepscikk/ (1997)
2. Kis Ádám: Az akadémiai helyesírási szabályzat és a számítógép. Magyar Nyelvőr 123 (1999) 149–168.
3. Laczkó Krisztina, Mártonfi Attila: Helyesírás. Osiris Kiadó, Budapest. (2004)

IV. Beszédtechnológia

Nagyszótáras híranyagok felismerési pontosságának növelése morfémaalapú, folyamatos beszédfelismerővel

Tarján Balázs, Mihajlik Péter, Tüske Zoltán

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{tarjanb, mihajlik, tuske}@tmit.bme.hu

Kivonat: Morfémaalapú beszédfelismerőnek tekintjük azokat a felismerőket, melyek szónál kisebb, morfémaszerű elemekre épülő nyelvi modellt használnak. Kísérleteink során öt különböző szegmentáló eljárással készített morfémaalapú felismerő teljesítményét hasonlítottuk egy standard, szó alapú rendszeréhez tervezett beszédű, híranyag felolvasásos feladaton. Megállapítottuk, hogy mind statisztikai, mind szabályalapú szegmentáló algoritmust használva, morféma alapon jelentős mértékben növelni lehet a felismerési pontosságot. Különösen alacsony hibaarányt értünk el egy hibrid eljárással, mely a statisztikai módszert nyelvspecifikus tudással egészíti ki. Felügyelet nélküli beszélőadaptációs technológiával kiegészítve, ily módon sikerült 20% alá csökkentenünk a szóhiba-arányt, mely tudomásuk szerint a legalacsonyabb eddig publikált eredmény magyar nyelvű, nagyszótáras, folyamatos beszédfelismerés területén.

1 Bevezetés

A nemzetközi gyakorlatban a **folyamatos, nagyszótáras beszédfelismerő rendszerekben** (LVCSR – Large-Vocabulary Continuous Speech Recognition) tipikusan szóalapú nyelvi modellezést alkalmaznak. Azonban a morfológiailag gazdag nyelveknél – mint amilyen a magyar – e szóalapú megközelítés alkalmazása a jelentős szóalaki változatosság miatt megkérdőjelezhető. A klasszikus nyelvi modell az egyes N-gramok (szó N-esek) relatív gyakorisága alapján becsüli meg egy N-1 szóból álló előtörténet ("history") után álló szavak feltételes valószínűségét. Sok szóalak esetén kevés tanítóminta áll rendelkezésre egy kontextus becsléséhez, így a nyelvi modell döntése is kevésbé megalapozott különösen, ha akusztikailag nehezen megkülönböztethető szavak között kell választania.

Megoldást kínál a nyelvi modellek új, nyelvünkhöz jobban illeszkedő alapokra helyezése. Ehhez morfémaalapú nyelvi modelleket kell létrehozni a tanítószövegek szegmentálásával, majd a mintaillesztési folyamatot szavak sorozata (W) helyett morfémák sorozatán (M) kell elvégeznünk. (**1**)

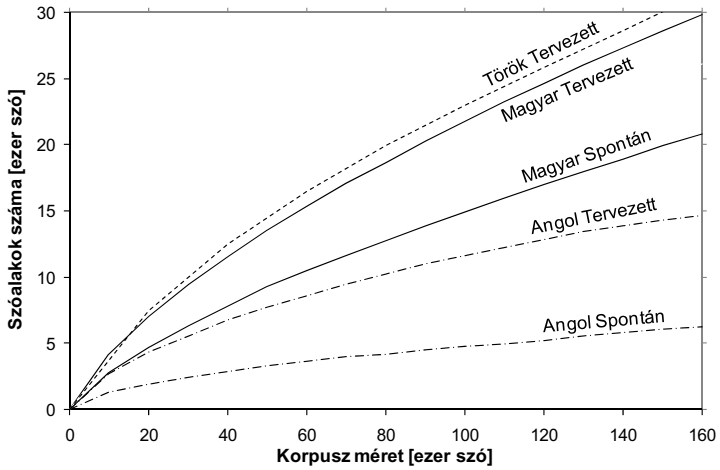
$$\hat{M} = \arg\max_{M} P(M) * P(O \mid M) \tag{1}$$

Ahol O a felismerendő beszédanyagból nyert jellemzővektorok sorozata.

A mintaillesztés morféma alapon is ugyanúgy elvégezhető, az egyetlen különbség, hogy a kimeneten is morfémasorozatot kapunk. Ezt újra szavakká összeilleszteni egyszerű feladat, ha előzetesen jelöltük a tanítószövegben a szóhatárokat. (**2**)

$$\hat{W} = f\left(\hat{M}\right) \tag{2}$$

A morfológiai gazdagságot jól jellemzi a felismerési feladat szótár-növekedési görbéje (1. ábra). Megfigyelhető, hogy az agglutináló nyelveknél, mint a magyar vagy a török mennyivel gyorsabb a szótárbővülés, és ennek üteme nem csak a nyelvtől, hanem a felismerési feladat jellegétől is erősen függ. Spontán beszédadatbázison alkalmazva a morfémaalapú megközelítést korábbi munkánkban [1] csak kismértékű javulást értünk el a szó alapú rendszerhez képest. Ezzel szemben [2] jelentős hibaarány csökkenésről számol be török nyelvű, olvasott híranyagokon végzett kísérletek alapján. Morfémaalapú rendszerükkel 20% körüli szóhiba-arányt értek el. Ennek hátterében az állhat, hogy a híranyag-felolvasás szóalakokban gazdagabb feladat, mint a spontán beszéd. Figyelembe véve, hogy a magyar és a török nyelv tervezett beszédadatbázisokon nagyon hasonló szótárnövekedést mutat (1. ábra), okkal feltételezzük, hogy olvasott híranyagon, magyar nyelven is jobban teljesíthetnek a morfémaalapú felismerők.



1. ábra. Szótárméret növekedés a korpusz méretének függvényében (források - spontán magyar: **[1]**; török és angol eredmények: **[3]**)

Cikkünkben öt különböző szegmentáló eljárást használó **morfémaalapú** felismerőt mutatunk be, melyek pontosságát egy standard, szó alapú felismerő pontosságával vetjük össze. Az említett rendszerek tanítása internetes híranyagok felhasználásával készült szöveges tanító-adatbázison történt. A teszteléshez egy kb. egy óra hosszúságú felolvasott híreket tartalmazó felvételt használtunk, mely egy országosan fogható televízió adásából került rögzítésre. A felismerési feladat részletes áttekintése után bemutatjuk hogyan nyertük a mintaillesztési folyamatban használt morfémákat, majd ismertetjük a tesztanyagon elért eredményeket. Végül összefoglalást adunk kísérleteink legfontosabb következményeiről.

2 A felismerési feladat

2.1 A szöveges tanító-adatbázis

A tanító-szöveg összegyűjtése

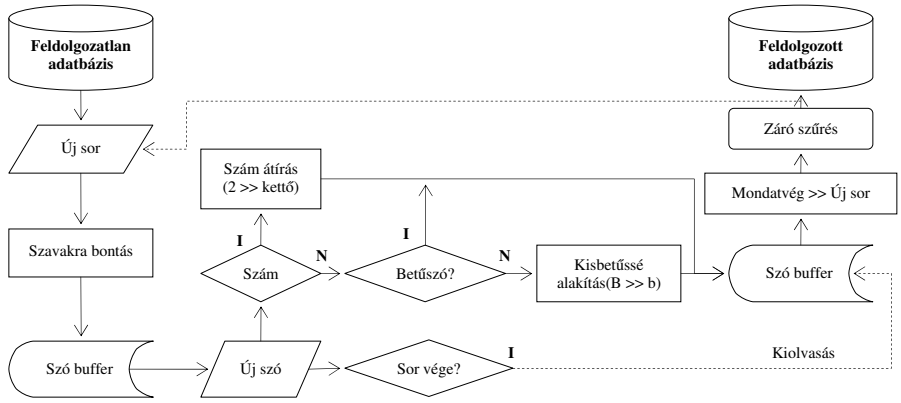
A cikkünkben ismertetett felismerő rendszer szöveges tanító-adatbázisa internetes gyűjtés eredménye, és egy országosan fogható televíziós csatorna portáljáról származik. Beszédfelismerők tanításnál a rendelkezésre álló beszédadatbázist általában több részre osztják, aszerint hogy tanításra, tesztelésre vagy modell paraméterek hangolására kívánják-e használni. Ebben az esetben általában a tanító beszédadatok leiratait használják fel a nyelvi modell tanításához. Ez azonban nagyon idő- és költségigényes, mivel kézi úton kell átírni a beszédjelet szöveggé.

Némi kompromisszum árán, de van hatékonyabb megoldás. Összegyűjthetőek különböző forrásokból a felismerési feladathoz tematikában, szókincsben, struktúrában jól illeszkedő szöveges tanítóadatok. Az ilyen módon készített nyelvi modellek természetesen valamelyest kevésbé illeszkednek a konkrét beszédadatbázishoz, azonban előnyük lehet, hogy gyorsabban, nagyobb méretben előállíthatóak, és kisebb célzottságuk miatt robosztusabb működést eredményezhetnek. Esetünkben az összegyűjtött tanítószövegek egy TV csatorna hírportálként is működő honlapjáról származnak. Az itt fellelhető belföldi híranyagok hat évre visszamenőleg kerültek összegyűjtésre, ami mintegy 54 ezer cikk feldolgozását jelenti. Ezekből állt össze a felismerő rendszer nyers tanító-adatbázisa.

A tanítószöveg előfeldolgozása

A beszédfelismerési feladat a szöveges tanítóadatok speciális előfeldolgozását követeli meg. Egy nyers internetes híranyag rengeteg olyan karaktersort tartalmazhat, melyeket nem kívánunk a nyelvi modellbe beépíteni, vagy a többi szótól eltérően szeretnénk kezelni. Az előbbire lehet példa a megjelenési dátum, forrásmegjelölés, esetlegesen bent maradt HTML tag, stb., mivel ezek csak feleslegesen rontják a nyelvi modell minőségét. Míg az utóbbira a betűszavak, vagy számok esete, melyekre a későbbiekben kitérek. Mindezek mellett az is elmondható, hogy azokat az írásjeleket is el kell távolítani a tanítószövegből, amelyeknek nincs egyértelmű megfelelője a hangképzésben, mint például a vessző, a kettőspont vagy a kötőjel. Az előfeldolgozás lépéseit szemlélteti a 2. ábra.

A nyers tanítószöveg feldolgozása soronként történik, **Perl** (Practical Extraction and Report Language) nyelven írt scriptek segítségével. Minden sor beolvasás után az egyes szavakon folytatjuk a feldolgozást. Erre azért van szükség, mert a számok és betűszavak az átlagos szavaktól eltérő eljárást igényelnek. Számok feldolgozásánál a nehézséget az jelenti, hogy nem rendelhető hozzájuk egyértelműen a kiejtett, fonetikus alakjuk. A probléma feloldásához azt a stratégiát választottuk, hogy átírjuk őket betűvel leírt alakjukra, így a felismerő ugyanolyan módon tanítható velük, mint a normál szavakkal. Ezt az átírást egy Perl szubrutin hajtja végre, tehát automatizáltan történik.



2. ábra. A tanító-adatbázis előfeldolgozása

A második megoldásra váró eset a betűszavaké. Ezekkel az a gond, hogy a magyar nyelvben egyszerű szabályokkal a kimondott alakjuk nehezen jósolható. Néha teljes értékű szóként ejtjük őket (pl.: APEH, KRESZ), néha viszont betűnként olvassuk ki (pl.: ÁNTSZ, DVD), nem is beszélve arról az esetről, ha idegen eredetű rövidítésről van szó (pl.: GDP, BBC). A számokkal ellentétben ezt egyszerű programozási eljárással nem lehet kezelni, így a fonetikai átiratukat a kiejtési modellben, kézzel adjuk meg. Ahhoz hogy ezt megtehessük, el kell kerülni a morfémákra bontásukat. Ezt legegyszerűbben úgy érhetjük el, ha meghagyjuk őket nagybetűs alakjukban, így formájukban elkülönülnek az átlagos szavaktól. Mindezek után minden egyéb szót kisbetűsítünk a tanítószövegben.

Hogy eljussunk a nyelvi modellezéshez alkalmas formához, már csak két lépést kell megtenni. Az elhangzó mondatokat egymástól függetlennek tekintjük a nyelvi modell szempontjából, ezért a mondatvégi írásjeleket „új sor” szimbólumra cseréljük. A második lépésben azt biztosítjuk, hogy a tanítószöveg végső alakjába ne kerülhessen semmilyen a felismerés folyamán nem értelmezhető karakter. Az ehhez alkalmazott szűrőfeltétel csak a magyar ábécé betűit engedi meg, minden más írásjelet töröl.

2.2 Akusztikai tanító-adatbázis

Az akusztikus modell tanításához összesen egy órányi átírt híranyag állt rendelkezésünkre. Ez önmagában kevés egy teljesen új akusztikus modell felépítéséhez, sőt mivel a teljes egy órát tesztelési célokra szerettük volna fenntartani a felügyelt adaptációról is le kellett mondanunk. Éppen ezért **beszélőfüggetlen** akusztikus modellként egy korábbi, a mostanitól független felismerési feladathoz illesztett modellt használtunk, mely eredetileg Magyar Referencia Beszédadatbázison (MRBA) [4] lett tanítva.

Bár felügyelt adaptálásra nem volt lehetőségünk, felügyelet nélkülit azonban végeztünk a rendelkezésre álló egy órás felvétel és a beszélőfüggetlen modell felhasználásával. Az így keletkezett akusztikus modellt használtuk **beszélőadaptált** kísérleteinkhez.

3 Morfémaszegmentálás

Mint a bevezetőben kitértünk rá a szóalapú nyelvi modellezés nehézségei főként nyelvünk szóalaki változatosságból erednek. A magyar nyelvben egyetlen szónak rengeteg képzett-ragozott formája létezik, így ugyanaz a szótő különböző kontextusban eltérő formákat vehet fel. Ennek következtében a tanítószövegben rendelkezésre álló információ elaprózódik, ami a szókapcsolatok pontatlan becslését eredményezi. E változatos morfológiával úgy küzdhetünk meg a legjobban, ha a szótári szavakat kisebb elemekre tudjuk bontani. Ha ezt a szegmentálást optimálisan hajtjuk végre, csökkenteni tudjuk a szótár méretét. Kisebb szótárméretnél a szótári elemek többször fordulnak elő, így több mintát szolgáltatnak a nyelvi modell tanításához, ami pedig végső soron hatékonyabb mintaillesztést tesz lehetővé.

A feladat elvégzéséhez azonban olyan módszerek bevezetése szükséges, amelyekkel a (**2**) képletben bevezetett f függvény inverze, a W-ről M-re képző f^{-1} optimálisan megvalósítható.

3.1 Szabályalapú eljárások

A szegmentáló eljárások közül először a nyelvspecifikus szabályokon és szótáron alapuló módszereket mutatjuk be röviden. Kísérleteinkben a magyar nyelvű **Hunmorph** [5], általános célú annotáló rendszert használtuk, melynek részeként futásidejű morfológiai elemzésre az ún. Ocamorph program szolgál. Tudásforrásként a Morphdb.hu [6] adatbázist használja, mely minden eddiginél mélyebben megalapozott morfológiai leírását tartalmazza a magyar nyelvnek.

Fontos jellemzője a szabályalapú módszereknek, hogy az egyes szavak egymástól függetlenül kerülnek elemzésre, és az elemző általában több lehetséges szegmentálást is megad. A megfelelő kiválasztására valamiféle stratégiát kell alkalmazni. Ennek megfelelően két változata született az elemzésnek a szegmentáló program beállítása és a szegmentált alak kiválasztása szerint.

Hunmorph Compound-Guessing (HCG)
A morfológiai elemzés ezen változatában a szegmentálás minden módja megengedett. A futásidejű elemző felbonthatja az összetett szavakat (*--compounds* kapcsoló) sőt, ha nincs egy szóalaknak semmilyen érvényes elemzése, akkor egy az adatbázisban nem szereplő tagot is leválaszthat róla feltéve, hogy így elemezhető alakhoz jut (*--guess Fallback* kapcsoló).

Hunmorph Strict Fallback (HSF)
A szabályalapú morfológiai elemzés második változata a feldolgozás lépcsőzetességére épül. Az első lépcsőben a Hunmorph számára csupán az egyértelműen feldolgozható szavak elemzése engedélyezett. A második lépcsőben az így nem elemezhető szavakra megengedjük, hogy összetett szóként legyenek figyelembe véve. Végül az ezután is felbonthatatlan szóalakokról az adatbázisban nem rögzített elemek is leválasztásra kerülhetnek.

3.2 Statisztikai alapú eljárások

A Morfessor család tagjai Minimum Description Length (MDL) elven alapuló statisztikai eljárások, amelyeket finn kutatók fejlesztettek ki. A statisztikai alapú szegmentáló eljárások nagy előnye, hogy működésük nem igényel emberi felügyeletet. Nem használnak sem nyelvspecifikus lexikont, sem toldaléklistát, hanem a bemenetül kapott szótár statisztikai tulajdonságai alapján bontják a szavakat kisebb elemekre. Céljuk olyan optimális felbontást találni, mely tömören képes a korpuszt reprezentálni. Ennek következtében a szegmentálás eredményeként kapott morfémaszerű egységek (ún. **morf**ok) nem feltétlenül rendelkeznek jelentéssel.

Morfessor Baseline (MB)
A Morfessor Baseline a finnek módszerének alapváltozata. Az optimális szegmentálás megkeresését a következőképpen önthetjük matematikai alakba. (**3**)

$$\arg\max_{M} P(M \mid korpusz) = \arg\max_{M} P(M) * P(korpusz \mid M) \qquad (3)$$

Ahol $M = \mu_1, \ldots, \mu_m$ a korpusz egy lehetséges morf felbontását jelöli. A mintaillesztéshez hasonlóan itt is két paraméter értékének szorzatát kell optimalizálni ahhoz, hogy a korpusz legnagyobb valószínűségű szegmentálását megkapjuk: az adott M morfkészletnek, mint lexikonnak a valószínűségét ($P(M)$) és M korpuszhoz való illeszkedési valószínűségét ($P(korpusz \mid M)$). E két változó közelítésről bővebb leírás a kapcsolódó irodalomban található [7].

Morfessor Categories-MAP (MC-MAP)
A Morfessor család második tagja a baseline módszer finomításaként született statisztikai alapú szegmentáló eljárás. Legnagyobb újításként az algoritmus megpróbálja kikövetkeztetni, hogy az egyes morfok prefixum, szótő vagy szuffixum szerepet töltenek-e be, és ezt a morf után helyezett címke (/PRE, /STM, /SUF) segítségével jelöli is a végeredményben. Bár ez tagadhatatlanul növeli a szótárméretet – ugyanis így ugyanaz a morf akár három lexikai elem szerepét is betöltheti – mégis megtérülhet ez a fajta megkülönböztetés a pontosabb nyelvi modellezésben. Részletekért lásd [8].

3.3 Hibrid eljárás

Mind a szabályalapú, mind a statisztikai alapú morféma szegmentálásnak lehetnek hátulütői. A Hunmorph-fal történő elemzések nem eredményeznek tömör szótárat, ami beszédfelismerési feladatnál nem előnyös, hiszen minél több szótári elem között kell a dekódolás során különbséget tenni, annál nagyobb a hiba lehetősége is. Ezzel szemben a statisztikai eljárások hatásfokát a nem elégséges mennyiségű tanítóadat ronthatja le. Éppen ezért felmerült az igény a két szemlélet egyesítésére.

A **Combined Hunmorph-Morfessor (CHM)** eljárás a HCG szabály alapú módszer szegmentálásán alapszik. Lényegében a MB algoritmus átalakítása oly módon, hogy a szegmentálás valószínűségi becslését a HCG által szolgáltatott felbontási

alternatívákon végzi el. Így tehát csak olyan morfémák keletkezhetnek, melyek a szabály alapú eljárással jöttek létre, viszont a statisztikai módszer biztosítja, hogy arra az alternatívára essen a döntés, mely globálisan tömör morfémakészletet eredményez. Az ehhez szükséges algoritmust [9] részletezi.

4 A beszédfelismerő hálózatok kiértékelése

4.1 Beszédfelismerési paraméterek és beállítások

Minden felismerő rendszer ugyanazon a kb. egy órás olvasott híranyagon lett kiértékelve. A 16 kHz-en mintavételezett és 16 biten kvantált felvételek lényegkiemeléséhez dinamikus Delta és Delta-Delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseket (**MFCC** – Mel-frequency cepstral coefficients) és ún. vak csatornakiegyenlítést alkalmaztunk. Akusztikus modellként balról-jobbra struktúrájú, háromállapotú rejtett Markov-modelleket használtunk, állapotonként 7 Gauss függvényből álló sűrűségfüggvényekkel.

A szóalapú nyelvi modell tanítása a normál, előfeldolgozott tanítószövegen, míg a morfémaalapú felismerők esetén, ugyanezen szöveg szegmentált változatain történt. Minden felismerő Kneser-Ney simított [10], trigram nyelvi modellen alapszik, mely az SRI-LM [11] nyelvi modellező toolkit segítségével lett előállítva. Entrópia alapú modell-metszést [12] csak a szó alapú hálózatban alkalmaztuk, memória takarékossági szempontból (1. táblázat).

1. táblázat: A tanító- és tesztadatbázis adatai

	Szavak száma [ezer szó]	Szóalakok száma [ezer szó]	OOV arány [%]
Tanító-	5600	285	–
Tesztadatbázis	7.6	3.5	3.6

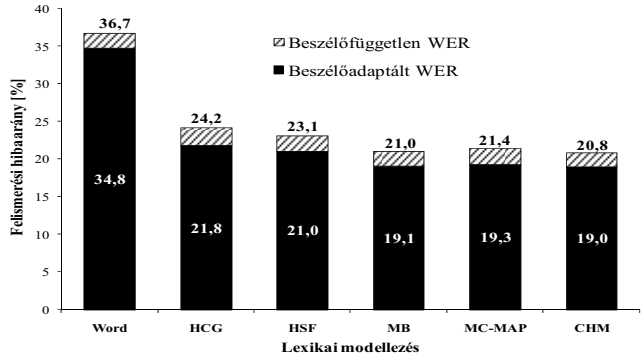
A felismerő hálózatok építését az Mtool WFST (Weighted Finite State Transducer) keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztési feladathoz a VOXerver [13] nevű dekódert használtuk. A dekódolási folyamat számításigénye az ún. Real Time Factor (RTF) tekintetében az egyes feladatok között kiegyenlítésre került.

4.2 Felismerési eredmények

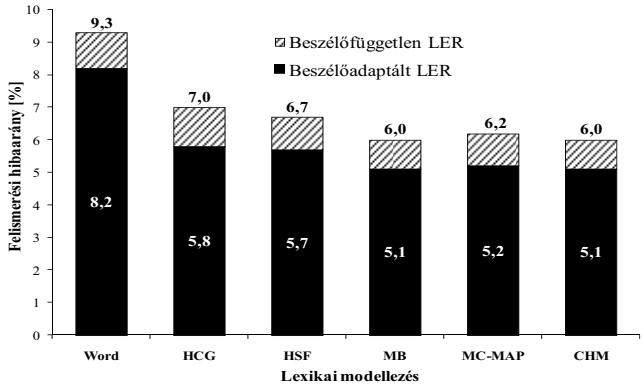
A felismerő rendszerek teljesítményének értékeléséhez szóhiba-arányt (**WER** – Word Error Rate), illetve betűhiba-arányt (**LER** – Letter Error Rate) számoltunk. Ezen kívül feltüntettük a szó alapú kiindulási rendszerhez képest elért relatív WER csökkeneséket beszélőfüggetlen és beszélőadaptált esetben is (2. táblázat, 3. ábra).

2. táblázat: Beszédfelismerési eredmények

Technika	Szótár mérete	Beszélőfüggetlen eredmények			Beszélőadaptált eredmények		
		WER [%]	-ΔWER$_{rel}$ [%]	LER [%]	WER [%]	-ΔWER$_{rel}$ [%]	LER [%]
Word	285 e	36.7	–	9.3	34.8	–	8.2
HCG	50 e	24.2	*37*	7.0	21.8	*37*	5.8
HSF	63 e	23.1	*34*	6.7	21.0	*40*	5.7
MB	31 e	21.0	*43*	**6.0**	19.1	*45*	**5.1**
MC-MAP	45 e	21.4	*42*	6.2	19.3	*45*	5.2
CHM	80 e	**20.8**	*43*	**6.0**	**19.0**	*45*	**5.1**



3.1 ábra. Beszélőfüggetlen és beszélőfüggő szóhiba-arányok



3.2 ábra. Beszélőfüggetlen és beszélőfüggő betűhiba-arányok

4.3 Értékelés

Az eredmények ismeretében elmondható, hogy a morfémaalapú felismerők szignifikánsan (Conf.=95%) jobban teljesítettek a felismerési feladaton, mint a szó alapú rendszer. Az újonnan bevezetett eljárásokkal elért átlag 40%-os relatív felismerési hiba csökkenés figyelemre méltó eredmény, és egyben felhívja a figyelmet arra, hogy nyelvünk jobban modellezhető morféma alapon egy szóalakokban gazdag feladat esetén. A bevezetett módszerek közül a legnagyobb felismerési pontosságot a hibrid szegmentáló eljárással (CHM) sikerült elérni, de fontos megjegyezni, hogy ettől szignifikánsan a statisztikai megközelítésekkel (MB, MC-MAP) nyert eredmények sem térnek el. A statisztikai és hibrid rendszerekhez képest viszont szignifikánsan magasabb felismerési hibát kapunk a szabályalapú szegmentálások (HCG, HSF) alkalmazásakor. Érdekes megfigyelni a beszélőadaptáció hatását is, miszerint jellemzően tovább növeli a szó és morfémaalapú felismerő közötti pontosság különbséget [14].

Vizsgáljuk meg, mitől pontosabbak a morfémaalapú rendszerek. Morféma alapon OOV szavak is felismerhetők. Míg a szó alapú felismerő csak a szótárában található szavakat képes helyesen felismerni, addig a morfémaalapú rendszerekben a lexikai elemek elvben tetszőlegesen összekapcsolódhatnak, így a tanítószövegben nem szereplő szavak is előállhatnak a mintaillesztés folyamán. Figyelembe véve azonban, hogy a tesztanyag mindössze 3.6%-a OOV szó, ez önmagában nem adhat választ az ennél jóval magasabb abszolút hiba csökkenésre. A döntő tényező valójában, a szó alapon fennálló adatelégtelenség kezelése, aminek köszönhetően a szótáron belüli szavak felismerésekor is jóval kevesebb helyettesítéses és törlődéses hiba keletkezik.

5 Összefoglalás

Cikkünkben öt különböző szegmentáló eljárással készített morfémaalapú gép felismerő teljesítményét hasonlítottuk egy standard, szó alapú rendszeréhez tervezett beszédű, híranyag felolvasásos feladaton. Minden újonnan bevezetett módszerrel szignifikáns, átlag 40%-os relatív hibaarány csökkenést sikerült elérnünk, mely a morfológiailag gazdag feladat pontosabb nyelvi modellezésére vezethető vissza. Különösen jól teljesítettek a statisztikai alapú szegmentáló technikák, ezen belül is legkiemelkedőbben egy hibrid eljárás, mely nyelvspecifikus tudást is felhasznált. Felügyelet nélküli adaptációs technológia segítségével a szóhiba-arányt 20% alá tudtuk szorítani, mely tudomásunk szerint egyedülállóan alacsony magyar nyelvű, LVCSR feladaton.

Korábbi munkáink [1],[15] is bizonyították, hogy más nyelvekhez hasonlóan magyar nyelven is eredményesen alkalmazható a morfémaalapú nyelvi modellezés, azonban ilyen mértékű javulás egyetlen korábbi feladat esetén sem volt mérhető. Ennek oka az lehet, hogy a szóalaki változatosság és relatív hibaarány csökkenés erős kapcsolatban áll egymással. Minél gazdagabb morfémákban a felismerési feladat, annál nagyobb szükség van olyan lexikai modellezés használatára, mely a szavaknál alacsonyabb szintű nyelvi elemeket is figyelembe veszi.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani az AITIA International Zrt.-nek és a THINKTech Kutatási Központ Nonprofit Kft.-nek a rendelkezésünkre bocsátott eszközökért és adatokért. Kutatásunkat részben az OM-00102-2007-es projekt támogatta.

Hivatkozások

1. Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T.: Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task. IEEE Transactions on Speech and Audio Processing (megjelenés alatt)
2. Arısoy, E., Can, D., Parlak, S., Sak, H., Saraçlar, M.: Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio Speech and Language Processing, vol. 17, no. 5 (2009) 874-883
3. Creutz, M. et. al.: Morph-Based Speech Recognition and Modeling Out-of-Vocabulary Words Across Languages. ACM Transactions on Speech and Language Processing, vol. 5, Issue 1, Article no. 3 (2007)
4. Vicsi K., Kocsor A., Teleki Cs., Tóth L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004)
5. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D.: Hunmorph: open source word analysis. In: Proc. ACL 2005 Software Workshop (2005) 77-85.
6. Trón V., Halácsy P., Rebrus P., Rung A., Simon E., Vajda P.: Morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005)
7. 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)
8. Creutz, M., Lagus, K.: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In: Proc. of AKRR'05, Espoo, Finland,15-17 June (2005)
9. Németh B., Mihajlik P., Tikk D., Trón V.: Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz. In: V. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (2007)
10. 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)
11. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing, Denver (2002) 901–904
12. Stolcke, A.: Entropy-based Pruning of Backoff Language Models. In: Proc. DARPA Broadcast News Transcription and Understanding Workshop (1998) 270-274
13. Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., Tatai, G.: VOXenter - Intelligent voice enabled call center for Hungarian. In: EUROSPEECH-2003 (2003) 1905-1908
14. Tüske Z., Mihajlik P., Fegyó T., Trón V.: Spontán, nagyszótáras, folyamatos beszéd gépi felismerési pontosságának növelése beszélőadaptációval a MALACH projektben. In: V. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2007)
15. Tarján B.: Large-Vocabulary Continuous Speech Recognition in Hungarian. In: Végzős Konferencia 2009, Budapest, 2009. május 20. (2009)

Zajszűrő eljárások alkalmazása, teljesítményük vizsgálata zajos beszéd automatikus felismerésénél

Sztahó Dávid, Szaszák György, Vicsi Klára

Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
1111, Budapest, Stoczek utca 2.
sztaho@tmit.bme.hu, szaszak@tmit.bme.hu, vicsi@tmit.bme.hu

Kivonat: A jelen cikk célja több zajszűrő eljárás teljesítményének összehasonlítása autók belső terében történő automatikus beszédfelismeréskor. A kutatást autóban felvett hanganyagon végeztük el német nyelvre. A zajszűrő eljárások teljesítményének összehasonlítását egy csatornán végeztük. Négyféle zajszűrő eljárást vizsgáltunk: Spectral Subtraction, Wiener-filter, Minimum Mean-Square Error Log-Spectral Amplitude Estimator, valamint modulációs spektrum szűrésén alapuló zajcsökkentés. Minden eljárásnál használtunk egy felülátereszto szűrőt is, amellyel az autó mélyfrekvenciás zaját tudtuk kiküszöbölni. A beszédfelismerési tesztekre Rejtett Markov-modell alapú felismerőt használtunk. A tesztsorozatokat két részre bontottuk. Az első tesztsorozat során megvizsgáltuk az egyes zajszűrő eljárások alkalmazhatóságát a beszédfelismerésben úgy, hogy a TELEAUTO személygépkocsi belterében rögzített hanganyagot használtuk tanításra és tesztelésre is, az adott zajszűrő eljárás alkalmazása után. A második tesztsorozat során pedig megvizsgáltuk, hogy a SpeechDat adatbázissal betanított HMM modellekkel a szűrt felvételek milyen eredményeket adnak a szűretlen személygépkocsikban rögzített felvételekkel történő felismeréshez képest. A kapott eredmények azt mutatják, hogy a zajszűrő eljárások közül az MMSE adja a legjobb felismerési százalékot az általunk vizsgált módszerek közül. Továbbá a teszteredményekből az is egyértelmű, hogy a felismerés szempontjából az a leghatékonyabb eljárás, ha a zajos beszédfelismerésnél a hasonló zajban felvett beszédadatbázissal történik a betanítás.

1 Bevezetés

Napjainkban a beszédjelek kiemelése a zajos környezetből, vagyis a zajjal terhelt beszédjelek javítása kiemelt kutatási téma. Ennek oka a számos felhasználási lehetőség, amellyel egy hatékony beszéd kiemelő rendszer rendelkezik. Ma már számos technológiai eljárás létezik ennek megvalósítására. Általánosságban elmondható, hogy a mai kifejlesztett beszédfelismerők megcélzott felhasználási környezete alacsony zajszintű. Éppen ezért egy ilyen felismerő alkalmazása zajos környezetben akkor lehetséges, ha a felismerő bemenetére már zajszűrt beszédjel kerül, vagy a felismerő akusztikai előfeldolgozó eljárását zajtűrő eljárásra cseréljük. A zajszűrés külön problémát jelent különösen olyan esetekben, amikor változó, hol állandó, hol

impulzusszerű, valamint változó hangszintű és -színezetű zaj váltakozva van jelen, mint például gépkocsik belső terében.

A jelen cikk célja több zajszűrő eljárás teljesítményének összehasonlítása autók belső terében történő, Rejtett Markov-modelleken alapuló automatikus beszédfelismeréskor.

Az egycsatornás zajszűrő eljárások összehasonlításánál az alábbi eljárásokat vizsgáltuk: Spectral Subtraction [2], Wiener-filter [3], Minimum Mean-Square Error Log-Spectral Amplitude Estimator [6][7], valamint modulációs spektrum szűrésén alapuló zajcsökkentés [1]. Miután a személygépkocsikban felvett zajos beszédet a különböző zajszűrő eljárásokkal megszűrtük, az anyagon beszédfelismerési teszteket végeztünk. A felismerési feladatok között német nyelvű információlekérés, rövidebb megerősítés jellegű mondatok, valamint hosszabb általános mondatok felismerése szerepelt. Mivel a rendelkezésre álló TELEAUTO-német adatbázis eredeti, zajos, német nyelvű felvételeit magyar anyanyelvű személyek mondták be, ezért az így betanított felismerő német anyanyelvű személyek bemondásainak felismerésére csak korlátozottan lesz alkalmas. Ezért a tesztek során két fázist különítettünk el. A tesztsorozatok első fázisában a TELEAUTO-német adatbázist használva megvizsgáltuk, hogy az egyes zajszűrő eljárások alkalmazása a felismerések során hogyan teljesít az eredeti zajos felvételekkel történő betanításhoz és felismeréshez képest. Ehhez mind a tanítás, mind pedig a tesztelés felvételeit zajszűrtük, majd beszéd-felismerési kísérleteket folytattunk.

A második fázisban az anyanyelvi német mobil telefon beszédet tartalmazó SpeechDat(II) adatbázissal [4] végeztük a HMM modellek betanítását, a tesztek pedig most is az autóban készült felvételek zajszűrt változataival történtek. A kapott eredmények megmutatták, hogy a szűrt felvételek milyen felismerési teljesítményt adnak a szűretlen felvételekkel történő felismeréshez képest.

A cikk részeiben először bemutatjuk a használt vizsgálati módszert, az alkalmazott adatbázisokat. Majd rövid áttekintést adunk az általunk alkalmazott zajszűrő eljárásokról. Utána pedig a két tesztsorozat módszerét, valamint azok eredményeit ismertetjük.

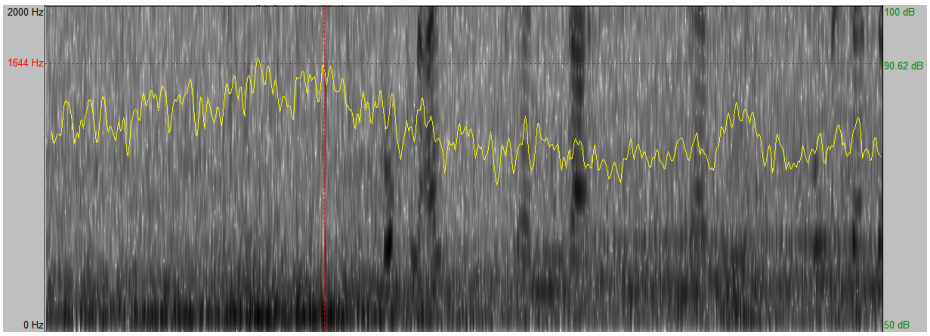
2 Vizsgálati módszer leírása

A kísérletek során két adatbázist használtunk fel:

A TELEUTO projekt kapcsán elkészült német nyelvű, autók belső terében készült hanganyag (TELEAUTO-német) négy mikrofonnal lett (2 szimmetrikusan elhelyezkedő vezetőülés melletti, 1 középső tükörnél lévő, és 1 headset mikrofon) rögzítve. Navigációs rendszereknél általában használt 400 darab rövid, megerősítés jellegű mondatot, 1386 rövid kérést kifejező mondatot, valamint 1432 hosszabb mondatot tartalmaz. A hanganyagot magyar anyanyelvű személyek mondták be német nyelven. A felvételek az autó haladása közben készültek, változó zajos környezetben, 48kHz-es sávszélességgel, 16 bites számábrázolással.

A jelen vizsgálatban csak egymikrofonos, a középső tükörnél lévő mikrofonnal készült felvételeket használtuk. A felvételek során a jel-zaj viszony -10 és 10 dB között változott. Gyakran előfordult, hogy a zaj átlagos intenzitásszintje nagyobb volt, mint

a beszédé. Az 1. ábrán egy zajos felvétel látható az adatbázisból. A piros vonalak egy tisztán zaj részt mutatnak. Jól látható, hogy ennek intenzitása (sárga görbe) nagyobb, mint ami a későbbi beszédnél tapasztalható.



1. ábra. Egy felvétel spektrogrammja a TELEAUTO-német adatbázisból.

A másik adatbázis a német nyelvű SpeechDat(II)-német adatbázis [4] volt, amely közel 5000 vezetékes telefonfelvételt és 1400 mobilhálózaton keresztüli beszélgetés felvételét tartalmazza, amelyek a telefon sávszélességének megfelelően 8000 Hz-es sávszélességűek és 16 bites számábrázolásúak. Ez a hanganyag mobilhálózaton keresztül már anyanyelvi német beszélőkkel készült, de nem személygépkocsikban került rögzítésre és ez igen lényeges akusztikai különbség a két adatbázis felvételei között. A tesztelésnél, a következtetések levonásánál ezt mindenképp figyelembe kell venni. Sajnos csak ilyen adatbázisok álltak a rendelkezésünkre a vizsgálathoz.

A felismerést Rejtett Markov Modell alapú felismerővel valósítottuk meg, amelyre a Sphinx szoftvert [5] használtuk. A modellek felépítéséhez az előfeldolgozás során a 8 kHz-es felvételeket 130Hz és 3700Hz közötti tartományra szűrtük, majd 25 ms-os Hamming ablakolást követően 512 pontos FFT-t számítottunk. A spektrumot kritikus sávok szerint szűrve MFC együtthatókká transzformáltuk, tehát a „klasszikus” 39 elemű jellemzővektorok lettek létrehozva (13 MFCC együttható, valamint ezek első és második deriváltja) 10 milliszekundumos kereteltolással. Ezekből 16 Gauss-keverékkel 3 állapotú trifón – tehát környezetfüggő – beszédhangmodellek készültek.

A tesztelésnél a 3. pontban részletezett zajszűrő eljárások alkalmazása után végzett felismerési kísérleteket hasonlítottuk össze egymással és a szűretlen zajos beszédfelvételekkel kapott felismerési eredményekkel.

3 Zajszűrő eljárások

Az autó belső terében készült zajos felvételek minőségének javításához, a beszédjel kiemeléséhez, a beszéd érthetőbbé tételéhez a következőkben szereplő zajszűrő eljárásokat alkalmaztuk. Minden eljárásnál használtunk egy felüláteresztő szűrőt is, amellyel az autó mélyfrekvenciás zaját tudtuk kiküszöbölni.

3.1 Wiener-szűrő [3]

A Wiener-szűrők központi szerepet játszanak számos alkalmazásban, például lineáris predikció, jelkódolás, visszhang kioltás, jelvisszaállítás és csatornakiegyenlítés megoldásaiban. A Wiener szűrőt úgy számoljuk, hogy a szűrő kimenete és a kívánt jel átlagos négyzetes távolsága minimális legyen. Általános esetben a szűrő azt feltételezi, hogy a jelek stacionárius folyamatok, ám a szűrő együtthatóinak időnkénti újraszámolásával adaptívvá tehető a jel karakterisztikájához. A Wiener-szűrő véges impulzus válaszú (FIR) szűrőként való megvalósítása adott számú lineáris egyenletet ad, amelyeknek létezik zárt alakú megoldása. A 2. ábra a Wiener-szűrőt ábrázolja, a $\boldsymbol{w}$ együttható vektorral, az $y(m)$ bemenő jellel, és az $\hat{x}(m)$ kimenő jellel, amely $\hat{x}(m)$ a kívánt cél $x(m)$ legkisebb átlagos négyzetes becslése. A szűrő bemeneti-kimeneti összefüggése:

$$\hat{x}(m) = \sum_{k=0}^{P-1} w_k y(m-k) = \boldsymbol{w}^T \boldsymbol{y} \tag{1}$$

A becsült és a cél jel közötti különbségből adódó hiba:

$$e(m) = x(m) - \hat{x}(m) = x(m) - \boldsymbol{w}^T \boldsymbol{y} \tag{2}$$

A legkisebb négyzetes hibájú Wiener-szűrőt a következő egyenlet alapján kapjuk meg:

$$\boldsymbol{R}_{yy}\boldsymbol{w} = \boldsymbol{r}_{yx} \tag{3}$$

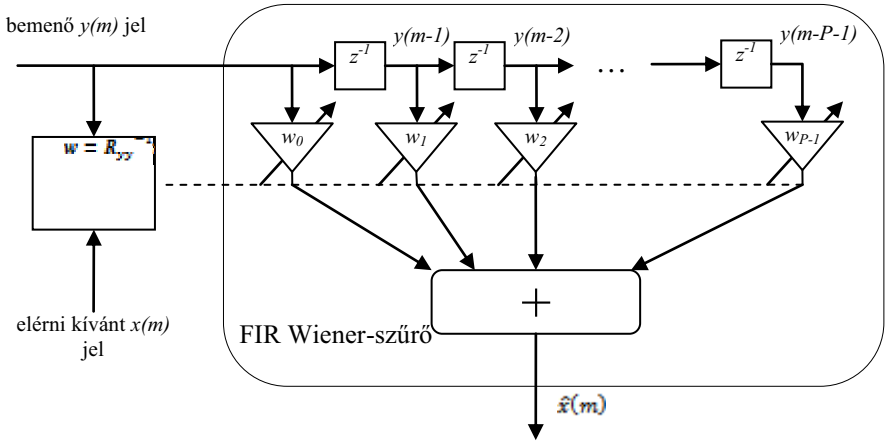
ahol szükségünk van a bemenő jel autókorrelációs mátrixára, valamint a bemenő és a kívánt jel keresztkorrelációs vektorára. Mivel az additív zajjal terhelt jel felírható $y(m) = x(m) + n(m)$ formában, a jel és a zaj korrelálatlanságából adódóan az autókorrelációs mátrix és a keresztkorrelációs vektor felírható

$$\boldsymbol{R}_{yy} = \boldsymbol{R}_{xx} + \boldsymbol{R}_{nn}, \text{ és} \tag{4}$$

$$\boldsymbol{r}_{xy} = \boldsymbol{r}_{xx} \tag{5}$$

formában, ahol az $\boldsymbol{R}_{yy}, \boldsymbol{R}_{xx}, \boldsymbol{R}_{nn}$ a zajos jel, a zajmentes jel és a zaj autókorrelációs mátrixai, $\boldsymbol{r}_{xy}$ pedig a zajos jel és a zajmentes jel keresztkorrelációs vektora. A (3), (4) és (5) egyenletekből a következő összefüggés adódik a Wiener-szűrő meghatározására:

$$\boldsymbol{w} = (\boldsymbol{R}_{xx} + \boldsymbol{R}_{nn})^{-1}\boldsymbol{r}_{xx} \tag{6}$$



2. ábra. A Wiener-szűrő felépítésének illusztrációja.

3.2 A „Spectral Subtraction" eljárás [2]

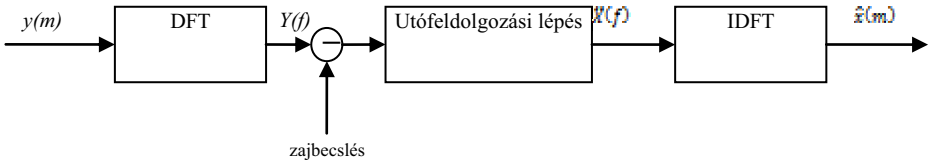
A spectral subtraction egy olyan eljárás, amely alkalmas egy additív zajban megfigyelt jel teljesítmény-spektrumának, illetve magnitúdó-spektrumának visszaállítására a zajos jelből. A zaj spektrumát becsülni lehet olyan időszeletekből, amikor nincs értékelhető jel, csupán a zaj van jelen. A spectral subtraction számításigénye kicsi, ám a zaj hirtelen változásai negatív teljesítmény-, illetve magnitúdó-spektrumot eredményezhetnek, amelyeket nemnegatív tartományba kell áttranszformálni. Ezen nemlinearitás a jelet torzítja.

A zajos jelet leírhatjuk $y(m) = x(m) + n(m)$-ként az időtartományban, ahol $y(m), x(m), n(m)$ a zajos jel, a jel és az additív zaj. Frekvenciatartományban a $Y(f) = X(f) + N(f)$ összefüggéssel írható le, ahol $Y(f), X(f), N(f)$ a zajos jel, az eredeti jel és a zaj Fourier-transzformáltjai. Az eljárás blokkdiagramja a 3. ábrán látható, ahol a spectral subtraction-t megvalósító egyenlet:

$$\left|\widehat{X(f)}\right|^b = |Y(f)|^b - \alpha\overline{|N(f)|^b} \tag{7}$$

A negatív teljesítmény- és magnitúdóspektrum kiküszöbölése érdekében egy utófeldolgozási lépést teszünk az inverz Fourier-transzformáció elé:

$$|\hat{X}(f)| = \begin{cases} |\hat{X}(f)|, \text{ ha } |\hat{X}(f)| > \beta|Y(f)| \\ fn[|Y(f)|], \text{ különben} \end{cases} \tag{8}$$



3. ábra. A spectral subtraction eljárás blokkdiagramja.

3.3 A „Minimum Mean-Square Error Spectral Amplitude Estimator"(MMSE) eljárás [6][7]

A spectral subtraction módszeren alapuló eljárások esetén a rövid távú spektrális amplitúdó (STSA) a jel spektrális komponenseinek varianciájának maximum likelihood becslő négyzetgyökeként adódik. A Wiener-szűrőknél az STSA becslő a jel spektrum-komponenseinek optimális legkisebb átlagos négyzetes hibájaként kapható meg. Mivel mindkét STSA becslő eljárás adott körülmények között kínál optimális megoldást, egyik sem általános optimális spektrális amplitúdó becslő. Ezért a következő eljárás, amelyet kipróbáltunk olyan, amely az STSA becslést közvetlenül a zajos megfigyelésből nyeri.

Jelölje $x(t)$ és $d(t)$ a beszéd és a zaj folyamatát, valamint $y(t)$ a megfigyelt jelet.

$$y(t) = x(t) + z(t) \tag{9}$$

Jelölje $X_k \triangleq A_k exp(j\alpha_k), D_k$ és $Y_k \triangleq R_k exp(j\vartheta_k)$ a beszédjel, a zaj és a megfigyelt jel *k.* spektrális összetevőjét. Y_k megadható a következőképpen is:

$$Y_k = \frac{1}{T}\int_0^T y(t) exp\left(-j\frac{2\pi}{T}kt\right)dt, \quad k = 0, \pm 1, \pm 2 \ldots \tag{10}$$

Az eljárás célja az A_k tényező becslése:

$$\hat{A}_k = E(A_k|Y_k) = \Gamma(1{,}5)\frac{\sqrt{v_k}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right]R \tag{11}$$

ahol $\Gamma(\cdot)$ jelöli a gamma függvényt, I_0 és I_1 jelöli a nullad-, és elsőrendű módosított Bessel-függvényeket, $\lambda_d(k) = E\{|D_k|^2\}$ a zaj varianciája, v_k definíciója pedig

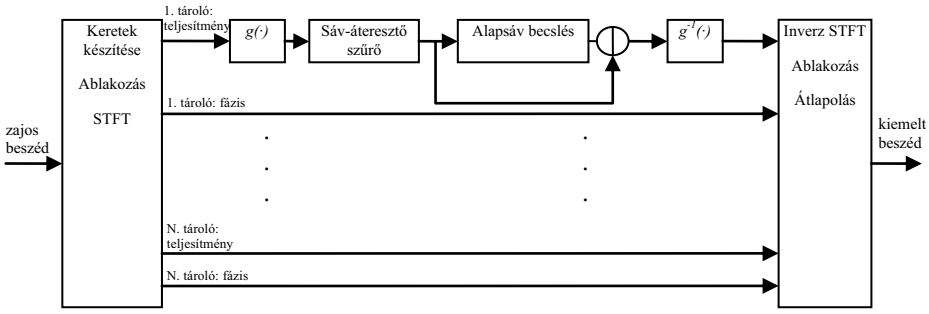
$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \text{ ahol } \xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \text{ és } \gamma_k = \frac{R_k^2}{\lambda_d(k)} \tag{12}$$

3.4 Modulációs spektrum alapú eljárás [1]

A modulációs spektrum a teljesítményspektrum (vagy annak egy tömörített változatának) komponenseiből felépített időbeli sorozatnak Fourier-transzformáltja. A kü-

lönböző modulációs spektrumok beszédérthetőségre gyakorolt hatását számos kutató vizsgálta, és általánosságban megállapították, hogy az 1 Hz és 16 Hz közötti modulációs spektrum játssza a legnagyobb szerepet az érthetőségben.

A 4. ábra illusztrálja a modulációs spektrum szűrésén alapuló zajszűrési eljárást. Az *x(t)* bemenő jelet keretenként 20 ms-os Hamming ablakkal és 10 ms-os időléptékkel elemzik. Jelölje $X_k(f)$ a short-time Fourier-transzformációt az *f.* frekvenciatárolóban. Definiáljuk a rövid távú teljesítményspektrumot $P_k(f) = |X_k(f)|^2$ -ként. Legyen *N* az FFT tárolók száma, valamint *K* a zajos beszéd megfigyelések keretszáma. Ekkor $P_k(f), k = 1, \dots K$ jelöli az *f.* frekvenciatároló idősorozatát. A modulációs szűrés során sávátereszto szűrőt alkalmazunk a $g\left(P_k(f)\right), k = 1 \dots K$ idősorozatra $f = 1 \dots N$ szerint. A 301 hosszúságú Parks-Mclellan módszerrel tervezett FIR szűrőt 1 és 16 Hz közötti tartományon alkalmaztuk. A $g(\cdot)$ egy tömörítő függvény, amelyet a teljesítményspektrumon alkalmaztunk a dinamika tartomány csökkentése érdekében. A tömörített időbeli burkológörbe alapsávú komponensét becsüljük a sávszűrt komponensből. A tömörítő függvény $g^{-1}(\cdot)$ inverzét alkalmazzuk a tömörítő hatás visszaállítására. A kapott közelítő burkolót felhasználva áll elő a rövid távú magnitúdóspektrum. Ezt a módosított magnitúdóspektrumot és az eredeti rövid távú fázisspektrumot használják fel a feljavított jel előállításához inverz FFT, ablakozás és átlapolás segítségével.



4. ábra. A modulációs szűrésen alapuló beszédkiemelő eljárás blokkdiagramja.

4 Zajszűrési eljárások kiértékelése

4.1 Tesztelési eljárások

A teszteket két adatbázissal, a TELEAUTO-német és a SPEECHDAT(II)-német adatbázissal végeztük, és a két adatbázis együttes kiértékeléséből vontuk le a következtetéseinket. Az autó zajának kiküszöböléséhez a zajszűrő eljárás mellett még egy felülátereszto szűrőt is alkalmaztunk, a mélyfrekvenciák eltávolítása érdekében. Így a fentebb bemutatott zajszűrő eljárásokat össze tudtuk hasonlítani aszerint, hogy beszédfelismerésre mennyire alkalmasak.

Az első fázisban a kísérletek során a TELEAUTO-német adatbázisa került felhasználásra. Az egyes zajszűrő eljárások tesztelése során az adatbázis három különböző hosszúságú mondatot tartalmazó részét két részre bontottuk és részenként 70%-ot használtunk a felismerő betanítására, a maradék 30%-ot pedig a felismerési hanganyagra. A tanítás előtt a felvételeken elvégeztük az adott zajszűrést, majd a tanítást már ezekkel a mintákkal kezdtük el.

Minden szűrő eljárást a változó zajhoz adaptálva alkalmaztunk a mondatok előtt vagy után található, csak zajból álló jel felhasználásával.

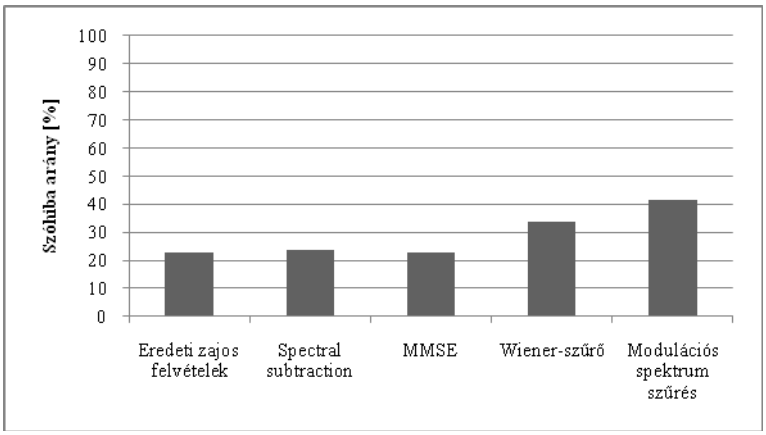
Az anyanyelvi német autós beszéd megfelelő biztonságú felismerése anyanyelvi német beszédadatbázis használatát követeli meg. Ezért a tesztelések második fázisában a tanításnál az anyanyelvi német SpeechDat(II) adatbázist használtuk fel. A teszteléskor továbbra is az előző fázisban felhasznált, autóban felvett hanganyaggal megegyező mintahalmazt használtunk már a megfelelő zajszűrő eljárás lefuttatása után. Az anyanyelvi különbségekből adódó alapvető kiejtésbeli eltérés miatt ezeknél a felismerési kísérleteknél az abszolút felismerési teljesítmény nem lehet mérvadó. Ezért itt mindig az eredeti zajos felvételekkel történő felismeréshez viszonyítottuk a különböző zajszűrő eljárásokkal kapott felismerési eredményt.

4.2 Beszédfelismerési eredmények

Az 1. táblázatban és 5. ábrán az első tesztsorozat eredményei láthatóak. Az 1. táblázat részletesen tartalmazza a TELEAUTO adatbázis három mondattípusával kapott felismerést. Az elvártaknak megfelelően az egyik legjobb felismerési eredményt akkor kaptuk, ha a betanítás és a tesztelés is az eredeti, zajos felvételekkel történt. Ehhez képest legjobb zajszűrők által elért felismerést az MMSE Spektrális Amplitúdó Becslő, valamint Spectral Subtraction eljárás adta, amelyek csupán legfeljebb 1%-kal tértek el az eredeti, zajos felvételekkel történő tesztek eredményétől, valamint közel 10%-kal jobb felismerést produkáltak, mint a további két zajszűrő módszer. Ezek a módszerek tehát egy felismerési feladat során alkalmazhatónak adódtak.

1. táblázat: Az első fázis tesztsorozatai során kapott felismeréseinek szóhiba arányai százalékban.

	Rövid, megerősítés jellegű mondatok	**Kérés jellegű mondatok**	**Hosszú, általános mondatok**	**Átlagos felismerés**
Eredeti zajos felvételek	18,5	15,3	34,4	**22,7**
Spectral subtraction	22,6	15,1	33,3	**23,7**
MMSE	17,8	16	34,1	**22,6**
Wiener-szűrő	35,6	19,5	46,7	**33,9**
Modulációs spektrum szűrés	37,7	28	59,1	**41,6**

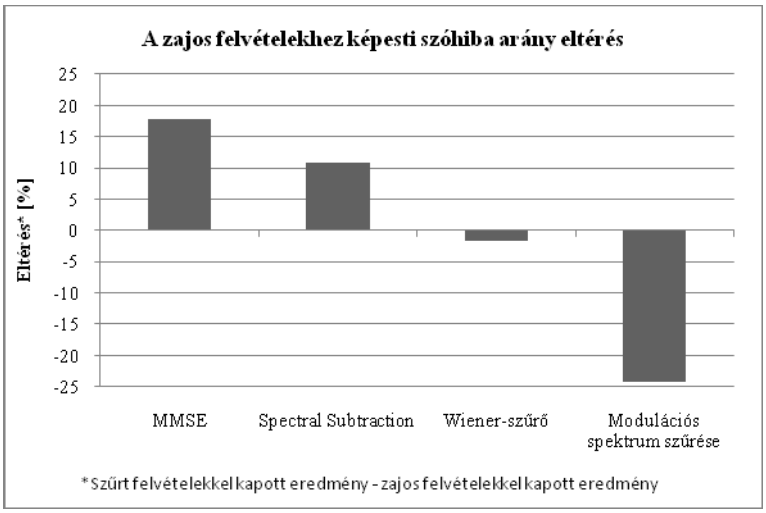


5. ábra. A TELEAUTO adatbázissal készített felismerés eredményei: az eredeti zajos felvételeket, valamint az adott zajszűrő eljárást alkalmazva a tanításra és tesztelésre

A 6. ábrán a második tesztsorozat eredményei láthatóak. Az ábra a szűretlen hangmintákkal végzett felismerési teljesítményhez képesti javulást vagy romlást mutatja az egyes zajszűrő eljárások esetében. A 2. táblázatban részletesen láthatóak az egyes mondattípusokkal kapott eredmények. Látható, hogy azok az eljárások, amelyek az első tesztsorozat esetében jól teljesítettek, itt is hasonló tulajdonságot mutatnak, de itt az MMSE kicsit kiemelkedik a többi közül.

2. táblázat: A német nyelvű SpeechDat adatbázissal végzett tanítás és a TELEAUTO adatbázisával végzett tesztelés során az adott zajszűrő eljárás és az eredeti zajos felvételek felismerési szóhiba arányai közötti eltérés százalékban.

	Rövid, megerősítés jellegű mondatok	**Kérés jellegű mondatok**	**Hosszú, általános mondatok**	**Átlagos eltérés**
MMSE	32,2	14,1	7,3	**17,8**
Spectral Subtraction	31,5	2,1	-0,9	**10,9**
Wiener-szűrő	10,9	-9,1	-6,5	**-1,6**
Modulációs spektrum szűrése	-25,4	-39,2	-7,9	**-24,2**



6. ábra. A német nyelvű SpeechDat adatbázissal végzett tanítás és a TELEAUTO adatbázisával végzett tesztelés során az adott zajszűrő eljárás és az eredeti zajos felvételek felismerési teljesítménye közötti eltérés.

A James G. Lyons, Kuldip K. Paliwal-féle modulációs spektrum alapú eljárás a vártnál lényegesen rosszabb felismerési eredményt adott mind a két adatbázissal végzett betanítás esetén. Annak ellenére, hogy az irodalomban lényeges szubjektív érthetőségnövekedésről számolnak be a kutatók e szűrés alkalmazása esetén.

5 Összefoglalás

A cikkben olyan zajszűrő eljárásokat hasonlítottunk össze, amelyek alkalmasak additív zaj szűrésére, a hasznos jel (beszéd) kiemelésére. Az összehasonlítás során személygépkocsiban felvett időben változó zajkörnyezetű folyamatos beszéd felismerését vizsgáltuk.

Az eljárások két fázisban kerültek tesztelésre. Az első fázisban ugyanazon személygépkocsi belső terében felvett hanganyaggal történt a betanítás is és a tesztelés is, a TELEAUTO-német adatbázissal. Ezekben az esetekben a szóhibaarány 20-25% közöttinek adódott a -10 és 10 dB közötti jel-zaj viszony határok között. Ez azt mutatja, hogy szűrés nélkül is a legjobb szűréssel kapott eredményhez közeli elfogadható eredményt kapunk abban az esetben, ha az akusztikus modell betanítása hasonló zajos körülmények között történik, mint ami a felismeréskor is előfordul.

Abban az esetben, amikor nem áll rendelkezésre megfelelő zajos adatbázis a betanításhoz, amit a kísérletben a SPEECHDAT(II) német adatbázis használatával modelleztünk, egyes zajszűrő eljárások a 4. fejezetben tárgyalt kísérletek szerint adaptív módon sikerrel alkalmazhatók személygépkocsikban adódó zajkörnyezetben.

A legjobban teljesítő zajszűrő eljárás a Minimum Mean-Square Error Spectral Amplitude Estimator (MMSE) volt, amelynek használatával közel 18%-kal adott jobb

felismerési eredményt, mint az eredeti zajos felvételek esetén. Az egyes módszerek egy beszéddetektor segítségével automatikusan adaptívvá tehetők.

Köszönetnyilvánítás

Ez a kutatás a Jedlik OM-00102/2007 számú "TELEAUTO" projekt keretén belül készült.

Hivatkozások

1. Lyons J. G., Paliwal K. K.: Effect of Compressing the Dynamic Range of the Power Spectrum in Modulation Filtering Based Speech Enhancement. Interspeech 2008 (2008) 387–390
2. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of Speech Corrupted by Acoustic Noise. IEEE ICASSP (1979) 208–212
3. Vaseghi, S. V.: Advanced Signal Processing and Digital Noise Reduction. Wiley & Teubner Communications (1996)
4. SpeechDat.: http://www.speechdat.org/
5. Sphinx.: http://cmusphinx.sourceforge.net/html/cmusphinx.php.
6. Yariv, E., Malah, D.: Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing (1984) 1109–1121
7. Yariv, E., Malah, D.: Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing (1985) 443–446

Beszédfelismerési kísérletek hangoskönyvekkel

Tóth László

MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport
tothl@inf.u-szeged.hu

Kivonat: Valós körülmények között a gépi beszédfelismerést számos tényező nehezíti, például a háttérzaj, a beszélő hangjának egyéni sajátosságai, a spontán artikuláció vagy a beszéd érzelmi töltete. A gyakorlatban is alkalmazható felismerőrendszerek készítéséhez természetesen ezeket a problémákat mind tudni kell kezelni, egyelőre azonban a jóval egyszerűbb feladatokat sem tudjuk tökéletesen megoldani. Jelen cikkben annak megvizsgálása a célkitűzésünk, hogy vajon mire képes a jelenlegi technológia „ideális" körülmények között. Az optimális viszonyok szimulálásához egy hangoskönyv hanganyagával dolgozunk, mivel ennek rögzítése során az említett hátráltató tényezők többsége nem, vagy csak minimális mértékben jelentkezik. A kiértékelés segítése érdekében a kapott eredményeket egy korábbi, telefonos adatbázison végzett hasonló kísérletsorozat eredményeivel állítjuk párhuzamba. Méréseink szerint a hangoskönyvön kapott fonetikai kimenet pontossága már minimális nyelvi támogatással is 86% fölött van, és emberi szemmel is majdnem tökéletesen olvasható.

1 Bevezetés

A piac használható alkalmazások iránti igénye a beszédtechnológiai kutatást erőteljesen kényszeríti az egyre nehezebb, komplexebb problémák irányába – példa erre a zajos beszéd felismerése iránti igény vagy az utóbbi időben a természetes, spontán beszéd vizsgálatának fókuszba kerülése. A piaci elvárások persze jogosak olyan értelemben, hogy a gyakorlati használhatósághoz valóban túl kell lépni a csak steril laboratóriumi körülmények között működő rendszereken. Ez nincs alapvető ellentmondásban a kutatók vágyaival, hiszen végcélnak ők is a teljesen kötetlen beszéd felismerését tekintik. A baj inkább az, hogy egyelőre még az egyszerűbb, „redukált" felismerési feladatok sincsenek teljesen megoldva, így az ipar egyfajta „előremenekülésre" kényszeríti a kutatókat – miközben a problémamegoldás íratlan szabályai sokkal inkább az egyszerűbb feladatokra való visszalépést írnák elő. Épp ezért azt gondoljuk, hogy nem szabad abbahagyni az egyszerűsített felismerési szituációk vizsgálatát sem, mivel a nehézségeket okozó tényezőket szétválasztva könnyebb azokat elemezni és megérteni. Az egyszerűbb, könnyebb feladatok vizsgálatát továbbá azért sem érdemes feladni, más számos olyan értelmes alkalmazás létezik, ahol ezeknek is létjogosultságuk lehet (például egy rádióhíreket figyelő vagy TV-híradót feliratozó rendszer esetén mind a stúdióminőségű felvétel, mind a fegyelmezett artikuláció feltételezhető).

Jelen cikkünkben hangoskönyveken végzünk beszédfelismerési kísérleteket. A tesztekkel annak megvizsgálása a célunk, hogy a jelenlegi beszédfelismerők (főleg az akusztikus komponens) mire lennének képesek ideális körülmények közt, vagyis ha a zavaró tényezők nagy részét ki tudnánk zárni. A hangoskönyvek tartalma „ideális" beszédnek tekinthető olyan értelemben, hogy a beszédfelismerést valós helyzetben megnehezítő tényezők közül a legtöbb nem jelentkezik a hanganyagukban. A 2. fejezetben áttekintjük ezeket a tényezőket, és megpróbáljuk érzékeltetni a beszédfelismerőkre tett hatásukat. Az érzékeltetést fogja szolgálni az is, hogy felismerési eredményeinket párhuzamba állítjuk a 2008-as Interspeech konferencián publikált értékekkel, melyeket ugyanazon felismerési technikával értünk el, de az MTBA telefonbeszéd-adatbázison. Az eredmények 5. fejezetbeli közlése előtt azonban természetesen részletesen ismertetjük a hanganyag feldolgozásának lépéseit a 3., majd az alkalmazott ún. „tandem" felismerési technológiát a 4. fejezetben. Cikkünk az eredmények elemzésével és a következmények levonásával zárul a 6-7. fejezetekben.

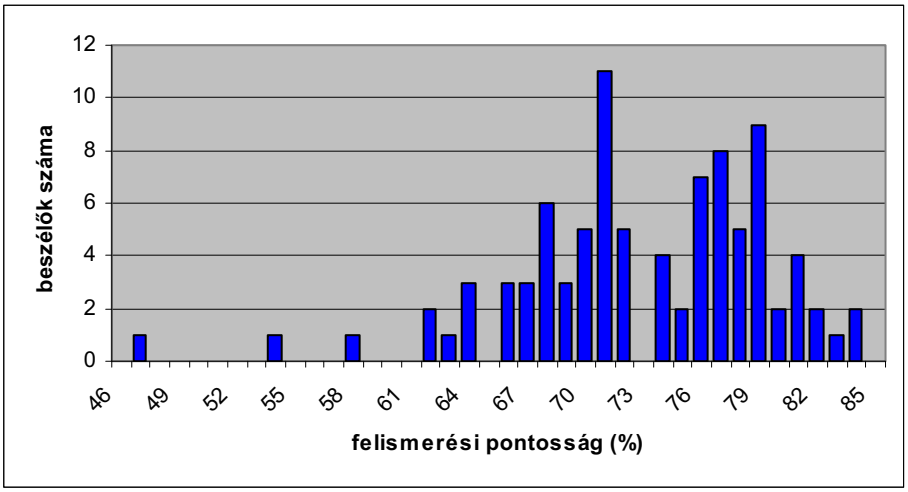
2 A beszédfelismerést megnehezítő tényezők

Az alábbiakban áttekintjük a beszédfelismerést valós szituációkban megnehezítő fő tényezőket, és az irodalomból vett adatokkal kíséreljük meg érzékeltetni jelentőségüket. Megvizsgáljuk továbbá, hogy a hangoskönyvre és az összehasonlítási alapként szolgáló MTBA adatbázisra az adott tényező milyen mértékben jellemző.

Gyakorlati körülmények lényegében nincs olyan helyzet, amelyben a háttérzaj beszűrődése teljesen megakadályozható lenne. Tapasztalataink szerint még a híradók stúdióban rögzített felvételein is akad háttérzaj, behallatszik például, ahogy a bemondó a papírjait rendezgeti. És sok olyan alkalmazás van, amely kifejezetten erős háttérzaj mellett kísérel meg beszédfelismerést használni (pl. egy vadászgép pilótafülkéjében). Hagyományosan zajnak (ún. konvolutív zaj) tekintjük továbbá az átviteli közeg (pl. telefonvonal) okozta torzítást is, amely bizonyos frekvenciakomponensek gyengülését-erősödését okozza. A háttérzaj a beszédfelismerők felismerési pontosságát drasztikusan csökkenti, főleg ha az emberi beszédpercepcióval párhuzamba állítva vizsgáljuk [6]. A konvolutív zaj hatása még kiábrándítóbb, ugyanis ezt mi emberek szinte nem is érzékeljük (legfeljebb a hangszín változásaként), miközben a felismerők hatékonyságát meglepő fokban le tudja rontani. Ezt jól példázza, hogy ha diktálószoftvert vásárolunk, általában mikrofont is kapunk hozzá, mivel már pusztán másik mikrofon használata is érzékelhető teljesítménycsökkenéssel járhatna. Esetünkben az összehasonlítási alapként szolgáló MTBA adatbázis különféle vonalakon rögzített telefonos felvételeket tartalmaz, a telefon szokásos torzításával és frekvenciavágásával. Háttérzaj is beszűrődik a felvételekbe, bár tapasztalataink szerint viszonylag ritkán (az adatközlők érzékelhetően nyugodt körülményeket választottak a híváshoz). Ezzel szemben a feldolgozott hangoskönyv stúdióban készült, így háttérzajt gyakorlatilag nem tartalmaz, és feltehetően professzionális mikrofonnal vették fel (bár azt nem tudhatjuk, hogy végig ugyanazzal-e).

A gépi beszédfelismerők közismerten érzékenyek a beszélő személyére, azaz az egyes beszélők hangja közt adódó eltérésekre. Az MTBA adatbázis 500 adatközlő felvételeit tartalmazza, és mindenkitől csak 12-12 mondatot, így a beszélő személye

igen gyakran változik. Az 1. ábrán bemutatott hisztogram a különböző adatközlőkre kapott felismerési pontosságok szórását érzékelteti egy konkrét, az MTBA adatbázison végzett kísérlet esetén. Látható, hogy a 74%-os átlaghoz képest a 10-10% körüli kitérés sem ritka egyik irányban sem, sőt, a legjobb és legrosszabb beszélő közti különbség több mint 36%! Habár az egyes felvételek közt nem csak a beszélő személye, hanem a telefonvonal, így a zajviszonyok is változnak, úgy véljük, hogy a kapott nagy szórást alapvetően a beszélők közti eltérések okozzák (mint mondtuk, a felvételek zajszintje jellemzően alacsony). Az MTBA 500 beszélőjével szemben a bemutatandó kísérletekben feldolgozott hangoskönyvet egyetlen ember olvassa fel, így a beszélőváltás mint zavaró tényező teljesen ki lesz zárva.



1. ábra. Beszédhang-felismerési pontosság eloszlása az MTBA adatbázison a beszélő személy függvényében.

Mivel a beszédadatbázisok sokáig úgy készültek, hogy kísérleti alanyokat kértek fel valamely szöveganyag felolvasására, így viszonylag későn tudatosult a kutatókban, hogy milyen jelentős eltérések vannak az olvasott és a spontán beszéd artikulációja között. Eleinte csak az tűnt fel, hogy a laboratóriumi körülmények közt elfogadhatóan működő felismerők a gyakorlatban sokkal rosszabbul teljesítenek, de csak az utóbbi 5-10 évben kezdték el a spontán beszéd jellegzetességeit közelebbről tanulmányozni. Hogy kézzelfogható értékeket is mondjunk, végeztek például olyan tesztet, melyben egy tárgyaláson felvett hanganyagot újraolvastattak ugyanazon résztvevőkkel. Az olvasott és a spontán felvételeken mért felismerési hiba között közel kétszeres faktort kaptak [12]. Magyar nyelvre Mihajlik és társai próbálkoztak spontán és tervezett beszéd (hírműsorok) ugyanazon technológiával való felismerésével [8]. Habár az eredmények nem precízen összemérhetők, hiszen a két feladat közt a beszédmódon kívül más eltérések is voltak, az általuk kapott bő kétszeres hibatényező is jól érzékelteti, hogy milyen jelentős hatékonyságromlás lép fel spontán beszéd esetén. Ez a hatékonyságromlás épp elég ahhoz, hogy a felismerők átessenek az „éppen használható” kategóriából a használhatatlanba, ezért olyan megoldással is találkozni – például egy japán tévéműsor-feliratozó rendszerben –, hogy a zajos vagy spontán részeket egy képzett beszélő megfelelő artikulációval újramondja [14]. Esetünkben mindkét

adatbázis olvasott beszédet tartalmaz, de míg a hangoskönyveket színészek olvassák lemezre, az MTBA adatbázisban bőven akadnak igénytelen beszédmódú adatközlők. Így ebből a szempontból is könnyebbnek ígérkezik a hangoskönyvek felismerése, habár az MTBA sem tartozik a legnehezebb (azaz spontán beszéd) kategóriába.

Egyetlen embertől származó hangfelvétel esetén is változhat a beszéd hangminősége, akár fizikai (pl. rekedtség), akár lelki okokból (pl. érzelmi felindultság). Ebből a szempontból talán kivételesen a hangoskönyv a rosszabb, az MTBA esetében ugyanis olyan rövid hanganyagunk van egy-egy beszélőtől, hogy ezt a jelenséget nem igazán van mód megfigyelni. Egy hangoskönyvben természetesen előfordulhat, hogy a színész hangszínének megváltoztatását kifejezőeszközként használja, de az általunk választott felvétel esetén ez kevésbé jellemző. Néhol fordul csak elő egyfajta suttogás jellegű, visszamerengő beszédstílus.

3 A hanganyag és feldolgozása

A viszonyítási alapként közölt felismerési eredményeket az MTBA adatbázison értük el, és részben már publikáltuk korábban [9]. Az MTBA adatbázisról is közöltünk már részletes leírást [10]; mint már kiderült, ez egy telefonon át rögzített korpusz, mely 500 beszélőtől tartalmaz felvételeket, melyekből mi itt az olvasott mondatokat és szavakat tartalmazó blokkot használtuk fel. A teljes adatbázis manuális fonetikai szegmentáláson és címkézésen esett át; az ennek során használt 58 címkéből viszont némelyik olyan ritkán fordul elő, hogy kénytelenek voltunk néhány összevonást eszközölni, így a kísérletekben 52 címkével dolgoztunk. A felvételekből elhagytunk bizonyos, a kézi címkézés során zajosnak talált felvételeket, így az eredeti 8000 fájl helyett csak 6935-öt használtunk fel. Ezt úgy osztottuk fel tanító és tesztelő részre, hogy előbbibe 408, utóbbiba 91 beszélő került (1 beszélő esetén az összes felvétel túl zajosnak bizonyult).

Feldolgozandó hangoskönyvnek olyan felvételt választottunk, amelynek eredeti, írott változata is jogdíjmentesen elérhető. Választásunk Krúdy Gyula Szindbád-történeteinek „Szindbád utazásai” című gyűjteményes kiadására esett, Gáspár Sándor előadásában (Kossuth kiadó – Mojzer kiadó). A felvétel teljes játékideje 212 perc, ami körülbelül fele az MTBA adatbázis időtartamának. A hanganyagot szinkronba kellett hoznunk a szöveganyaggal, ennek lépéseit ismertetjük az alábbiakban. Először is a hanganyagot végighallgattuk, a szöveghez képesti esetleges eltéréseket keresve. Ilyet kb. tucatnyi esetben találtunk csak, és viszonylag rövid szavakat érintve (többnyire indulatszavak, pl. „óh”, „ah” elhagyása vagy beszúrása a felolvasó által). A lehallgatás során vágtuk ki az egyes fejezetek végén elhangzó zenei szignált, valamint az idegen szavakat is kigyűjtöttük a fonetikai átírás előkészítéseként.

Mivel az MTBA-éhoz hasonló fonetikai szintű címkézést szerettünk volna készíteni a hangoskönyvhöz, így a következő lépés a szöveganyag fonetikai átírása lett volna. Erre egy elég sajátos megoldást alkalmaztunk, több szempontot is figyelembe véve. A szokványos út az előforduló szóalakok kigyűjtése, majd azok átírása. Az átírás azonban nem triviális dolog, több okból sem [7]. Az egyik problémát a kettős betűk okozzák, melyek azonosításához morfológiai elemzésre lenne szükség (lásd pl. „pácsó”). A másik probléma, hogy bizonyos hasonulási folyamatok fellépése szintén

függ a morfémahatárok helyétől (erre példa a /tj/ kapcsolat az „látják”, illetve „átjáró” szavakban). Ráadásul a hasonulás sok esetben opcionális, azaz többféle ejtés is helyes lehet. Erre tényleg csak az a megoldás létezik, hogy az adott szóhoz több lehetséges kiejtést is megadunk. Tipikus ilyen opcionális hasonulási pozíció a szóhatár, ahol akár kis szünetet is tarthatunk, de kiejthetjük a szomszédos szavakat szünet nélkül is, sőt a szóvégi hangok hasonulásával is. Hogy melyik következik be, az leginkább az artikuláció igényességén múlik, azaz a szövegből többnyire megjósolhatatlan. A szóhatárokon fellépő hasonulásokat a szavak izoláltan történő átírásával dolgozó módszerek többnyire nem is képesek kezelni.

A fenti okból, valamint mivel nem állt rendelkezésünkre egy kifinomult, morfológiai elemzést is figyelembe vevő fonetikus átíró, a szavankénti átírás helyett egy mássalhangzó-kapcsolatokra épülő fonetikai átírást alkalmaztunk. Ehhez abból indultunk ki, hogy a szóköz csak az írott szövegben jelent triviális tagolási határt – az akusztikumban viszont a szóhatár az egyik legkiszámíthatatlanabbul viselkedő jelenség. Miért nem választunk hát inkább olyan tagolást, amelynek határai akusztikailag stabilak? Ebből kiindulva a szöveget nem a szóközöknél, hanem a magánhangzóknál tördeltük el. Egyrészt azért esett a magánhangzókra a választásunk, mert szép artikuláció esetén nem jellemző, hogy kiesnek vagy redukálódnak (a hossz módosulásától eltekintve). Másrészt pedig a hasonulás alapvetően a mássalhangzó-klasztereket érinti, a magánhangzókon nem terjed át, így egyfajta természetes határt képez. Előnyt jelentett továbbá az is, hogy mássalhangzó-kapcsolatból jóval kevesebb van, mint szóalakból: esetünkben a 7186 különböző szóalakhoz képest csak 809 különböző mássalhangzó-kapcsolatot találtunk (a szóhatárokon átívelő kapcsolatokat is beleértve!). Így az elemek automatikus, szabályalapú fonetikai átírása után az összes elemet át tudtuk nézni, és szükség esetén kézileg korrigálni. Ezzel a megoldással a szóhatárokat kényelmesen tudtuk kezelni, például a „T SZ” betűsorhoz három lehetséges átiratot rendeltünk:

t sil s
t s
ts:

ahol “*sil*” a csend fonetikai címkéje. A módszernek természetesen van egy olyan hátránya, hogy mivel a teljes szót nem látja, így olyankor is megenged alternatívákat, amikor nem kellene, például a *pácsó* szóhoz a helyes [pa:tʃo:] mellett a hibás [pa:tʃo:] átírást is fel fogja kínálni. Mindenesetre úgy gondoltuk, hogy ez kevésbé rontja a felismerő hatásfokát, mint ha egy szóhoz csak egyetlen, de esetleg hibás átirat van megengedve.

A fonetikai átírással kapott, a fentiekben ismertetett módon alternatívákat is megengedő szimbólumsorozatnak a hanganyaghoz való legjobb illeszkedését ún. kényszerített illesztéssel [5] határoztuk meg. Ehhez a HTK beszédfelismerő csomagot használtuk [13], melyet az MRBA adatbázison tanítottunk be. Ez az adatbázis szerkezetében nagyon hasonlít az MTBA-hoz, a lényeges különbség, hogy nem telefonvonalon, hanem személyi számítógépekbe dugott mikrofonokon keresztül rögzítettük [11]. Emiatt úgy éreztük, hogy felvételi körülményei jobban igazodnak a hangoskönyvéhez, és ezért talán megfelelőbb a feladathoz.

A kényszerített illesztés révén előállt annotált adatbázist kb. 80%-20% arányban osztottuk fel tanító és tesztelő részre, egész pontosan a hangoskönyv tíz Szindbád-történetéből nyolcat jelöltünk ki tanításra és kettőt tesztelésre.

4 Akusztikai modellezés a tandem technológiával

A hanganyag előfeldolgozása eltérőképpen zajlott a két adatbázis esetén, ugyanis az MTBA-s kísérletekben alkalmazkodnunk kellett egy angol rendszerhez [9]. Így ott PLP-vektorokkal reprezentáltuk a beszédjelet, míg a hangoskönyv esetében a szokványos 39 elemű kepsztrális (MFCC) együtthatóvektorok sorozatát nyertük ki [5]. Korábbi tapasztalataink alapján ez nem okoz nagy eltérést, egyik reprezentáció sem nevezhető szignifikánsan jobbnak a másiknál.

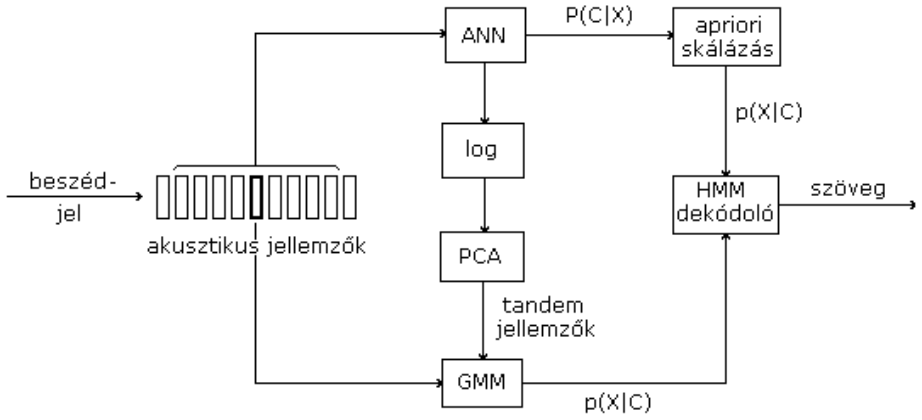
A hagyományos rejtett Markov-modelles (HMM) technológia a jellemzővektorok alapján, Gauss-keverékeloszlások illesztésével ad közelítést az egyes építőelemek (modell-állapotok) valószínűségére [5]. Mi egy másik fajta technikát használtuk, amely a gaussos modellek helyett mesterséges neuronhálót alkalmaz a lokális valószínűségek becslésére. Ez a megoldás két fő előnyt kínál a hagyományoshoz képest: egyrészt a neuronháló tanítása diszkriminatív – szemben a Gauss-keverékmodell hagyományos generatív tanításával –, ezért általában valamivel nagyobb osztályozási pontosságot tud elérni. Másrészt a neuronhálót általában nem csak egyetlen adatvektoron, hanem több (esetünkben 9) szomszédos vektoron szokták tanítani, a nagyobb környezet figyelembe vétele pedig szignifikáns javulást tud hozni. Meg kell jegyeznünk azonban, hogy a Gauss-keverékmodellhez is létezik diszkriminatív tanítási algoritmus, és annak sincs elvi akadálya, hogy több szomszédos vektoron tanítsák – egyszerűen csak valami oknál fogva ez nem terjedt el.

A neuronháló által adott kimenetek bizonyos kritériumok teljesülése esetén valószínűségi becslésként értelmezhetők, és egy apró módosítással beépíthetők a hagyományos HMM-sémába; így kapjuk az ún. HMM/ANN hibrid modellt [2]. A hibrid technológiát – főként kisebb feladatok esetén – sokan találták jobbnak, mint a hagyományos HMM-et, de a nagyobb rendszerekben mégsem bírt elterjedni. Saját tapasztalatunk az, hogy bár akusztikai szinten tényleg pontosabb, a nyelvi modellel kombinálva mégis leromlik a teljes rendszer hatékonysága. Ennek oka sejtésünk szerint az lehet, hogy a másfajta modellezési és tanítási technika miatt a neuronhálós akusztikus modellt máshogy kellene kombinálni a nyelvi modellel, mint ahogy azt a hagyományos HMM teszi.

Ezt a problémát egy huszárvágással oldja meg az ún. HMM/ANN tandem technológia [4]. Ez a neuronhálótól kapott értékeket nem valószínűségi becslésként értelmezi, hanem úgy tekinti, hogy a neuronháló egy nemlineáris transzformációt hajtott végre az akusztikus jellemzőkön; vagyis a kimenet továbbra is akusztikus jellemzővektor, pusztán egyfajta transzformált formában. Ez esetben viszont be lehet rajta tanítani egy teljesen hagyományos, Gauss-komponensekkel dolgozó rejtett Markov-modellt. Ezzel a trükkös megoldással azt mondhatjuk, hogy a rendszerben csak az akusztikai előfeldolgozó modult cseréltük le, így semmit nem kell módosítani a hagyományos, jól bevált és ezerszer letesztelt rejtett Markov-modellünkön. A megoldás

egyetlen hátránya az, hogy a rendszert duplán kell tanítani, és nyilván a kiértékeléskor is lassabb lesz.

A 2. ábra blokkdiagramja összefoglalja a hagyományos, a hibrid és a tandem modellek számítási lépéseit.



2. ábra. A hagyományos modell (alsó útvonal), a hibrid (felső útvonal) és a tandem modell (középső útvonal) sematikus összevetése.

Az elvi áttekintés után lássuk a tandem modell megvalósításának technikai részleteit. Az alkalmazott neuronháló 9 szomszédos jellemzővektoron tanult, kimenetként pedig az 52 fonetikai címke mindegyikéhez rendeltünk egy-egy kimenő neuront. A MTBA-n végzett tesztek során a rejtett réteg neuronjainak száma 4800 volt, ugyanis szinkronban kellett lennünk az említett angol modellel. A hangoskönyv esetén csupán 500 rejtett neuronnal dolgoztunk, mivel a neuronszám további növelése nagyobb futásidő-növekedéssel jár, mint amennyit az eredményeken javít. A neuronhálót mindkét esetben backpropagation algoritmussal tanítottuk be, az adatok 10%-án számított keresztvalidációt használva megállási kritériumként. A tanítási célértékeket természetesen a kényszerített illesztés során kapott fonetikai címkék képezték.

A neuronháló által kiadott vektorokon a HTK csomag rejtett Markov-modelljét tanítottuk be [13]. Akusztikus komponensként 3-állapotú monofón beszédhangmodelleket képeztünk, állapotonként 9-9 Gauss-eloszlással. Az irodalom javaslata szerint a neuronháló kimenő értékeit a HMM-be való beengedés előtt érdemes logaritmizálással Gauss-görbéhez jobban igazodó alakúra hozni, valamint főkomponens-analízissel dekorrelálni. Mi is így tettünk, ugyanis saját méréseink is alátámasztották az említett trükkök hasznosságát [9]. Egy további trükk a neuronhálókimeneteknek az eredeti akusztikai vektorokkal együtt való használata, azaz a két vektor konkatenálása. Habár a két vektor elvileg redundáns, a gyakorlatban egy minimális javulást ez a fogás is tud hozni, így mi is alkalmaztuk. Így összességében a HMM inputját képező jellemzővektor 91 komponensű volt a szokványos 39 helyett. A beszédhangmodellek tanításához a hagyományos, maximum-likelihood kritériumot optimalizáló algoritmusok mellett diszkriminatív (MMI-hibakritériumot alkalmazó)

tanítást is bevetettük [3]; szerencsére a HTK csomag tartalmazza ennek implementációját.

A novellák szöveganyagát kevésnek éreztük egy szószintű nyelvi modell (*N*-gram) betanításához, egy általános, kortárs korpuszokon tanított nyelvi modell pedig nem igazán illett volna a regény majd' száz éves szókincséhez. Ezért nyelvi modellként beszédhangszintű modellezéssel próbálkoztunk: a HTK eszköztárát használva a tanítókorpusz fonetikai címkéiből beszédhang-bigramokat számoltunk. Továbbá mivel a kényszerített illesztésnél alkalmazott módszer miatt rendelkezésünkre állt a szöveganyag magánhangzó-mássalhangzókapcsolat elemekre való felbontása, kézenfekvően adódott, hogy ezekből is megpróbáljunk bigramot képezni. Erre a nyelvi modellre jobb híján „szótag"-bigramként fogunk hivatkozni, bár az elemei csak méretükben hasonlítanak a nyelvészeti értelemben vett szótagokhoz.

5 Eredmények és diszkusszió

Legelső lépésként a rejtett Markov-modellt teljesen hagyományos módon, azaz közvetlenül az akusztikus jellemzővektorokon tanítottuk be. Az így kapott értékeket viszonyítási alapként használhatjuk a tandem-reprezentáció, azaz a neuronháló segítségével végzett transzformáció hasznosságának megítélésében. Az első tesztekben semmiféle nyelvi modellt nem használtunk, hogy az eredmények tisztán az akusztikus modellek hatékonyságát tükrözzék. Az MTBA adatbázison 53,37%-os, míg a hangoskönyv esetén 72,18%-os pontossággal egyezett a felismerő által kiadott és a címkézés szerint a fájlhoz tartozó leirat (pontosságon a két sztring szokványos, angol terminológiával „accuracy"-nek nevezett illeszkedését értve). Már magában ez az érték is jól mutatja, hogy a hangoskönyv mennyivel könnyebb felismerési feladatot jelent.

A következő lépés a hagyományos jellemzőkről a tandem jellemzőkre való áttérés volt. Ennek első fázisa a neuronháló betanítása az osztálycímkék felismerésére. Ennek eredményessége a rejtett Markov-modellbe való beépítés előtt is tesztelhető, bár ilyenkor persze még csak az egyes adatvektorokra vonatkozó osztályozási pontosságot tudjuk vizsgálni. A neuronháló 74,11%-os felismerést tudott elérni az MTBA esetén, míg a hangoskönyvön 85,24%-ot produkált. Mivel ezek a pusztán lokális értékek jóval magasabbak, mint a HMM-mel kapott globális eredmények, jó eséllyel várhattuk, hogy az ezekre épülő teljes modell is lényegesen jobb lesz.

A rejtett Markov-modell tandem jellemzőkkel történt betanítása után kapott eredményeket az 1. táblázat 2. sorában találhatjuk. Látható, hogy a tandem technikának köszönhetően mindkét adatbázison jelentősen, és körülbelül ugyanolyan mértékben (kb. 25%-kal) csökkent a felismerési hiba.

Harmadik finomítási lépésként a nyelvi modellek bevetésével folytattuk. A táblázat 3. sora mutatja a beszédhang-bigrammal kapott eredményeket. Mivel a „szótag"-jellegű felbontást csak a hangoskönyvön csináltuk meg, így az ezekre épülő bigramot is csak ezen az adatbázison értékeltük ki; az eredmény a táblázat 5. sorában található. Mint az várható volt, a kétféle nyelvi modell közül a nagyobb egységekkel dolgozó szótagalapú hozott nagyobb javulást.

1. táblázat: beszédhang-felismerési pontosságok a két adatbázison, különféle akusztikai és nyelvi modellek esetén.

	MTBA	Hangoskönyv
HMM hagyományos jellemzőkkel (nyelvi modell nélkül)	53,37%	72,18%
HMM tandem jellemzőkkel (nyelvi modell nélkül)	65,09%	79,49%
Tandem beszédhang-bigram nyelvi modellel	69,67%	83,62%
Tandem + beszédhang-bigram + diszkriminatív tanítás	73,93%	86,26%
Tandem szótag-bigram nyelvi modellel	---	84,58%
Tandem + szótag-bigram + diszkriminatív tanítás	---	86,33%

Utolsó lépésként bevetettük a HTK diszkriminatív tanítási algoritmusát. Mivel ez a módszer a teljes rendszert finomítja, így mindkét nyelvi modell mellett le kellett futtatnunk a tanítást. A diszkriminatív tanítás újabb 13-15 százalékkal csökkentette a hiba mértékét, ennek köszönhetően a telefonos adatbázison sikerült megközelíteni a 75%-os pontosságot. A hangoskönyvön a kétfajta nyelvi modell között csökkent a különbség, a végeredményként kapott 86,26% és 86,33% közt nincs jelentős eltérés.

A táblázat értékei jól mutatják a két adatbázis által prezentált felismerési feladat nehézségi különbségét: az MTBA adatbázison elért legjobb eredmény alig jobb, mint a hangoskönyvön a legegyszerűbb megoldással kapott érték! Érdekességképp megjegyezzük, hogy a korábban az 1. ábrán bemutatott hisztogram az MTBA-n elért 73,93%-os átlaghoz tartozik, és az ábrán szereplő legmagasabb, 85%-os érték gyakorlatilag megegyezik a hangoskönyvön kapott pontossággal. Tehát az MTBA-n betanított modell is el tudta érni ugyanazt a hatékonyságot, de csak a számára „legszimpatikusabb” beszélőn – a többiek sajnos lehúzták az átlagot.

SZ I N B Á D A Z E L Ő T M E SS Ú T A K A I S H A J D O N D Ó
V O L T H E CC O K NY A F O D R O CS K A CS A L G O TT A -
- M E N P E S T Ö BB U D Á R A -- A N É P R G E TT Ó L A M A
R I CC G E T I V A TT A L E MM É K T O V A B I S -- D E M O S
T -- A L E KK Ö Z E L E BB É S A L O K I K S E M E N T H A --
E S O N M E LL E TT E TY K E D V E S T É S N Ő -- A K I N E K
F E H É R F Á TTY A L A V O L T É SS A L G O S F É R C I P Ő

3. ábra. Példa a beszédfelismerő fonetikai szintű kimenetére.

Az eredmények jól érzékeltetik a tandem technológia hasznosságát is. Meg kell azonban jegyeznünk, hogy az összes kísérletben kizárólag monofón HMM-eket alkalmaztunk. A táblázat 1. sorában összehasonlításként szereplő eredmények feltehetően sokkal magasabbak lennének trifón modelleket használva. A tandem eredmények viszont kevésbé javulnának, ugyanis a neuronháló tanítása elég nehezen házasítható össze a trifón modellezéssel, és ennek optimális megoldása jelenleg a tandem-jellegű módszerekkel foglalkozók egyik legfontosabb kutatási problémája (lásd pl. [1]).

A tandem technológia egyik sajátossága, hogy a neuronháló révén rögtön az adatvektorok szintjén is tudunk mondani részeredményt; a hagyományos HMM-es technológiában ez nem szokás (bár megoldható lenne). Pedig érdekes tanulságokat kínálna annak részletes kielemzése is, hogy vajon a bigram modellel is megtámogatott globális eredmény miért nem jobb lényegesen, mint a neuronháló által a puszta adatvektorokon elért pontosság (73,93% vs. 74,11%, illetve 86,33% vs. 85,24%). Ez a meglepő megfigyelés azt sejteti, hogy a lokális hibák nem egyenletesen oszlanak el, hanem bizonyos hosszabb-rövidebb szakaszokon felhalmozódnak. E hipotézis igazolása azonban mélyre hatóbb kivizsgálást igényelne.

A fonetikai szintű kimenet mellett természetesen nagyon érdekes lenne szószintű eredményeket is látni, a fent kapott értékekből ugyanis nem lehet tudni, hogy vajon a szavakat milyen arányban tudná eltalálni egy szómodelleket is tartalmazó rendszer. A korábban ismertetett okok miatt sajnos nem állt módunkban komolyabb nyelvi modellel is kipróbálni a felismerést; végeztünk azonban egy olvasási tesztet, melynek során a kísérleti személy azt a feladatot kapta, hogy „fejtse meg” a felismerő kimenetét, azaz legjobb tudása szerint javítsa értelmes magyar szöveggé. Erre tetszés szerinti idő állt rendelkezésére, és a szövegben is oda-vissza ugrálhatott. Feladványként a tesztadatbázisba került két Szindbád-történet egyikét kapta meg (melyet korábban még nem olvasott). A 3. ábra egy részletet mutat a dekódolandó betűsorozatból. Kísérleti alanyunknak a szöveg 1337 szövegszavának 94,24%-át sikerült eltalálnia. Meg kell jegyezzük, hogy bár szigorú értelemben 77 szót nem talált el, a hibák túlnyomó többsége csak egyetlen betű vagy morféma eltéréséből állt, és értelemzavarónak csak szűk tucatnyit lehetne nevezni. Ez elég elgondolkoztató arra nézve, hogy az eltalált szavak száma mennyire értelmes mérőszáma a felismerés pontosságának. További észrevételünk, hogy habár az ember által szimulált „nyelvi-szemantikai modell” nyilván összehasonlíthatatlanul ügyesebb, mint a gép, valós szituációban az utóbbi annyival könnyebb helyzetben van, hogy az akusztikai modelltől nem csak a legvalószínűbb megoldást kapja meg, hanem további lehetőségeket is (ún. *N*-best list vagy lattice). Hasonló segítség birtokában feltehetően kísérleti személyünk is még jobb eredményt tudott volna elérni.

6 Összegzés

Cikkünkben egy hangoskönyvön végeztünk beszédfelismerési kísérleteket annak felmérésére, hogy egy ilyen gyakorlatilag optimálisnak nevezhető hangfelvétel esetén milyen felismerési pontosságra képes rendszerünk. Az eredményeket az MTBA telefonbeszéd-adatbázison végzett hasonló tesztek eredményeivel párhuzamba állítva

igazolódott sejtésünk, hogy a hangoskönyv lényegesen egyszerűbb felismerési feladatot jelent. Fonetikai szinten 86%-os pontosságot sikerült elérnünk, ami már szabad szemmel is jórészt értelmezhető kimenetnek felel meg. További legfontosabb feladatnak a tesztek magasabb szintű nyelvi modellel való megtámogatását tartjuk, illetve tervezzük a felismerési hibák jellegzetességeinek elemzését is, ami rálátást adhat az akusztikai modell további javításához.

Hivatkozások

1. Aradilla, G., Bourlard, H., Magimai-Doss, M.: Using KL-based Acoustic Models in a Large Vocabulary Recognition Task. In: Proceedings of Interspeech 2008 (2208) 928–931
2. Bourlard, B., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic (1994)
3. He, X., Deng, L.: Discriminative Learning for Speech Recognition: Theory and Practice. Morgan & Claypool (2008)
4. Hermansky, H., Ellis, D., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of ICASSP 2000 (2000) 1635–1638
5. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing. Prentice Hall (2001)
6. Lippmann, R. P.: Speech Recognition by Machines and Humans. Speech Communication, 22(1) (1997) 1–15
7. Mihajlik P., Tatai, P.: Automatikus fonetikus átírás magyar nyelvű beszédhez. Beszédkutatás 2001 (2001) 172–185
8. Mihajlik P., Tarján B., Tüske Z., Fegyó T.: Investigation of Morph-based Speech Recognition Improvements across Speech Genres. In: Proceedings of Interspeech 2009 (2009) 2687–2690
9. Tóth L., Frankel, J., Gosztolya G., King, S.: Cross-lingual Portability of MLP-Based Tandem Features - A Case Study for English and Hungarian. In: Proceedings of Interspeech 2008 (2008) 2695–2698
10. Vicsi K., Tóth L., Kocsor A., Gordos G., Csirik J.: MTBA - magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika, Vol. LVII, No.8 (2002) 35–43
11. Vicsi K., Kocsor A., Teleki Cs., Tóth L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: II. Magyar Számítógépes Nyelvészeti Konferencia (2004) 315–318
12. Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A.: Effect of speaking style on LVCSR performance. In: Proceedings of ICSLP 1996 (1996) 16–19
13. Young, S. et al.: The HMM Toolkit (HTK) – software and manual. http://htk.eng.cam.ac.uk (1995)
14. Zhao, Y.: Speech-Recognition Technology in Health Care and Special-Needs Assistance. IEEE Signal Processing Magazine, 26(3) (2009) 87–90

Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban

Vicsi Klára, Sztahó Dávid
Budapesti Műszaki és Gazdaságtudományi Egyetem

Távközlési és Médiainformatikai Tanszék Beszédakusztiai Laboratórium,
1111 Budapest, Sztoczek utca 2.
vicsi@tmit.bme.hu, sztaho@tmit.bme.hu

Kivonat: A cikkünkben egy érzelem-felismerési kísérletről számolunk be, ahol a spontán társalgás során a semlegesről idegesre, feszültre megváltozott érzelmi állapotot kívánjuk automatikusan detektálni, telefonon keresztül. A cél egy automatikus figyelőrendszer kifejlesztése, amely meghatározza az ügyfél elégedettségének, vagy elégedetlenségének a mértékét. Ehhez a munkához létrehoztuk, 1000 telefonhívás-felvételből az ún Magyar Telefonos Ügyfélszolgálati Beszéd Adatbázist (MTÜBA), amelyben a spontán dialógusok nyelvi tartalmát, valamint frázisonkénti érzelmi tartamát jelöltük be. Az akusztikai előfeldolgozás után az érzelem-felismerést support vector machine (SVM) osztályozó segítségével végeztük. Az SVM osztályozóval végül is csak 2 állapotot, egy semleges, és egy elégedetlenséget kifejező (ideges és panaszkodó együtt) állapotot különböztettünk meg. Az automatikus figyelőrendszer részére kiválasztottunk 15 másodperc hosszú figyelő ablakot, amelyen belül összeszámoltuk az elégedetlenséget jelző frázisok számát. Ez adta meg az elégedetlenség mértékét. Az ablakot 10 másodpercenként léptettük előre a beszélgetés folyamán. Kísérletezéssel beállítható volt egy olyan elégedetlenségi mérték küszöb, amely felett jelzés (riasztás) történik. Amennyiben ez a küszöb a 30%-os elégedetlenségi mérték, akkor az átlagos riasztási pontosság 89,6% volt, ami legtöbbször csak a kézi és az automatikus riasztás közötti időcsúszásból eredt. Így a kifejlesztett automatikus figyelőrendszer hasznos eszköz lehet diszpécser központokban.

1 Bevezetés

Az emberi beszédkommunikációban a beszéd információfeldolgozásának két egymástól elkülönült feldolgozási módjáról beszélhetünk. Az egyik feldolgozási mód esetében speciális szemantikai tartalmú üzeneteket dolgozunk fel (verbális csatorna); a másik információfeldolgozási mód az, ahol a beszélő általános érzelmi, egészségi állapotát, hangulatát dolgozzuk fel (a nem verbális csatorna) [1]. Az utóbbi évtizedekben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna jelentősége ez ideig kisebb volt, és működését kevésbé értjük.

Az emberi beszéddel nagyon sok mindent ki lehet fejezni a nyelvi tartalmon kívül, amelyeket különböző beszédváltozatok jelenítenek meg, például a beszédstílus, rit-

mus, hangerő, hangszín, intonáció – ezek mind széles körben használatosak arra, hogy a beszélő érzelmi, egészségi állapotát egyidejűleg kifejezzék. Csak az utóbbi években növekedett meg a jelentősége a beszéd különböző paralingvisztikai és extralingvisztikai nézőpont szerinti vizsgálatának. Az irodalomban található néhány kutatási leírás, amely a beszéd érzelemtartalmának vizsgálatával, és az érzelem automatikus felismerésével foglalkozik, de ezek az eredmények mind laboratóriumi körülmények között elhangzó tiszta beszédre vonatkoznak [2, 3, 4, 5]. A publikációk legtöbbjében szimulált különböző érzelemtartalmú beszédet használnak, leggyakrabban művészek bemondásmintáit. Az érzelem jellemzésére a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban, például az MPEG-4 szabvány leírásában [6] hagyományos érzelemkategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban e kategóriákat az arcmimika jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták.

A valóságban rendszerint spontán beszédet használunk, és a spontán beszédre jellemző adatok igen nagymértékben különböznek a színészek által produkált szép beszédtől [7], és a beszédtechnológiai alkalmazásokban a valóságos spontán beszéd alkalmazása az, ami szükséges. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával [8] és információtartalmainak felismerésével [9] foglalkozik.

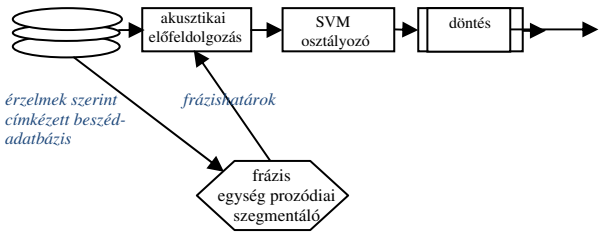
Jelen cikkünkben a telefondiszpécser és az ügyfél közötti hétköznapi spontán társalgási adatbázis alapján végzett automatikus érzelem-felismerési kísérletekkel foglalkozunk. Az akusztikai előfeldolgozásnál támaszkodtunk a korábban végzett, imitált érzelemtöltetű beszéd felismerési kísérleteink eredményeire [10].

A cikkünkben a beszéd érzelmet kifejező akusztikai paramétereinek a felismerését tárgyaljuk, de tervezzük a verbális csatornán keresztül is a nyelvi tartalom érzelemre vonatkozó statisztikai jellegzetességeinek vizsgálatát is.

2 Rendszerleírás

Egy beszélgetés során, különösen, ha az hosszan tartó, a beszélő érzelmi állapota, hangulata változik. Ha követni akarjuk a beszélő érzelmi változásait, szegmensekre kell felosztanunk a beszédfolyamot, így meg tudjuk vizsgálni, hogyan változik szegmensről szegmensre a beszélő érzelmi állapota a beszélgetés alatt. Rendszerünkben a frázist választottuk szegmentálási egységként, a korábbi tanulmányaink során nyert tapasztalatok alapján [10]. A frázis méretű egységek szegmentálásakor az egységekre való osztást prozódiai szegmentálónk végezte el [11]. (Ezt a szegmentálót a folyamatos beszéd felismerés részeként a beszéd szemantikai feldolgozására fejlesztettük ki, amelyet a frázis- és mondathatárok detektálására és a modalitás (mondattípus) felismerésére használtunk.)

Az akusztikai előfeldolgozás után a frázis méretű szegmenseket azok érzelmi töltete szerint osztályoztuk, SVM (support vector machine) gépi osztályozót használva. A rendszerünk folyamatábráját az 1. ábra szemlélteti.



1. ábra. Beszédérzelem osztályozónk blokkvázlata.

Kezdetben négy különböző érzelmi állapot került megkülönböztetésre a rögzített dialógusokban: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Később, a kísérletek tapasztalata alapján ezeket az érzelmeket összevontuk, már csak összesen két érzelmi osztályt különböztetve meg.

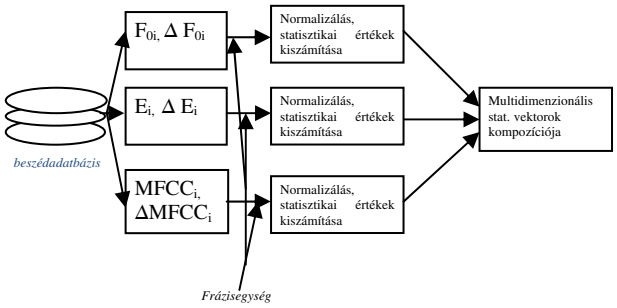
Végezetül ahhoz, hogy az érzelmi döntéshozás biztosabb legyen, egyszerre több frázis együttes kezeléséből alkotunk végleges döntést a beszélő érzelmi állapotáról.

2.1 Akusztikai előfeldolgozás

Általánosságban az alapfrekvencia, az intenzitás és annak időbeli függése a leghagyományosabban használt fizikai jellemző az érzelmek kifejezésére, mind a beszédfelismerés, mind a beszédszintézis területén. Azonban a korábbi automatikus beszéd felismerési kísérleteink során kiderült, hogy spektrális információ hozzáadása nagymértékben javítja az érzelem-felismerési eredményeket [10]. Ennek megfelelően az alapfrekvenciákat (F_{0i}), az intenzitásértékeket (E_i), 12 MFCC-t és deriváltjaikat mértük, 150 ms időablakot használva 10 ms időkeretekben, összesen 28 tulajdonságvektorral 10 ms-ként. Ezután a frázis prozódiai szegmentáló kijelöli a frázishatárokat a beszédben, frázisok sorozatát hozva ezzel létre. A 10 mszekundumonkénti tulajdonságvektorok alapján minden egyes frázist egy multi-dimenzionális statisztikai tulajdonságvektor jellemez, amint azt a 2. ábra mutatja. Ezeket a statisztikai tulajdonságvektorokat a következők szerint számítottuk ki: először F_{0i} értekeit az első időkeret F_{0i} értékeire, az E értékeket pedig az E maximum érték szerint normalizáltuk minden egyes frázis esetében. Majd e normalizált paraméterekből számítottuk ki a következő statisztikai adatokat minden egyes frázisnál:

- F_{0i} maximum, minimum, közép, medián értékei
- ΔF_{0i} maximum, minimum, közép, torzulás (skew) értékei
- E_i közép, medián értékei
- ΔE_i maximum, minimum, közép, torzulás (skew) értéke

- $MFCC_i$ maximum, minimum, közép értékei
- $\Delta\ MFCC_i$ maximum, minimum, közép értékei



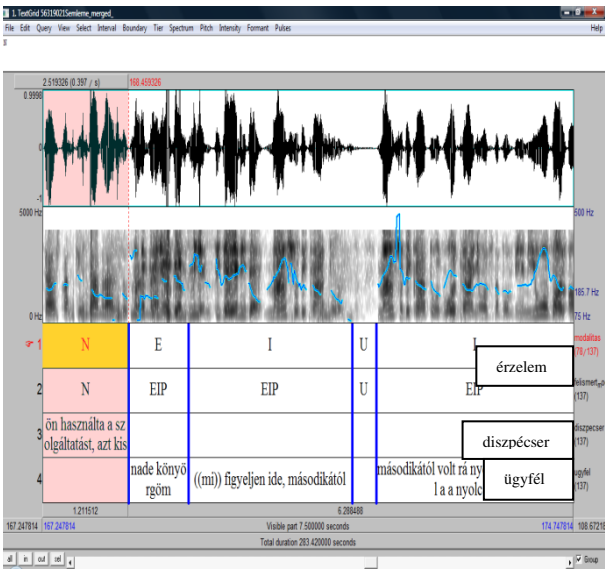
2. ábra. Akusztikai előfeldolgozás

2.2 Telefonos Ügyfélszolgálati Beszéd Adatbázis (TÜBA)

A TÜBA egy telefonos ügyfélszolgálat dialógusainak gyűjteménye, amely telefonvonalon keresztül lett rögzítve, 250-3500 Hz közötti frekvenciasávban, 8000 Hz-es mintavételi sebességgel és 16 bites amplitúdó felbontásban. A diszpécserek és ügyfelek közötti párbeszédek időtartalma változó, 1 és 30 perc közötti volt. A hanganyag feldolgozásához, a szegmentáláshoz és a címkézéshez a közismert Praat fonetikai feldolgozó programot [13] használtuk, mivel ez az eszköz megfelelő a párhuzamos feldolgozáshoz. A frázishatárok bejelölése után frázisonként bejegyzésre került a nyelvi tartalom, és a hozzá tartozó érzelem is párhuzamosan. A beszélő, a beszélő neme szintén bejegyzésre került.

A frázishatárok automatikus kijelölésére a prozódiai szegmentálónkat [11] használtuk , amint azt már az előző fejezetben is említettük. Azután szakértők kézzel javították a határokat, érzelem szerint felcímkézték a frázisszegmenseket. Négy különböző érzelmi állapotot különböztettek meg a rögzített párbeszédekben: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Gyakorlatilag nem volt több érzelemtípus az 1000 hívásban, csupán ez a négy. Sajnos sok esetben az ügyfél beszéde semleges volt. Összesen 346 ideges, 603 panaszkodó, és 225 egyéb frázis volt az ügyfelek beszédében, valamint több ezer semleges, amelyből 603 tipikusan semleges frázist választottunk ki a négy érzelem betanítására az osztályozási kísérletben.

A párbeszédek szegmentálásának és címkézésének egy példáját a 3. ábra mutatja be. A kézi szegmentálás és címkézés a harmadik sorban jelenik meg, osztályozónk címkézési eredménye pedig alatta látható. Az ügyfél és a diszpécser beszédének szövege a beszéddel és az érzelemmel párhuzamosan került lejegyzésre.



3. ábra. Példa a TÜBA szegmentálására és címkézésére. U: szünet, N: semleges, I: ideges, P: panaszkodó és E: egyéb.

2.3 A rendszer tesztelése

Frázisok érzelem szerinti osztályozása

Érzelmi osztályozónk betanítására és tesztelésére az úgynevezett „leave-one-out cross-validation” (LOOCV) módszert használtuk [12], amely egyetlen frázist használ értékelési adatként, a hívás fennmaradó frázisait pedig betanítási adatként. Majd ez úgy ismétlődik, hogy végül is minden egyes frázis egyszer értékelési adatként kerül felhasználásra. Az 1. táblázat mutatja a négy érzelem esetében kapott hibamátrixot.

1. táblázat: E, I, P, N érzelmek felismerési hibamátrixa.

	E	I	N	P	Pontosság
E	**49**	26	62	88	22%
I	9	**153**	60	124	44%
N	14	38	**398**	153	66%
P	11	70	157	**365**	60%
				átlag	54%

Az I és P érzelmeket nemcsak az osztályozó, de az emberek is alig tudták differenciálni. Így az I, P és E osztályok egy osztályba kerültek, mint elégedetlenséget kifejező érzelmek. Tehát végül az „elégedetlen” osztályt és a semleges érzelmek osztályát

különböztettük meg, és így tanítottuk be az SVM osztályozót. A teszteredményeket a 2. táblázat szemlélteti.

2. táblázat: Az (E, I, P), mint elégedetlen érzelmű összevont osztály, és az (N) semleges érzelmű osztály felismerési hibamátrixa.

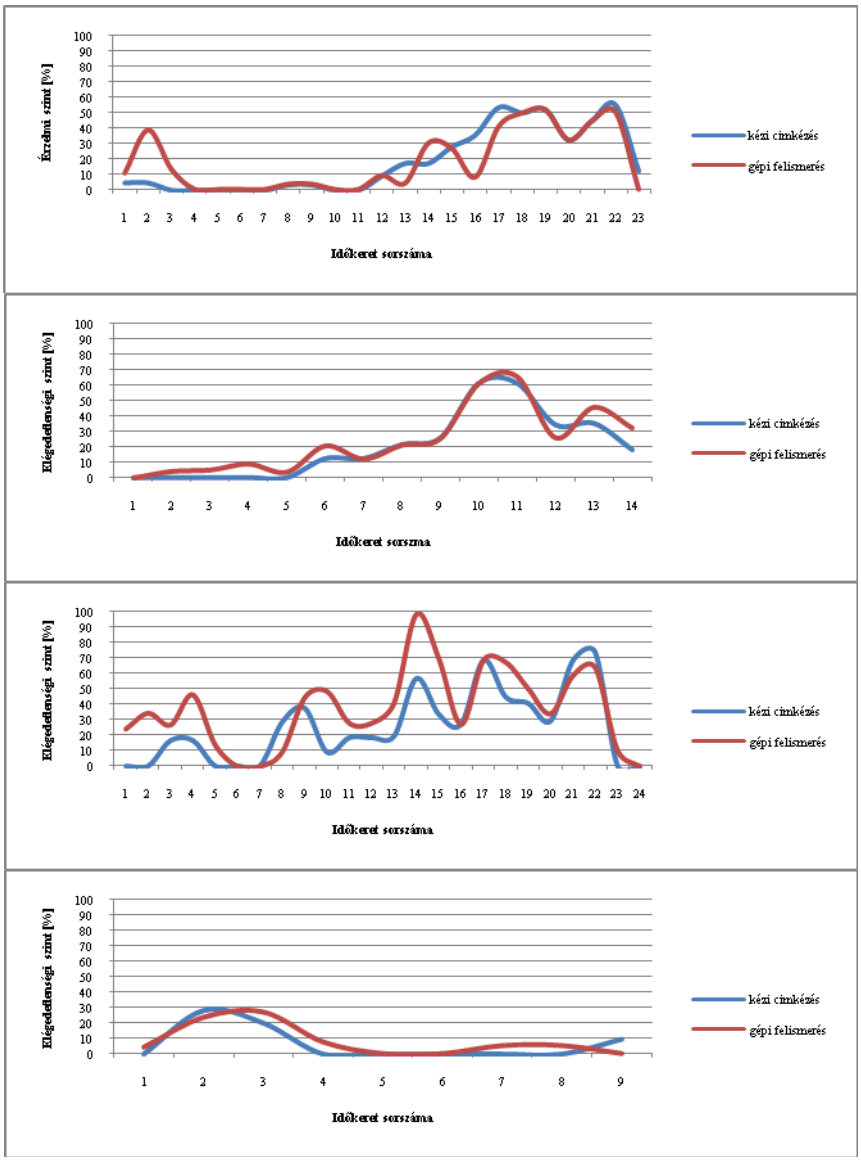
	EIP	N	Pontosság
EIP	**887**	287	76%
N	335	**839**	71%
		átlag	**73%**

Az ügyfél érzelmi állapotának detektálása

E kutatómunkának a célja annak a feltárása, hogy egy beszélgetés során hogyan lehetséges az ügyfelek érzelmi állapotát automatikusan felismerni. Frázisonként változhat, ugrálhat a megítélt érzelem. Biztos döntés akkor hozható, ha több frázison keresztül többségében egy típusú érzelem fordul elő. Ehhez előzetes kísérletezgetés alapján 15 másodperc hosszúságú időablakot választottunk, és mértük az ablakon belül az „elégedetlen"-nek osztályozott frázisok számát. Ez a szám %-ban kifejezve adta meg az „elégedetlenség" mértékét. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve.) Azután az ablakot továbbmozgattuk, 10 másodperc időlépéssel. A 4. ábrán néhány példa jelenik meg arról, hogyan változik meg az ablakban mért szám, vagyis az elégedetlenség mértéke a beszélgetés során. Az automatikusan nyert eredményeket összehasonlítottuk a kézzel felcímkézett eredményekkel.

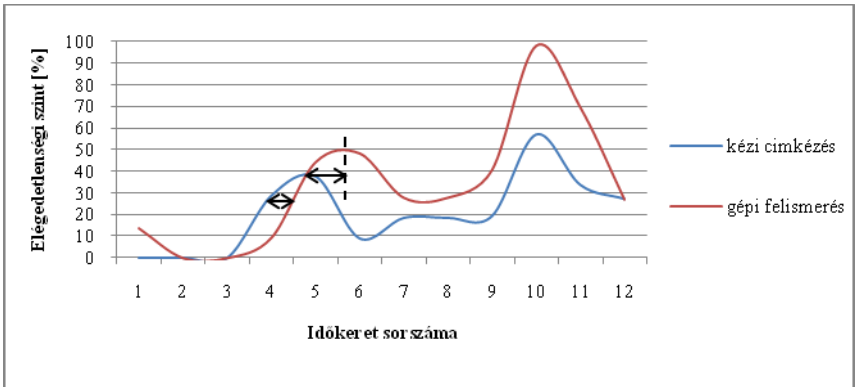
Egészében véve folyamatos megfigyelés esetében az automatikusan nyert, és a kézzel címkézett eredmények között az átlagos távolság 11,3% volt, összehasonlítva minden 10 másodperces időlépésben a megfigyelt eredményeket, és átlagolva a kapott különbségeket az egész adatbázishoz.

A valós felhasználásban az automatikus felismerés fő célja jelezni, amikor az elégedetlenségi szint elérte a kritikus szintet. Mi ezt „riadószint"-nek nevezzük. Ez a „riadószint" manuálisan beállítható. Például, válasszuk 30 százalékra a „riadószint"-et (ez azt jelenti, hogy 30% felett van a mért elégedetlenség). Vizsgáljuk meg ebben az esetben a riasztási pontosságot. Ezt úgy végezhetjük el, hogy összehasonlítjuk, az automatikus riasztást azzal a riasztással, amit az előzetesen kézzel címkézett anyagon számolunk. Az összehasonlítás keretről keretre történt. A különbségeket riadódetektálási hibának tekintettük. Az átlagos riadódetektálási hiba 10,4%-os volt.



4. ábra. Az ügyfél elégedetlenségének mértéke egy beszélgetés során. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve.) Az automatikusan nyert eredményeket összehasonlítottuk a kézzel felcímkézett eredményekkel.

Ha csak azokat a párbeszédeket nézzük, ahol egyáltalán nem volt „riadószint” (semleges párbeszédek), a „riadószint” detektálási hiba 6,8%-os volt. Ez azt jelenti, hogy ha csak a több mint 30 százalékos elégedetlenséget tekintjük „elégedetlen” érzelmi állapotnak, az automatikus felismerési arány 93,2%-os. Egyéb párbeszédek érzelmi töltete (ahol a kézi felcímkézés legalább egy esetben elérte a „riadószint”-et) 77,2%-ban került felismerésre. A hibák fő oka az automatikus adatfelismerés és a kézi felcímkézés közötti kismértékű eltolódás. Ezt illusztrálja az 5. ábra.



5. ábra. A 4. ábra 3. diagramjának kinagyítása, példa a kézi felcímkézés és az automatikus felismerés görbéi közötti kismértékű eltolódásra.

3 Összegzés

A kísérletsorozat kezdetén, a 2.3.1. bekezdésben négy különböző érzelmi állapotot különböztettünk meg a rögzített párbeszédekben: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Ezeknek az érzelmeknek az átlagos osztályozási pontossága csupán 54%-os volt. Az osztályozási pontosság természetesen bizonyos mértékig növelhető a betanítási anyag növelésével, de az érzékelési kísérletek során, még művészek által előadott beszédnél is az emberi érzelem-felismerés (nem verbális csatornákon) általánosságban kevesebb volt, mint 70% (hat alapérzelem esetében) [2, 3, 5, 10] specifikus szemantikai tartalom nélkül (verbális csatornák). Ebből következik, hogy aligha várható sokkal jobb eredmény az automatikus érzelem-felismerés esetében, spontán beszédnél. Világos, hogy sokkal jobb eredmény érhető el, ha a verbális csatorna néhány információja a rendszerhez integrálódik. Ez az oka annak, hogy a lingvisztikai tartalmat is rögzítettük az adatbázis feldolgozáson keresztül, amint azt a 2.2 bekezdésben leírtuk. A jövőben azt tervezzük, hogy néhány lingvisztikai információt is feldolgozunk, és a két csatorna információit fogjuk integrálni.

A 2.3.2. bekezdésben leírt második kísérletünk során az osztályozott frázisokat egy időablakon keresztül figyeltük meg, hosszabb ideig, mint ameddig a frázis tart, hogy specifikusabb döntést hozhassunk a beszélő érzelmi állapotát illetően. Ez a megfigyelési technika képesnek látszik arra, hogy riasztást adjon, ha az ügyfél elégedetlensége

túlmegy egy bizonyos küszöbön, még verbális csatorna használata nélkül is. Ennek megfelelően a leírt döntési technika hasznos lehet a diszpécserközpontokban.

Köszönetnyilvánítás

Ezúton kívánunk köszönetet mondani az SPSS Hungary Ltd.-nek és az INVITEL Telecom Zrt.-nek a rendelkezésünkre bocsátott 1000 dialógusért.

Hivatkozások

1. Burkhardt, F., Paeschke A. et al.: A database of German Emotional Speech. N: Proc. Of Interspeech2005 (2005) 1517-1520
2. Campbell, N.: Getting to the heart of the matter. Keynote Speech in Proc. Language resources and Evaluation Conference (LREC-04), Lisabon, Portugal (2004)
3. Campbell, N.: Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. COST Action 2102 International Conference on Verbal and Nonverbal Features….Patras, Greece, (2007) 107-120
4. Douglas-Cowie, E. – Campbell, N. – Cowie, R. – Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication 40. (2003) 33–60
5. 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. May, Vol. 119, Issue 5. (2006) 3109-31206
6. Kohavi, R.: "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12) (1995) 1137–1143
7. Kostoulas, T., Ganchev, T., Fakotakis, N.: Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data, COST Action 2102 International Conference on Verbal and Nonverbal Features….Patras, Greece, October 2007. (2007) 235-242.8
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 Transaction on Audio, Speech, and Language Processing, vol. 14, no. 4, July, 2006 (2006)
9. MPEG-4: ISO/IEC 14496 standard. http://www.iec.ch, (1999)
10. Tóth Sz. L., Sztahó D., Vicsi K.: Speech Emotion Perception by Human and Machine. Proceeding 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
11. Praat, http://www.fon.hum.uva.nl/praat/
12. Vicsi, K. Szaszák, Gy.: Using Prosody for the Improvement of ASR: Sentence Modality Recognition. In: Interspeech 2008. Brisbane, Ausztrália 2008.09.23-2008.09.26. ISCA Archive, http:www.isca-speech.org/archive, (2008)
13. Wilting, J., Kramber, E., Swerts, M.: Realvs. Acted emotional speech.In:Proc. Of the Interspeech 2006 (2006) 805-808

Mássalhangzó-magánhangzó kapcsolatok automatikus osztályozása szubglottális rezonanciák alapján

Csapó Tamás Gábor[1], Németh Géza[1]

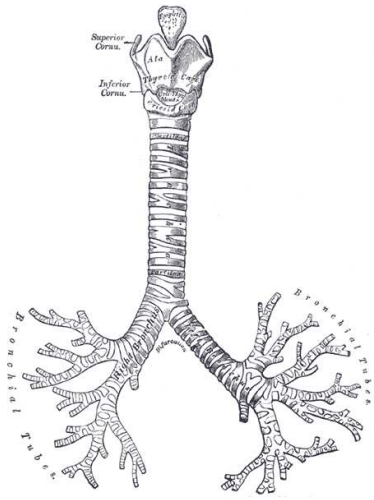
[1]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
Budapest, 1117, Magyar tudósok krt. 2.,
e-mail: {csapot,nemeth}@tmit.bme.hu

Kivonat A nemzetközi szakirodalom az elmúlt években kezdett intenzíven foglalkozni a szubglottális rezonanciák vizsgálatával, melyek az alsó légutak rezonanciái. Korábbi kutatásokban kimutatták, hogy ezek a magánhangzókat természetes osztályokra tagolják. A mássalhangzó-magánhangzó kapcsolatokban a magánhangzó formánsértékei nem állandóak a koartikuláció miatt. A zárhangok például képzési helyüktől függően módosítják a szomszédos magánhangzó formánsait. A mássalhangzó végén és a magánhangzó közepén mérhető második formáns értékét összevetve rajzolható meg a locus egyenlet tér, melyben az egyes beszédhangosztályok az artikulációs helyük szerint elkülönülve jelennek meg. Hipotéziseink szerint a csoportok elkülönüléséhez a szubglottális rezonanciák is hozzájárulnak, hasonlóan a magánhangzókban okozott kategorikus elválasztáshoz. Jelen kutatás során egy magyar anyanyelvű beszélő alapján tovább vizsgáljuk a mássalhangzó-magánhangzó kapcsolatok helyét a locus egyenlet térben, valamint a szubglottális rezonanciák csoportelválasztó szerepét is elemezzük. Bemutatjuk egy automatikus osztályozó működését, amely a szubglottális rezonanciák és a második formáns viszonya alapján csoportosítja a mássalhangzó-magánhangzó beszédhang-kapcsolatokat.

Kulcsszavak: szubglottális rezonancia, SGR, CV-kapcsolat, locus egyenlet

1. Bevezetés

A nemzetközi szakirodalom az elmúlt években kezdett intenzíven foglalkozni a szubglottális rezonanciák (SGR) vizsgálatával, melyek az alsó légutak (pl. tüdő, légcső, hörgők, l. 1. ábra) rezonanciái [15]. Ezek a formánsokhoz hasonlóan alakítják a zöngés hangok spektrumát, de a formánsokkal ellentétben nem erősítik a rezonanciafrekvencia körüli harmonikusokat, hanem gyengítik őket. Mivel az alsó légúti szervek viszonylag keveset mozognak a beszéd során, a rezonanciafrekvenciák közel állandóak egy-egy ember beszédében.



1. ábra. Az alsó légúti rendszer [4]. Tipikus rezonanciafrekvencia értékei 600 Hz, 1400 Hz és 2100 Hz körüliek [15].

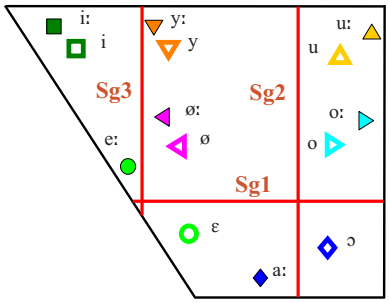
1.1. A szubglottális rendszer rezonanciáinak szerepe

Korábbi kutatásokban kimutatták, hogy a szubglottális rezonanciák a magánhangzókat kategorikusan természetes osztályokra tagolják [8]. Az angol nyelven végzett vizsgálatok alapján az derült ki, hogy a második szubglottális rezonancia ($Sg2$) természetes határként (fonológiai megkülönböztető jegy, [14]) szolgál az elöl és hátul képzett magánhangzók között: ha a második formáns frekvenciája ($F2$) magasabb, mint a második alsó légúti rezonancia, akkor elöl képzett magánhangzóként érzékeljük, ha alacsonyabb, akkor hátul képzettként. Az első ($Sg1$) és harmadik ($Sg3$) alsó légúti frekvencia elválasztó szerepére is utalnak bizonyos eredmények [8].

Az eddigi eredmények szerint a szubglottális rezonanciák a formánsmenetekben a folytonosság megszakadását okozhatják [2], észrevehetőek a beszédpercepció számára [7], valamint hasznosak lehetnek a beszélőnormalizálásban [16,17]. Eddig azonban csak néhány nyelvre vizsgálták a magánhangzó-formánsok és SGR-ek kapcsolatát. Wang és kollégái angol-spanyol kétnyelvű gyermekek beszédével foglalkoztak [16]. Lulich egy felnőtt férfi és kilenc gyermek amerikai angol beszélő $Sg2$ és $F2$ kapcsolatát elemezte [8]. Madsack és társai az $Sg1$-$F1$ és $Sg2$-$F2$ közötti összefüggést kutatta két német dialektus néhány beszélőjén [11], Jung pedig hasonlót végzett a koreai nyelvre [6].

A szubglottális rezonanciák magánhangzó-elkülönítő szerepével kapcsolatban magyar nyelvre eddig kezdeti kutatások történtek csak. Az első kísérletek alapján az $Sg1$, $Sg2$ és $Sg3$ szerepet játszhat a beszédhangok produkciójában [3]. Az eredmények szerint az $Sg1$ az alsó és nem alsó, az $Sg2$ az elöl és hátul képzett magánhangzók közötti határon található, míg az $Sg3$ az elöl képzett ajakréses nem alsókat különíti el a többi elöl képzett magánhangzótól. A magyar magánhangzócsoportok között feltételezett elválasztó szerepet a 2. ábra mutatja.

A vízszintes és függőleges vonalak utalnak a szubglottális rezonanciák helyére a formánstérben. [3] eredményei szerint két férfi és két nő logatom-olvasása alapján nagyrészt teljesülnek ezek a hipotézisek.



2. ábra. A magyar magánhangzók elméleti formánstere. A vízszintes és függőleges vonalak a szubglottális rezonanciák által feltételezett elkülönülést mutatják.

A különböző nyelvekre történt kutatásokat viszonylag kevés adaton végezték el, de az eredmények konzisztensek abban, hogy mindegyik vizsgált nyelvben az alsó légúti rezonanciák határként szolgálnak különböző magánhangzó-csoportok között.

1.2. Formánsmenetek mássalhangzó-magánhangzó kapcsolatokban

A mássalhangzó-magánhangzó (CV) kapcsolatokban a magánhangzó formánsértékei nem állandóak a két hang közötti koartikuláció miatt [5]. A zöngés és zöngétlen zárhangok képzési helyüktől függően kisebb-nagyobb mértékben módosítják a szomszédos magánhangzó formánsait. A második formáns változása alapján ezen hangkapcsolatokat regressziós egyenesek (ún. locus egyenlet) segítségével jellemezhetjük [9]. A regressziós egyenesekből megrajzolható az ún. locus egyenlet tér, mely a zárhang végén és a mássalhangzó közepén mérhető második formáns értékét veti össze [9, 2. ábra]. Ezen ábrán az egyes beszédhangosztályok az artikulációs helyük szerint elkülönülő csoportokban jelennek meg az $F2$ változása miatt. Néhány korábbi kísérletben kimutatták, hogy ezen csoportok elkülönüléséhez a szubglottális rezonanciák is hozzájárulnak, hasonlóan a magánhangzókban okozott kategorikus elválasztáshoz [9,10].

Jelen kutatás során tovább vizsgáljuk a mássalhangzó-magánhangzó kapcsolatok helyét a második formáns által meghatározott locus egyenlet térben, valamint az alsó légúti rezonanciák csoportelválasztó szerepét is bemutatjuk. A kísérleteink során egy magyar anyanyelvű beszélő hangfelvételeit és szubglottális felvételeit elemezzük. Bemutatjuk egy automatikus osztályozó eljárás működését, amely az alsó légúti rezonanciák és a második formáns viszonya alapján csoportosítja a mássalhangzó-magánhangzó beszédhangkapcsolatokat. Az eredmények segíthetik a fonológiai megkülönböztető jegyek szerepének megértését,

illetve alkalmazásra kerülhetnek a beszélőnormalizálásban és beszédfelismerésben.

2. Módszerek

A kísérleteink során egy magyar anyanyelvű beszélő ("B1", 29 éves, férfi) beszédfelvételeit és szubglottális felvételeit elemeztük. A rögzített hanganyagot elsősorban akusztikai szempontból vizsgáltuk.

2.1. Beszédfelvételek

A felvétel során "B1" beszélő "ɔCVbɔ" típusú logatomokat olvasott fel egy csendesszobában. A logatomok első mássalhangzója az összes zöngés és zöngétlen zárhangot tartalmazta (labiálisok: [b,p], alveolárisok: [d,t], velárisok: [g,k] és palatálisok: [ɟ,c]). A középső hangsúlytalan szótagban mind a 14 magyar magánhangzó szerepelt ([ɔ,aː,o,oː,u,uː,ɛ,eː,i,iː,ø,øː,y,yː]). A logatomokat a beszélő véletlenszerű sorrendben olvasta fel, mindegyiket tízszer, így összesen 1120 logatomot kiejtve. A beszédhangot EMC 100 kondenzátor mikrofonnal rögzítettük, mely a felvétel során a beszélő ajkaitól kb. 15 cm-re helyezkedett el. Az elhangzott anyagot 48 kHz-es mintavételezéssel digitalizáltuk Terratex DMX 6 Fire USB külső hangkártyával, a Wavesurfer programot használva [13].

2.2. Formánsmérések

Az "ɔCVbɔ" logatomok hanghatárait a beszédfelvétel és a felolvasott szöveg alapján automatikus módszerrel határoztuk meg, egy beszédfelismerőt kényszerített üzemmódban használva [12]. A második formánsfrekvenciákat Praat segítségével [1] mértük automatikusan, a zárhang végén ($F2_{msh}$, a hangidőtartam 95%-ánál), valamint a második magánhangzó közepén ($F2_{mgh}$, a hangidőtartam 50%-ánál). Az automatikus formánsmérés eredményén ezután kézi ellenőrzést végeztünk: manuálisan megvizsgáltuk az átlagostól jelentősen eltérő eseteket, külön-külön az egyes CV csoportokra. A mássalhangzókban és magánhangzókban mért formánsértékek mediánjait az 1. táblázat tartalmazza.

2.3. Szubglottális felvételek

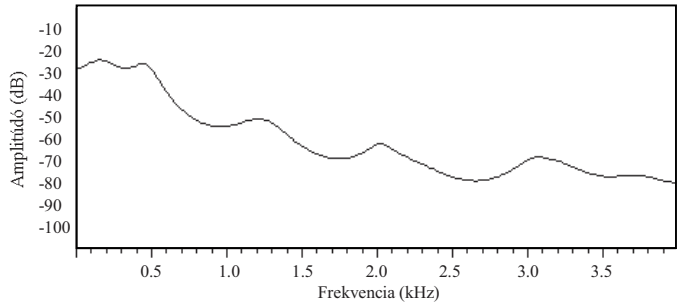
Egy másik felvétel során az alsó légúti rendszer jelét is felvettük csendesszobában, "B1" beszélőtől. Amíg a beszélő felolvasott néhány mondatot, a beszédhangját és alsó légúti jelét rögzítettük. A beszédfelvételeket jelen kísérlet során nem használtuk fel. A szubglottális jelet egy K&K HotSpot gyorsulásmérő eszköz segítségével vettük fel, amely a beszéd során a beszélő nyakára volt szorítva, a pajzsporc fölé. A jelet 8 kHz-es mintavételezéssel, Terratex DMX 6 Fire USB külső hangkártyával digitalizáltuk a Wavesurfer programmal.

1. táblázat. "B1" beszélő beszédfelvételein mért $F2_{msh}$ és $F2_{mgh}$ értékek mediánjai (az értékek Hz-ben értendők). Az $F2_{msh}$ értékeket a zárhangok 95%-ánál, az $F2_{mgh}$ értékeket a magánhangzók 50%-ánál mértük.

		$F2_{msh}$								$F2_{mgh}$							
		Labiális		Alveoláris		Veláris		Palatális		Labiális		Alveoláris		Veláris		Palatális	
		b	p	d	t	g	k	ɟ	c	b	p	d	t	g	k	ɟ	c
Hátsó	ɔ	1045	1435	1074	2022	1066	1001	1560	2058	1056	1251	1114	1295	1095	1037	1197	1322
	o	830	1304	843	1651	878	841	1514	1714	797	978	875	1003	845	786	958	1036
	oː	817	1374	853	1632	782	793	1499	1860	651	691	675	720	661	633	674	703
	u	805	1486	852	1703	807	789	1587	1899	691	878	749	919	712	686	849	976
	uː	825	1435	805	1690	784	798	1526	2035	619	712	640	691	644	552	678	728
	aː	1236	1655	1714	2001	1752	1266	1638	2106	1478	1506	1593	1560	1564	1504	1527	1541
Első	ɛ	1518	1726	2021	2181	2101	1542	1753	2179	1678	1716	1798	1846	1795	1706	1678	1812
	ø	1348	1661	1374	2076	1524	1390	1695	2018	1433	1500	1475	1583	1477	1446	1500	1613
	øː	1518	1726	1635	2055	1688	1525	1729	2007	1659	1680	1600	1702	1621	1602	1703	1663
	y	1594	1841	1730	2149	1809	1569	1860	2116	1803	1904	1740	1909	1782	1824	1975	1881
	yː	1708	1899	1796	2198	1961	1774	1934	2149	1953	2002	1824	1927	1849	1878	1911	1848
	eː	1769	1894	2112	2242	2299	1997	1880	2264	2278	2302	2288	2306	2300	2308	2287	2296
	i	1939	2022	2242	2225	2292	1956	1947	2244	2209	2281	2300	2255	2258	2240	2274	2190
	iː	2014	2025	2235	2217	2266	2308	1945	2309	2317	2380	2409	2334	2357	2312	2358	2358

2.4. Szubglottálisrezonancia-mérés

A szubglottális jelből manuális módon, a Wavesurfer program segítségével mértük az első három szubglottális rezonancia értékét. A 3. ábra egy példa spektrumot mutat "B1" gyorsulásmérő felvételéből, melyen látható, hogy az SGR-mérés a formánsméréshez hasonlóan, a spektrumbeli csúcsok leolvasásával történik. Az SGR meghatározásának módszeréről részletesebb leírás olvasható [2,8]-ben. A hullámformában 20 helyen mértük meg az SGR-értékeket, az összesített adatok a 2. táblázatban találhatóak.



3. ábra. Példa LPC spektrum "B1" beszélő gyorsulásmérő felvételéből. A spektrális csúcsok (454 Hz, 1211 Hz, 2023 Hz és 3067 Hz) a szubglottális rezonanciák értékei. Az ábrán látható $Sg1$ értéke meglehetősen alacsony a szakirodalmi adatokhoz képest [15].

2. táblázat. "B1" beszélő gyorsulásmérő felvételében mért SGR értékeinek adatai. A továbbiakban a medián értékeket használtuk fel.

	$Sg1$	$Sg2$	$Sg3$
Átlag	545 Hz	1241 Hz	2027 Hz
Medián	554 Hz	1244 Hz	2022 Hz
Szórás	60 Hz	42 Hz	145 Hz

3. Eredmények

A formánsmérések alapján megvizsgáltuk "B1" beszélő locus egyenlet terét, a szubglottális rezonanciák szerepét kiemelve. Ezután egy osztályozó segítségével vizsgáltuk a különböző CV-csoportok elválaszthatóságát.

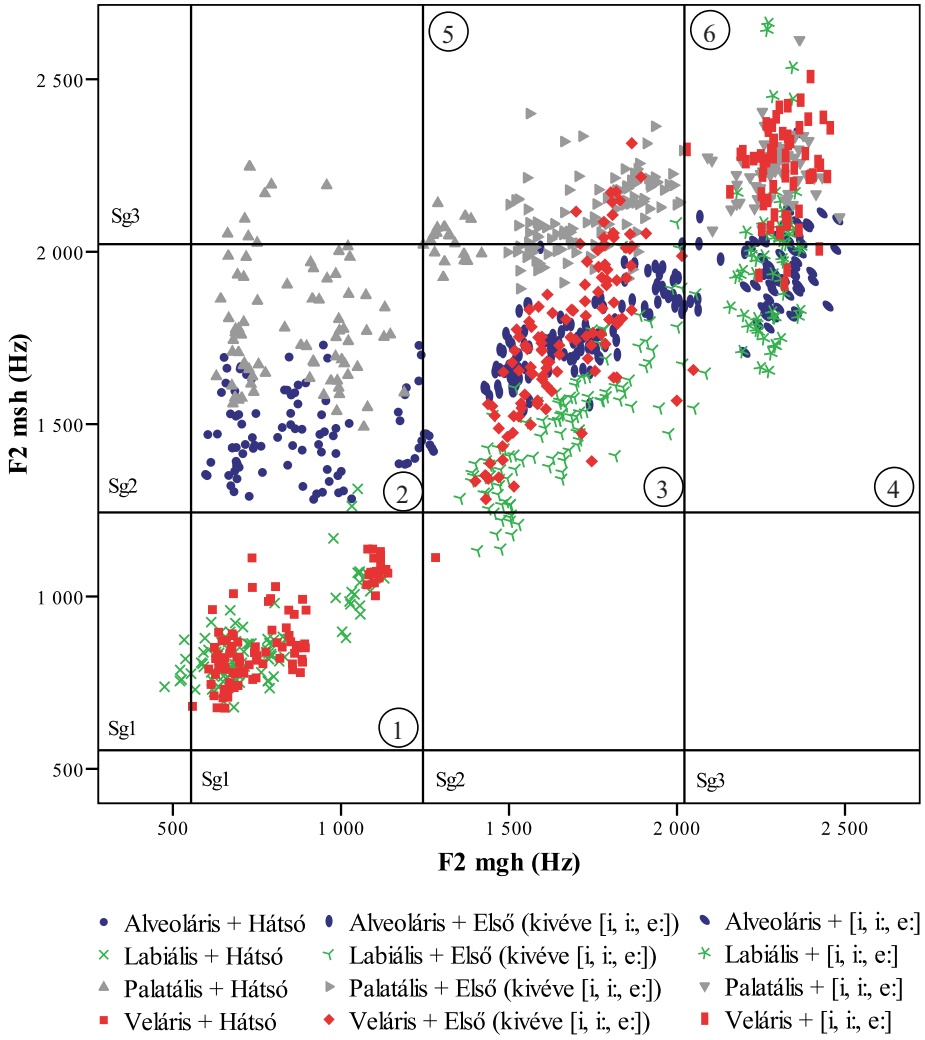
3.1. Locus egyenlet tér

"B1" beszélő $F2$- és SGR-adatai alapján elkészítettük a locus egyenlet terét, amely a 4. ábrán látható módon veti össze az $F2_{msh}$ és $F2_{mgh}$ értékeket. Amint az ábra mutatja, a locus egyenletek terében a CV-kapcsolatok elkülönülnek, az $F2_{msh}$-$F2_{mgh}$ párok a mássalhangzó és a magánhangzó képzési helyének megfelelő csoportokban jelennek meg. Ezen csoportokat a szubglottális rezonanciák határolják: a függőleges $Sg2$ az elöl, illetve hátul képzett magánhangzók közé ékelődik, az $Sg3$ az elöl képzett ajakréses nem alsó magánhangzókat választja el a többi elöl képzettől. A vízszintes $Sg2$ azokat a labiális és veláris mássalhangzókat különíti el, amelyeket hátul képzett magánhangzó követ. A vízszintes $Sg3$ szerepe kisebb mértékű.

Az ábrán hat tartományt jelöltünk számokkal, mindegyik téglalap egy-egy CV-osztálynak felel meg, melyeket az SGR-ek határolnak:

1. Labiális és veláris mássalhangzók, hátul képzett magánhangzókkal
2. Alveoláris és palatális mássalhangzók, hátul képzett magánhangzókkal
3. Alveoláris, labiális és veláris mássalhangzók, elöl képzett magánhangzókkal, kivéve [i, iː, eː]
4. Alveoláris és labiális mássalhangzók, elöl képzett ajakréses nem alsó magánhangzókkal ([i, iː, eː])
5. Palatális mássalhangzók, elöl képzett magánhangzókkal, kivéve [i, iː, eː]
6. Palatális és veláris mássalhangzók, elöl képzett ajakréses nem alsó magánhangzókkal ([i, iː, eː])

Ezek a tartományok részben különböznek az amerikai angol nyelven végzett kísérletben bemutatotthoz képest [10]. Az angol nyelvben a veláris mássalhangzó - elöl képzett magánhangzó kapcsolatokban a $F2_{msh}$ érték nagyobb, mint az $Sg3$. A magyar nyelvre végzett kísérletünkben csak az [i, iː, eː] magánhangzók esetén igaz ez, a többi veláris - első kapcsolatban az $F2_{msh}$ érték kisebb $Sg3$-nál. A palatális mássalhangzókat is vizsgáltuk kísérletünkben, amelyek az angol nyelvben nem fordulnak elő.



4. ábra. "B1" beszélő locus egyenlet tere. 1120 adatpont látható, melyek a logatomokban vizsgált CV-kapcsolatok második formánsai alapján kerültek ábrázolásra. A különböző képzési helyű mássalhangzókat és magánhangzókat eltérő színnel és alakkal jelöltük. A CV-kapcsolatok $F2_{msh}$-$F2_{mgh}$ párjai a mássalhangzó és magánhangzó képzési helyének megfelelően elkülönülő csoportokban jelennek meg, melyeket 1–6 számokkal jelöltünk. A vízszintes és függőleges vonalak a mért szubglottális rezonanciák helyét jelzik.

A 4. ábra alapján az SGR-ek jól elkülöníthető csoportokra osztják a CV-kapcsolatokat második formánsuk alapján. Ez csak néhány kisebb CV-halmaz esetén nem teljesül. A palatális - hátsó kapcsolatok az $F2_{msh}$ irányban nagy teret foglalnak el, néhány adatpont esetén az $F2_{msh}$ érték magasabb az $Sg3$-nál. A palatálisok egy jól elkülönülő csoportja található a függőleges $Sg2$ és a vízszintes $Sg3$ között (melyekre az $F2_{mgh}$ érték nagyobb $Sg2$-nél). A legtöbb palatális - elöl képzett CV-kapcsolat esetén az $F2_{msh}$ nagyobb $Sg3$-nál, míg a "palatális - első (kivéve [i, i:, e:])" csoportnak körülbelül harmada nyúlik $Sg3$ alá. Azokban a CV-kapcsolatokban, amelyekben a magánhangzó [i, i:, e:] volt, a mássalhangzók $F2$ értéke 1600–2600 Hz között szóródik, így a 4-es és 6-os tartomány adatpontjai nehezen elkülöníthetőek (elsősorban a labiálisok találhatók meg a tartomány szélső értékeinél is).

3.2. Mássalhangzóosztályok locus egyenletei

A különböző artikulációs helyű CV-hangkapcsolatokra jellemző formánsmeneteket lineáris regresszió segítségével vizsgáltuk. Ezen locus egyenletek együtthatói, valamint a korrelációs mérőszámok a 3. táblázatban találhatóak. A lineáris regressziós vizsgálatok eredményeként kiderült, hogy a formánsmenetet leíró egyenlet meredeksége (m) és y-metszete (b) eltérő a különböző mássalhangzócsoportokra. Az alveolárisok és palatálisok meredeksége 0,3 körüli, míg a labiálisok és velárisok esetében ez az érték 1-hez közelít. A labiálisok és velárisok $F2_{msh}$ és $F2_{mgh}$ értékei között erősebb a korreláció, melyet a 4. ábrán látható lineárishoz közelítő elhelyezkedésük is mutat.

3. táblázat. A különböző artikulációs helyű mássalhangzóosztályok locus egyenleteinek lineáris regressziós együtthatói és Pearson-féle korrelációs mérőszámai. $F2_{msh} = m \cdot F2_{mgh} + b$

	m	b	R^2
Alveoláris	0,333	1184,350	0,768
Labiális	0,732	301,220	0,915
Palatális	0,307	1552,820	0,628
Veláris	0,912	179,195	0,936

3.3. CV-kapcsolatok osztályozása

A kísérletek során a [10]-ben bemutatott osztályozó algoritmust használtuk fel, melynek segítségével lehetséges a CV-kapcsolatok automatikus osztályozása, szubglottális rezonanciák alapján. Mivel a magyar nyelv mássalhangzó-magánhangzó kapcsolatai részben különböznek az angolétól, az algoritmuson kisebb változtatásokat végeztünk, így például a palatális mássalhangzókat is vizsgáltuk. Az osztályozás során a 4. ábrán látható 1–6 tartományokat vettük

figyelembe. Ezen régiók határait a formánsok ($F2_{msh}$ és $F2_{mgh}$) és szubglottális rezonanciák ($Sg2$ és $Sg3$) közötti egyenlőtlenségek segítségével írhatjuk le, melyek a 4. táblázatban találhatóak.

4. táblázat. A CV-osztályok határait megadó egyenlőtlenségek.

Tartomány	CV-osztály	1. egyenlőtlenség	2. egyenlőtlenség
1	Labiális, Veláris + Hátsó	$F2_{msh} < Sg2$	$F2_{mgh} < Sg2$
2	Alveoláris, Palatális + Hátsó	$Sg2 < F2_{msh} < Sg3$	$F2_{mgh} < Sg2$
3	Alveoláris, Labiális, Veláris + Első	$F2_{msh} < Sg3$	$Sg2 < F2_{mgh} < Sg3$
4	Alveoláris, Labiális + [i, iː, eː]	$F2_{msh} < Sg3$	$Sg3 < F2_{mgh}$
5	Palatális + Első	$Sg3 < F2_{msh}$	$Sg2 < F2_{mgh} < Sg3$
6	Palatális, Veláris + [i, iː, eː]	$Sg3 < F2_{msh}$	$Sg3 < F2_{mgh}$

Célunk az 1–6 tartományok optimális klasszifikációjának megtalálása volt. Ennek érdekében az $\widetilde{Sg2}$ értékét 1000 és 1500 Hz között folyamatosan növelve vizsgáltuk az osztályozás találati és téves riasztási arányait. Az $\widetilde{Sg3}$ értékét "B1" beszélő $Sg2/Sg3$ aránya alapján számítottuk ($\widetilde{Sg3} = 1,6254 \cdot \widetilde{Sg2}$).

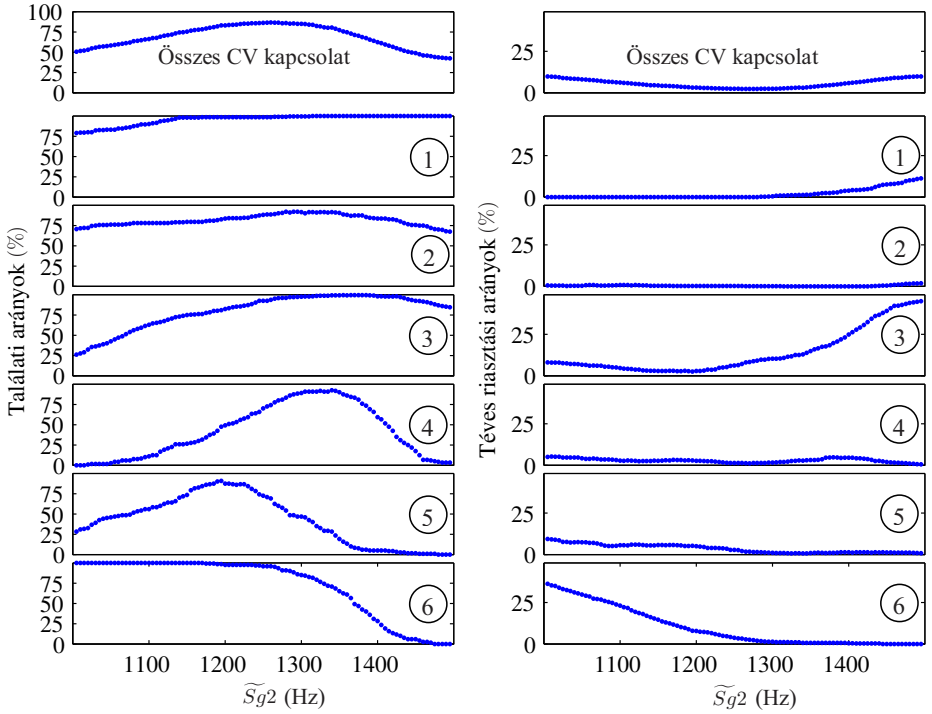
Az 5. ábra mutatja az osztályozás találati és téves riasztási arányait az $\widetilde{Sg2}$ függvényében. Az összes CV-kapcsolatra vonatkozó optimális klasszifikáció 1260 Hz esetén történik. Emellett eltérő a legmagasabb találati arány a 4-es tartomány esetén. A 4. ábrát megvizsgálva azt láthatjuk, hogy a 4-es és 6-os tartományban lévő CV-kapcsolatok között sok az átfedés, emiatt fordulhat elő, hogy a 4-es tartomány esetén az optimális szeparáció magasabb $\widetilde{Sg2}$ esetén (kb. 1350 Hz) megy végbe.

Az optimális osztályozás ($\widetilde{Sg2}$=1260 Hz) esetén az összesített eredményeket az 5. táblázat mutatja. Az összes CV-kapcsolatra a találati arány 86,6%, míg a téves riasztási arány 2,3%. A 4-es tartomány esetén a legalacsonyabb a találati arány, a korábban leírtak miatt.

Ezután megvizsgáltuk az osztályozást "B1" beszélő szubglottális jelében manuálisan mért SGR-értékek mediánjai alapján ($Sg2$=1244 Hz, $Sg3$=2022 Hz). A találati és téves riasztási arányokat a 6. táblázat mutatja külön-külön az egyes kategóriákra, illetve összesítve is. Az összes vizsgált CV-kapcsolatra a találati arány 85,5%, míg a téves riasztási arány 2,4%. Ezek az értékek nagyon közel vannak az optimális elválasztáshoz, mivel a mért $Sg2$ értéke (1244 Hz) szinte megegyezik az optimális osztályozás során kapott $\widetilde{Sg2}$-vel (1260 Hz).

4. Következtetések

Jelen kutatás során egy kísérletsorozatot mutattunk be, amely egy beszélő logatomfelvételeiből származó CV-kapcsolatok által definiált locus egyenlet teret elemzett, illetve vizsgálta a szubglottális rezonanciák által okozott elválasztást. Először megvizsgáltuk "B1" beszélő locus egyenlet terét, majd a CV-csoportok



5. ábra. Az osztályozás eredménye az $\widetilde{Sg}2$ függvényében. A legfelső részábrák mutatják az összesített találati és téves riasztási arányokat, az alsóbb ábrák pedig az 1–6 tartományokhoz tartozó eredményeket.

5. táblázat. CV-kapcsolatok osztályozásának találati és téves riasztási arányai, az optimális $\widetilde{Sg}2$ értékkel számolva. ($\widetilde{Sg}2$ = 1260 Hz, CV jelöli az összes eredményt, 1–6 az egyes tartományokat.)

	CV	1	2	3	4	5	6
Találati arány	86,6%	98,5%	88,5%	93,6%	74,2%	69,2%	95,8%
Téves riasztási arány	2,3%	0%	0,2%	7,1%	1,3%	2,1%	3,2%

6. táblázat. CV-kapcsolatok osztályozásának találati és téves riasztási arányai, a mért SGR-értékekkel számolva. ($Sg2$=1244 Hz)

	CV	1	2	3	4	5	6
Találati arány	85,5%	98,5%	87%	91,7%	65%	74,2%	96,7%
Téves riasztási arány	2,4%	0%	0,2%	5,7%	1,3%	2,9%	4,3%

artikulációs helye alapján hat tartományt definiáltunk, melyek hipotéziseink szerint az SGR-ek segítségével elkülöníthetőek. A különböző mássalhangzó osztályok lineáris regressziós egyenleteit (ún. locus egyenlet) is vizsgáltuk. A [10]-ben bemutatott osztályozó algoritmust a magyar nyelv hangjainak megfelelően módosítottuk, és alkalmaztuk a CV-adathalmazra. A mért SGR-értékek alapján történő osztályozást összehasonlítottuk az optimális találati arányt és téves riasztási arányt okozó klasszifikációval. A szubglottális rezonancia alapú automatikus mássalhangzó-magánhangzó hangkapcsolat osztályozás "B1" beszélő esetén az optimálishoz képest mindössze 1%-kal alacsonyabb a találati arányt eredményezett. Ez a [10]-ben bemutatott amerikai angol kísérlethez hasonló eredményt jelent.

A további kutatás célja más SGR mérési lehetőségek keresése. [10] szerint a beszédfelvételből is meghatározható az $Sg2$ értéke, de ez viszonylag pontatlan, távol van a gyorsulásmérővel mért SGR-értékektől. [18]-ban egy egyedi eszközt készítettek a szubglottális jel felvételére, és az alsó légúti rezonanciák automatikus mérésével kísérleteztek, azonban ez távol volt a manuálisan mért értékektől. Amennyiben a szubglottális rezonanciák mérése egyszerűbben megoldható lesz, az itt bemutatott SGR alapú CV-osztályozás tetszőleges beszélő esetén alkalmazhatóvá válik.

Az itt bemutatott kísérlet során csak egy magyar nyelvű beszélő felvételeit vizsgáltuk. A továbbiakban érdemes lenne több beszélő hangfelvételeit is elemezni, megvizsgálva mások locus egyenlet terének és szubglottális rezonanciáinak kapcsolatát.

A szubglottális rezonanciákat már sikerrel alkalmazták a beszélőnormalizálásban [16,17], az eredményeink ezen kívül hozzájárulhatnak a beszédfelismeréshez is.

5. Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki a támogatóknak (NKFP 2/034/2004, Jedlik OM-00102/2007, TÁMOP-4.2.2-08/1/KMR-2008-0007), a kísérletben részt vevő adatközlőnek, valamint Bőhm Tamásnak a hangfelvételek rendelkezésre bocsátásáért. Külön köszönet illeti Steven M. Lulichot a szubglottális rezonanciák témájának részletes ismertetéséért, valamint a cikk javítására irányuló javaslataiért és megjegyzéseiért.

Hivatkozások

1. Boersma, P., Weenink, D.: Praat (Version 5.1.19). http://www.praat.org (2009)
2. Chi, X., Sonderegger, M.: Subglottal coupling and its influence on vowel formants. JASA 122 (2007) 1735–1745
3. Csapó, T. G., Bárkányi, Zs., Gráczi, T. E., Bőhm, T., Lulich, S. M.: Relation of formants and subglottal resonances in Hungarian vowels. In: Proc. Interspeech (2009) 484–487
4. Gray, H.: Anatomy of the human body. Philadelphia: Lea & Febiger. (1918)

5. Gósy, M.: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest. (2004)
6. Jung, Y.: Subglottal effects on the vowels across language: Preliminary study on Korean. JASA 125 (2009) 2638
7. Lulich, S. M., Bachrach, A., Malyska, N.: A role for the second subglottal resonance in lexical access. JASA 122 (2007) 2320–2327
8. Lulich, S. M.: Subglottal resonances and distinctive features. J. Phon. doi:10.1016/j.wocn.2008.10.006 (2009)
9. Lulich, S. M.: On the relation between locus equations and subglottal resonances. POMA 5, 060003 (2009)
10. Lulich, S. M., Chen, N. F.: Automatic classification of consonant-vowel transitions based on subglottal resonances and the second formant, POMA 6, 060005, (2009)
11. Madsack, A., Lulich, S. M., Wokurek, W., Dogil, G.: Subglottal resonances and vowel formant variability: A case study of High German monophthongs and Swabian diphthongs. In: Proc. LabPhon11 (2008) 91–92
12. Mihajlik, P., Révész, T., Tatai, P.: Phonetic Transcription in Automatic Speech Recognition. Acta Linguistica Hungarica, Vol. 49. (3-4), (2002) 407–425
13. Sjölander, K., Beskow, J.: Wavesurfer (Version 1.8.5). http://www.speech.kth.se/wavesurfer (2009)
14. Stevens, K. N.: On the quantal nature of speech, J. Phon. 17 (1989) 3–45
15. Stevens, K. N.: Acoustic Phonetics. MIT Press: Cambridge, MA. (1998)
16. Wang, S., Lulich, S. M. Alwan, A.: A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation. In: Proc. Interspeech (2008) 1717–1720
17. Wang, S., Lee, Y.-H. Alwan, A.: Bark-shift based nonlinear speaker normalization using the second subglottal resonance. In: Proc. Interspeech (2009) 1619–1622
18. Wokurek, W. Madsack, A.: Comparison of Manual and Automated Estimates of Subglottal Resonances. In: Proc. Interspeech (2009) 1671–1674

A magyar nyelv betűstatisztikája beszédfeldolgozási szempontok figyelembevételével

Zainkó Csaba

Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
zainko@tmit.bme.hu

Kivonat: A cikkben bemutatok egy új típusú betűstatisztikát, amely a klasszikus 44 betűs magyar ábécén alapuló eljárás továbbfejlesztése és egyesíti a betű- és a hangstatisztika előnyeit. A betűstatisztika készítését olyan módon egészítem ki, hogy figyelembe veszem a beszédfeldolgozás igényeit is. A módszer megkülönböztet betű szinten olyan jelenségeket is, amelyek csak a hangstatisztika szintjén lehet kezelni. Az új módszert a Magyar Nemzeti Szövegtáron tesztelem, összehasonlítom a módszert a klasszikus betűstatisztikával és a beszédfeldolgozásban használt hangstatisztikával.

1 Bevezetés

A magyar nyelvre sok célból készítenek különböző statisztikákat, például betű- és hangstatisztikákat, hangkapcsolódások statisztikáját vagy szótag- és szóstatisztikákat. Az első átfogó hangstatisztikát Szende Tamás közölte 1976-ban [8]. Betűstatisztikákat használtak például a titkosítás tudományában, de nyelv- és beszédfeldolgozási kutatásokban is fontos szerepe van. A betűstatisztikák általában a magyar 40 vagy a kiterjesztett 44 betűs ábécéből indulnak ki. A hangstatisztika két úton készülhet. Egyik módszer, hogy valamilyen hangzó anyagot gépi és kézi módszerrel annotálunk, a másik, hogy szövegből kiindulva fonetikus átiratot készítünk.

A jelen cikkben bemutatok egy újfajta megközelítést, amely alapjaiban egy betűstatisztika, de felhasznál olyan információkat is, amelyek általában csak a fonetikus átiratban állnak rendelkezésre. Ez azt jelenti, hogy figyelembe vesszük az osztályozáskor, a betűsorozat fonetikus reprezentációját is. Ehhez a betűstatisztikához újra definiálom a betű ilyen értelmű fogalmát. Például a *pech* szó klasszikus értelemben vett ch betűkapcsolatát az *sz* betűhöz hasonlóan kezelem, egy külön két karakterből álló betűnek tekintem, és másként kezelem például a *lánchíd* ch betűkapcsolatától, ahol külön c és h betű szerepel. A betű szintű osztályozás miatt viszont megmaradnak olyan információk is, amelyek a fonetikus átírás közben elvesznek. Például rendelkezésre áll az új típusú betűstatisztikában a /j/ hangként kimondott j és ly betű, vagy az /i/ hangként kimondott i és történelmi nevek végén gyakran szereplő y betű.

A cikk második részében ismertetem a Magyar Nemzeti Szövegtár állományaiból készített különböző statisztikákat, és összehasonlítom ezek eredményeit egymással.

A dolgozatban a hangokat ferde zárójelek közötti betűvel jelölöm, vagyis azzal a betűképpel, amelyik kimondása az adott hanghoz tartozik.

A kutatást a TELEAUTÓ projekten keresztül a Jedlik program támogatta.

2 Motiváció

A magyar helyesírás szabályai [1] szerint az ábécénk 40 betűt tartalmaz, amely egy vagy több írásjegyből áll (a, á, b, c, cs stb.). A számítógépes dokumentumokban az írásjegyeket karakterek formájában tároljuk, amelyek szintén egy vagy több karakteres betűt képviselnek. A továbbiakban – az egyszerűbb szóhasználat érdekében – a karakter elnevezést használom az írásjegy értelemben is. A szabályzat az ábécébe sorolja még – az idegen szavakban gyakran előforduló – q, w, x, y betűket is. Ez a 44 betűs ábécé forma, amit a gyakorlatban széles körben használnak.

A helyesírási szabályzat rendelkezik a régi magyar és az idegen eredetű nevek írásáról is. Ezekben lehetnek olyan betűk, amelyek két vagy több karakterből is állhatnak, ennek ellenére a karaktereket különálló betűkként kell a szabályok szerint kezelni bizonyos esetekben. Például a *Czuczor* /cucor/ családnevet betűrendbe soroláskor c + z betűk szerint kell besorolni, annak ellenére, hogy mai formában c betű lenne, illetve /c/ hangnak is ejtjük. A szabályok más esetekben viszont azt mondják ki, hogy az ilyen több karakterű betűket egyetlen betűként kell kezelni. Ilyen az elválasztási rendszer, amelynek szabályai előírják, hogy ezek a betűk nem elválaszthatóak. Például: *Ri-chárd*, *Mün-chen*, *Ben-czúr*.

A beszéd gépi feldolgozása során sok esetben célszerű a hangzó anyag mellett megadni annak valamilyen hangszintű, írott reprezentációját is. Ez az írott forma származhat a fonetikus átírásból, amely az elhangozó beszédhangok leirata hangjelölési szimbólumokkal megvalósítva. Sok esetben a feldolgozandó beszéd valamilyen szöveg felolvasása során keletkezik, az írott szöveg és a beszéd kapcsolata ilyenkor jól szinkronizálható, a fonetikus gépi algoritmusokkal és utólagos manuális feldolgozási lépésekkel elkészíthető. Az ilyen átírási folyamat összetett, időigényes és egyszerűsítéseket is tartalmazhat (például hangok különböző variánsait nem kezeli).

Az új típusú betűstatisztika használható kutatási feladatokra, a magyar szöveges állományok statisztikai tulajdonságainak vizsgálatára. Például: „Milyen gyakran jelent a ch betűkapcsolat /h/ hangot?” Felhasználható beszédadatbázisok készítésekor felolvasandó szövegállományok elemzésére, válogatására. Például megbecsülhető, hogy egy adott szöveg felolvasása esetén, a felolvasott szöveg egy kiválasztott hangból elegendő számút fog-e tartalmazni.

3 Módszer

A betűstatisztika készítésekor betűként a következő tulajdonságú karaktert vagy karaktersorozatokat értem:

- a 44 betűs magyar ábécének tagja, pl.: a, á, b, c, cs, …, sz, zs …
- a régi magyar családnevekben gyakran előforduló kettős betűk, pl.: cz, ch ejtése: /cs/ …
- idegen szavakban előforduló több karakteres betűk pl.: sch, ch ejtése:/h/ …
- kiejtve eltérő hangokhoz tartozhat. például.: h betű, amelynek kétfajta ejtését különböztetjük meg /$h_{néma}$, $h_{zöngés}$/ …

Néhány betűnél a különböző változatokat a szerint különböztetjük meg, hogy milyen hang keletkezik az adott szó kimondása esetén. Fontos megjegyezni, hogy ezek az osztályozások elsősorban beszédtechnológiai szempontok figyelembevételével történnek, nyelvészeti szempontból bizonyos döntések indokolatlanoknak tűnhetnek.

Szövegnek tekintem az általános szabályok szerint leírt magyar szövegeket, amelyek tartalmazhatnak számokat, írásjeleket és egyéb karaktereket. A statisztika készítésekor a nem betű típusú karaktereket figyelmen kívül hagyom (számok, relációs jelek stb.).

A szöveg feldolgozása szabályalapú algoritmussal történik, amely felhasználja a Profivox szövegfelolvasó rendszer fonetikai átíróját és szabálygyűjteményét [7], valamint különböző egyéb szótárakat is. Ilyenek a nagyméretű, magyar elektronikus kiejtési szótár [2], a névmondó tulajdonnév kiejtési szótárai [6], Huhyphn – magyar elválasztásiminta-gyűjtemény szószedete [5].

Az algoritmus a Profivox szövegfelolvasó fonetikai átírója szabálygyűjteményének egy részhalmazát használja. Ezek nagy részét a kettő vagy több karakterből álló betűk meghatározására vonatkozó szabályok teszik ki. A további szabályok azokra a betűkre vonatkoznak, amelyek kiejtésekor nem a betűhöz tartozó, nyelvileg szabályos fonéma realizáció (hang) keletkezik, hanem annak valamelyik speciális variánsa. Ilyen például a szóvégi zöngétlen /j/ hang (*lépj, hívj*). A szabályrendszer a magyar elektronikus kiejtési szótár információival van kiegészítve. Ez a szótár 1,5 millió magyar szóalak helyesírását és kiejtését adja meg párhuzamosított formában.

A régi magyar családnevekben előforduló betűk kezelését a névmondó szótár segítségével dolgozza fel az algoritmus. A régi betűvariációk nagy száma miatt, csak a gyakran előforduló személynevekben található eseteket kezeli a rendszer.

A magyarra jellemző a szóösszetétel. Az összetett szavak határán előfordulnak olyan karakterkombinációk, amelyek megegyeznek több karakteres betűkkel, de valójában nem azok. Ilyen például a *malacsült*, amelyben c + s betű található és nem cs. Az ilyen félreértelmezések elkerülésére az algoritmus a Huhyphn elvalasztásiminta-gyűjteményben található szószedetet használja. Az szószedet tartalmazza a szavak elválasztási lehetőségeit. Az algoritmus kihasználja ennek a szószedetnek azon tulajdonságát, hogy a helyesírási szabályok nem engedik meg a több karakteres betűkön belüli elválasztást, így az elválasztási helyeken korlátozza ezek hibás észlelését. Például a ma-lac-sült szó elválasztásából látható, hogy cs betű nem szerepelhet ebben a szóban. Az algoritmus figyelembe veszik a két karakterből álló hosszú változatát is pl.: tty, ssz, nny. Ezek két betűként szerepelnek a statisztikában zzs -> zs + zs.

A Profivox szabályokat tartalmaz a magánhangzók rövidülésére is, amely szintén használható a betűstatisztika finomítására. A magánhangzó rövidülése nagy változatosságot mutat a különböző beszélőknél, ezért ezek az információk nem adnak pontos eredményt, de tájékoztató adatnak megfelelnek. A további felhasználásuk esetében ezt figyelembe kell venni.

3.1 Speciális betűk, hangok

A ch betű az eredetétől függően /h/, /cs/ vagy /k/ hangot is jelölhet. Ennek megfelelően háromféle jelölést alkalmazunk: ch_h, ch_cs, ch_k (ezeknél a jelöléseknél az aláhúzás utáni betű jelöli a kiejtési formát)

A h betű hangalakja is többféle lehet. Néma /h/ valósul meg például a *cseh* /cse/ szóban. Csak ragozott formánál ejtjük a /h/ hangot (*csehül* /csehül/) Jelölése: $h_{néma}$

A h betű másik értelmezési formája a zöngés /h/ hang. Jelölése: $h_{zöngés}$

A j betű egyes esetekben zöngétlen /j/ hangként jelenik. Jelölése: $j_{zöngétlen}$

Az sch német eredetű betű is gyakran előfordul, a 3 karakteres hossza miatt fontos a külön kezelése. Jelölése: sch

Az y betű többnyire régi nevekben és idegen eredetű szavakban fordul elő általában /i/ vagy /j/ hangként valósul meg ejtéskor. Jelölésük: y_i, y_j

Az olyan rövid magán hangzókat is megkülönböztetjük, amely az átlagos hanghossznál rövidebbek. Jelölése: $a_{röv}$, $á_{röv}$ …

A régi írásmódú betűk jelölése: cz_c, ts_cs, eö_ö, tz_c, ck_k

3.2 Nyelvi anyag

A statisztikai elemzésekhez a Magyar Nemzeti Szövegtár (MNSZ) [4] anyagát használtam fel. A szövegtár 187,6 millió szövegszót tartalmaz, 5 nagyobb témát dolgoz fel. Tartalmaz sajtószövegeket, szépirodalmi műveket, tudományos, hivatalos és személyes szövegeket. A vizsgálatokhoz a teljes szövegtárat felhasználtam. A karakter és betűstatisztikához a vizsgált leghosszabb betűsorozat a szó volt, a beszédet reprezentáló fonetikai hangszimbólumok statisztikájához mondatokat használtam, ugyanis figyelembe vettem a szavak határán történő hangváltozási jelenségeket is.

3.3 A módszer korlátai

A statisztika egy gépileg gyűjtött és ellenőrzött szövegen alapul, amely tartalmaz hibákat. A összeállított szövegek mérete miatt manuális ellenőrzés nem jöhet szóba.

A felhasznált kivételszótárak szintén részben gépi módszereken alapulnak, egy része manuálisan ellenőrzött, de ennek ellenére tartalmazhatnak hibákat vagy hiányosak is lehetnek.

A kiejtéshez kapcsolódó szabályok a magyar nyelvi normát képviselik.

A vizsgált betűk meghatározása önkényes, a beszédfeldolgozás egyes szempontjait tartotta szem előtt, más felhasználás esetén a vizsgált betűk kiválasztása korlátozást jelenthet. Például a régi írásmódú betűk vizsgálata nem teljes körű, amely beszéd szempontjából megengedhető, de névelemzés esetén már további finomítás szükséges.

4 Eredmények

A vizsgált szöveg statisztikáját 3 formában készítettem el és az 1. táblázatban foglaltam össze. Az első két oszlop a karakterstatisztika, a második kettő a betűstatisztika, az utolsó két oszlop a hangstatisztika. A táblázatban szereplő számértékek megadják, hogy átlagosan 1000 elemből hány adott elem fordul elő. A hangstatisztika esetén a könnyebb összehasonlíthatóság miatt a betűképpel jelöltem a hangokat. A hangstatisztika teljesen gép módszerrel készült, manuális ellenőrzés nem történt a fonetikus átiratokon. A hangok esetében nincsenek megkülönböztetve a speciális esetek, variánsok. Az üres mezők azt jelentik, hogy az adott típusú statisztikában olyan elem nem szerepelt.

A betűstatisztika esetén az 1. táblázat csak a 44 betűs ábécében szereplő betűket tartalmazza, a speciális betűket a 2. táblázat tartalmazza. A második táblázatban a számértékek 1 millió elemre vonatkoznak.

A különböző statisztikák elkészítése nagyságrendileg eltérő erőforrást igényelt. A karakterstatisztika másodpercek alatt elkészült, a betűstatisztika több tíz perc, míg a hangstatisztika elkészítése 3-4 órát vett igénybe. A karakterstatisztika használata tehát akkor előnyös, ha gyors működés elengedhetetlen.

A karakterstatisztika hátránya, hogy csak 36 karakterre tartalmaz információkat, azokat is erősen torzítva. Az 1. táblázat s karakteréhez és betűjéhez tartozó gyakoriságokat összevetve látható, hogy a s karakter jóval gyakrabban fordul elő, mint az s betű, amelyet a kettős betűk szétdarabolása okozott. A karakterstatisztika tehát nyelv- és beszédfeldolgozási szempontokból egyáltalán nem vagy alig használható. Ennek ellenére az egyszerű programozhatósága miatt sok helyen használják betűstatisztika helyett.

Az itt szereplő hangstatisztika egyszerűsítéseket tartalmaz, csak 38 beszédhang szerepel benne. Ennek ellenére a karakterstatisztikához képest, jobban tükrözi a nyelv tulajdonságait, mert az egyszerűsítések fonetikailag megengedhető helyeken történtek. A betűstatisztikával összehasonlítva az elemek hasonló gyakorisággal szerepelnek, néhány esetben van csak eltérés, például az sz betű és az /sz/ hang között. Gósy [3] spontán beszédre készített hangstatisztikát, amelyben a magánhangzó-mássalhangzó arány 43% és 57% volt. Itt ez az arány 42% és 58% volt. A leggyakoribb hangokat összehasonlítva szintén hasonló számokat kaptunk, például a leggyakoribb /e/ hang gyakorisága Gósy statisztikájában 11.4%., itt 10.7%.

1. táblázat: karakter-, betű- és hangstatisztika

Karakter	1000-ből	Betű	1000-ből	Hang	1000-ből
a	89.37	a	92.85	/a/	90.21
á	35.95	á	37.58	/á/	37.99
b	19.66	b	20.56	/b/	18.28
c	7.64	c	3.97	/c/	6.10
		cs	3.91	/cs/	3.85
d	19.74	d	20.42	/d/	19.49
		dz	0.03		
		dzs	0.02		
e	98.70	e	101.31	/e/	106.59
é	33.46	é	35.02	/é/	35.69
f	9.18	f	9.59	/f/	9.04
g	33.80	g	22.69	/g/	19.82
		gy	12.70	/gy/	11.45
h	15.32	h	13.07	/h/	17.56
i	44.06	i	46.39	/i/	47.28
í	5.82	í	5.60	/í/	5.51
j	11.19	j	11.98	/j/	14.27
k	49.22	k	51.46	/k/	53.63
l	62.27	l	60.78	/l/	58.46
		ly	3.77		
m	35.00	m	36.56	/m/	36.61
n	58.12	n	53.78	/n/	54.37
		ny	7.02	/ny/	8.21
o	40.93	o	40.21	/o/	42.26
ó	10.03	ó	10.49	/ó/	9.95
ö	10.90	ö	11.39	/ö/	11.75
ő	8.94	ő	9.35	/ő/	9.68
p	11.14	p	11.65	/p/	12.42
q	0.04	q	0.04		
r	42.47	r	44.41	/r/	44.02
s	60.35	s	39.08	/s/	35.89
		sz	19.27	/sz/	24.55
t	79.42	t	82.72	/t/	81.58
		ty	0.27	/ty/	4.09
u	10.18	u	10.73	/u/	11.19
ú	3.01	ú	3.06	/ú/	2.69
ü	5.51	ü	5.85	/ü/	5.54
ű	1.86	ű	1.86	/ű/	1.74
v	19.89	v	20.80	/v/	21.52
w	0.28	w	0.29		
x	0.36	x	0.38		
y	22.71	y	0.21		
z	43.48	z	26.48	/z/	24.51
		zs	0.73	/zs/	2.21

2. táblázat: betűstatisztika speciális betűkre

hang	1000000-ból
ch_cs	28.12
ch_h	129.04
ch_k	2.33
ck_k	4.96
cz_c	25.24
eö_ö	7.16
$h_{néma}$	60.27
$h_{zöngés}$	2470.19
$a_{röv}$	627.99
$á_{röv}$	111.70
$e_{röv}$	2010.52
$é_{röv}$	15.02
$i_{röv}$	1000.29
$o_{röv}$	2667.03
$ö_{röv}$	36.57
$u_{röv}$	144.33
sch	85.09
ts_cs	34.25
tz_c	11.84
y_i	65.27
y_j	111.27
$j_{zöngétlen}$	3.59

A betűstatisztika 1. táblázatban szereplő részén mind a 44 betű statisztikáját megtalálhatjuk. Ez a 44 betű az összes betűstatisztikában szereplő betű 99%-at adja, a speciális betűk csak 1%-ot tesztnek ki a vizsgált szövegekben. A ábécé betűi közül a dz, dzs, q szerepel nagyon ritkán, a 1 millió szóban átlagosan 20-40 db található meg. Leggyakrabban a vártnak megfelelően az e betű szerepelt.

A 2. táblázatban szereplő rövid magánhangzók közül a rövid á 1 millió szóból átlagosan 111-szer szerepel. Ez a kis érték több okra vezethető vissza. A rövid /á/ hang a *fájl*, *bájt* szavakban található, amit gyakoribbnak volt várható. Egyik ok, hogy a szöveg jelentős részben tartalmaz irodalmi alkotásokat, amelyekben ez a szó nem szerepel. A másik ok, hogy a szöveggyűjteményben nagy számban helyesírásilag hibásan szerepelnek a *bájt* és *fáj*l szavak, az angol *file* és *byte* formában.

A szó végén szereplő zöngétlen /j/ hang kis számban szerepel a szövegekben. Ennek oka az lehet, hogy a felszólító módú igék írott szövegben kevésbé gyakoriak, inkább a beszélt nyelvben találhatóak meg.

Az y betű /j/ hangként való realizációja gyakoribb, mint az /i/ hangként való megjelenése. Ez abból adódik, hogy idegen nevek többször szerepelnek (például Toyota) mint a történelmi nevek (például Desseffy).

A ch betű leggyakrabban /h/ hangként jelenik meg, majd /cs/ hang a második leggyakoribb formája, /k/ hangként ritkán ejtjük.

A 3. táblázatban a betűstatisztika található gyakorisági sorrendben.

3. táblázat: Betűstatisztika gyakorisági sorrendben

Betű	db/1000	Betű	db/1000	Betű	db/1000	Betű	db/1000
e	101.31	d	20.42	ú	3.06	y_i	0.065
a	92.85	sz	19.27	$o_{röv}$	2.67	$h_{néma}$	0.060
t	82.72	h	13.07	$h_{zöngés}$	2.47	q	0.040
l	60.78	gy	12.70	$e_{röv}$	2.01	$ö_{röv}$	0.037
n	53.78	j	11.98	ű	1.86	ts_cs	0.034
k	51.46	p	11.65	$i_{röv}$	1.00	ch_cs	0.028
i	46.39	ö	11.39	zs	0.73	dz	0.026
r	44.41	u	10.73	$a_{röv}$	0.63	cz_c	0.025
o	40.21	ó	10.49	x	0.38	dzs	0.017
s	39.08	f	9.59	w	0.29	$é_{röv}$	0.015
á	37.58	ő	9.35	ty	0.27	tz_c	0.012
m	36.56	ny	7.02	y	0.21	eö_ö	0.007
é	35.02	ü	5.85	$u_{röv}$	0.14	ck_k	0.005
z	26.48	í	5.60	ch_h	0.13	$j_{zöngétlen}$	0.004
g	22.69	c	3.97	$á_{röv}$	0.11	ch_k	0.002
v	20.80	cs	3.91	y_j	0.11		
b	20.56	ly	3.77	sch	0.09		

5 Összegzés

Az új típusú betűstatisztika alkalmas arra, hogy szövegekről, korpuszokról olyan statisztikai információkhoz jussunk egy lépésben, amelyhez csak a klasszikus betűstatisztika és a hangstatisztika (fonémastatisztika) együttes elemzésével juthatunk. Megadtam egy lehetséges betűosztályozást, amellyel egy kibővített statisztikát lehet készíteni magyar nyelvre. A cikkben továbbá összehasonlító elemzést adtam karakterstatisztikára, az általam módosított értelmű betűstatisztikára és az ugyanazon szövegkorpuszból készített hangstatisztikára. A statisztikák a Magyar Nemzeti Szövegtár alapján készültek.

Hivatkozások

1. A magyar helyesírás szabályai. MTA Budapest: Akadémiai. Kiadó (1985)
2. Abari K., Olaszy G., Kiss G., Zainkó Cs.: Magyar kiejtési szótár az Interneten. In: Alexin Z., Csendes D. (szerk.) MSZNY (2006) 223-230
3. Gósy M.: Fonetika, a beszéd tudománya. Osiris. Budapest (2004)
4. Magyar Nemzeti Szövegtár. MTA – Nyelvtudományi Intézet http://corpus.nytud.hu/mnsz/
5. Nagy B.: Huhyphn: magyar elválasztásiminta-gyűjtemény. http://www.tipogral.hu/ (2008)
6. Németh G., Zainkó Cs., Kiss G., Fék M., Olaszy G., Gordos G.: Language Processing for Name and Address Reading in Hungarian In: IEEE NLP-KE Beijing, Kína (2003) 238-243
7. Olaszy G., Németh G., Olaszi P., Kiss G., Zainkó Cs., Gordos G.: Profivox - a Hungarian TTS System for Telecommunications Applications In: IJST 3-4: 201-215 (2000)
8. Szende T.: A beszédfolyamat alaptényezői. Akadémiai Kiadó (1976)

Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel

Tóth Bálint, Németh Géza

Távközlési és Médiainformatikai Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem
1117 Budapest, Magyar Tudósok krt. 2.
{toth.b, nemeth}@tmit.bme.hu

Kivonat: Napjainkban számos automatikus szövegfelolvasási módszer létezik, de az elmúlt években a legnagyobb figyelmet a statisztikai parametrikus beszédkeltési módszer, ezen belül is a rejtett Markov-modell (Hidden Markov Model, HMM) alapú szövegfelolvasás kapta. A HMM-alapú szövegfelolvasás minősége megközelíti a manapság legjobbnak számító elemkiválasztásos szintézisét, és ezen túl számos előnnyel rendelkezik: adatbázisa kevés helyet foglal el, lehetséges új hangokat külön felvételek nélkül létrehozni, érzelmeket kifejezni vele, és már néhány mondatnyi felvétel esetén is lehetséges az adott beszélő hangkarakterét visszaadni. Jelen cikkben bemutatjuk a HMM-alapú beszédkeltés alapjait, a beszélőadaptációjának lehetőségeit, a magyar nyelvre elkészült beszélőfüggetlen HMM adatbázist és a beszélőadaptáció folyamatát félig spontán magyar beszéd esetén. Az eredmények kiértékelése céljából meghallgatásos tesztet végzünk négy különböző hang adaptációja esetén, melyeket szintén ismertetünk a cikkünkben.

1 Bevezetés

Napjainkban már számos lehetőség áll rendelkezésre gépi szövegfelolvasásra: a beszédkeltés mechanizmusát modellező formáns és artikulációs szintézistől kezdve a diádos és triádos hullámforma összefűzéses szintézisen át a hullámforma-elemkiválasztó (korpusz) szintézisig. A beszédszintetizátor által kiadott hangot érthetőség és természetesség szempontjából szokták minősíteni, a technológiai megoldást pedig olyan további műszaki paraméterekkel jellemzik, mint például számításigény, tárhely igény. Napjaink vezető technológiája a korpusz alapú hullámforma-elemkiválasztásos módszer, azonban adatbázisának a mérete igen nagy (gigabyte-os nagyságrendbe esik), az elemkiválasztás sok számítási kapacitást igényel és a beszélő hangkarakterét az adatbázis határozza meg. Így új beszédhangokhoz új, több gigabyte-os stúdióminőségű hangfelvételek vagy beszélő transzformációs eljárások szükségesek, melyek minőségromláshoz vezetnek.

A statisztikai parametrikus szintézis, ezen belül is a beszédfelismerő rendszerek technológiáját használó rejtett Markov-modell (Hidden Markov Model, HMM) alapú beszédszintézis [1] igen jelentős fejődésen ment keresztül az elmúlt években. Az általa generált beszéd minősége és természetessége megközelíti a korpuszos rendsze-

rek minőségét, de emellett számos előnnyel rendelkezik: a futáshoz szükséges adatbázis mérete kicsi (néhány megabyte) [2], könnyen lehet vele új beszédhangokat létrehozni [3], alkalmas érzelemkifejezésre [4] és beszélőadaptációra [5], [6]. A HMM-alapú beszédszintézis beszédépítési eljárása lényegesen különbözik az elemkiválasztásos technológiáktól, mivel nem közvetlenül a hullámformával dolgozik, hanem a hullámformából spektrális és prozódiai jellemzők sokaságának kinyerése után (tanító fázis) ezekből válogatva alakítja ki a szintézishez szükséges adatsorozatot. A válogatást a tanítás során előállított rejtett Markov-modellek végzik.

A HMM-ek tanítására alapvetően két típusú eljárás létezik: a beszélőfüggő tanítás és beszélőadaptációs eljárás.

Az első esetben szükség van egy beszélőtől rögzített, minél hosszabb hanganyagra. A rendelkezésre álló hanganyagból kinyerjük a hullámformára jellemző spektrális, gerjesztési és a hangidőtartam paramétereket, majd ezekből egy – a hanganyagra jellemző – statisztikus modellt építünk.

A második esetben több beszélőtől kell minél hosszabb hanganyagokat gyűjtenünk, továbbá szükségünk van egy adott célbeszélőtől (akinek a hangkarakterisztikáját próbáljuk majd visszaadni a beszédelőállítás során) származó rövidebb felvételre. Az összegyűjtött szövegkorpuszból az első esethez hasonlóan kinyerjük a hullámformára jellemző spektrális, gerjesztési és fonéma hangidőtartam paramétereket, majd a több beszélőtől gyűjtött hosszabb felvételekből kinyert paraméterek segítségével megépítjük az ún. átlaghangra (*average voice*) jellemző statisztikus modellt, melyet az adott célbeszélő rövidebb felvételéből kinyert paraméterek segítségével a célbeszélő hangkarakterére adaptálunk.

Mindkét esetben az előállt modelleket adatbázisban tároljuk, majd a beszédelőállítása során az adatbázisban tárolt modellekből kinyert paramétereket használjuk fel.

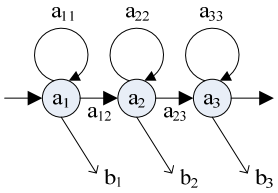
A beszédelőállításához beszédkódolási eljárást használunk, ahol a gerjesztési, szűrő és esetleges egyéb (pl. maradékjel) paramétereket HMM modellek generálják. A HMM-alapú magyar szövegfelolvasó beszélőfüggő tanításának lépéseiről korábban beszámoltunk [7], jelen cikkben röviden bemutatjuk a rejtett Markov-modellt, ismertetjük az átlaghang kialakítását, ennek adaptációs lehetőségeit és a betanított modellekből a beszédelőállításának folyamatát, továbbá bemutatjuk az általunk megvalósított szövegfelolvasó szubjektív méréséhez tervezett meghallgatásos teszt felépítését és eredményeit.

2 A rejtett Markov-modell

Gyakran használnak rejtett Markov-láncokat fizikai folyamatok modellezésére, ahol különböző megfigyelések alapján kell a folyamatot szimulálni. A beszédtechnológiában igen előnyösen lehet használni a rejtett Markov-modelleket, ekkor a beszédre jellemző, abból kinyert paramétereket kell tárolni, mely jelentősen hatékonyabb, mint a hangminta alapú rendszerek esetén a minták tárolása, hiszen a paraméterek jóval kevesebb helyet foglalnak el és jobban lehet belőlük általánosítani, mint az eredeti hullámformák esetén. A paraméterek (például spektrális jellemzők) kinyeréséhez úgynevezett akusztikus modelleket alkalmaznak. Régebben hangonkénti (ún.

monofón) akusztikus modellt használtak, manapság már a hangkörnyezetet is figyelembe vevő akusztikus modellek (pl. hanghármasok, ún. trifónok) a leggyakoribbak (Mihajlik et al. 2006).

Napjainkban a beszédtechnológia területén a rejtett Markov-modellek a beszédfelismerés alapjait képzik, szinte minden komoly rendszer erre a technológiára épül. A modell működését egy egyszerű példán keresztül mutatjuk be. A szavakat úgy tekintjük, hogy azok beszédhangok sorozataként állnak elő. Minden beszédhangra három állapotot feltételezünk: a hang eleje, közepe, vége. Az egyes állapotok között, és az egyes állapotokból saját magukra mutató, úgynevezett élek határozzák meg, hogy az adott állapotból mely következő állapotokba lehet lépni (1. ábra). Az ábrán az a_1 jelöli a beszédhang elejét, az a_2 a közepét és az a_3 pedig a végét. Az a_{12}, a_{23} élek a belső állapotok közötti átmeneti valószínűségeket jelentik, az a_{11}, a_{22}, a_{33} pedig azt jelzi, hogy milyen valószínűséggel maradunk az adott belső állapotban. A modell betanítása során az élekhez valószínűségek rendelhetők, melyek a helyben maradás (a_{11}, a_{22}, a_{33}), illetve továbblépés (a_{12}, a_{23}) valószínűségét határozzák meg. A b_1, b_2, b_3 jelöli a megfigyelési valószínűségeket.



1. ábra. Három állapotú rejtett Markov-modell

Az egyes állapotok tartalmazzák az akusztikus modellek készítése során becsült sokdimenziós Gauss-eloszlások paramétereit. Általában egy adott környezetben lévő beszédhang többször előfordul a tanító adatbázisban, a tanítás során pedig az ehhez tartozó spektrális paraméterhalmazt próbáljuk becsülni Gauss-eloszlással. A mintaillesztő eljárás ezen akusztikus modellekhez illeszti a bejövő paramétereket, hogy eldöntse, megegyezik-e az a felismerendő szóval. A rejtett Markov-modelleket [8] mutatja be részletesen.

A rejtett Markov-modell alkalmazása a beszédszintézis területén az elmúlt évtizedben merült fel és napjainkra egyre nagyobb figyelmet kapott. Az erre kidolgozott eljárás három lényegi ponton tér el a beszédfelismerésre kidolgozott megoldástól. A legjelentősebb különbség az, hogy a két eljárás esetében a bemeneti és a kimeneti paraméterek felcserélődnek, tehát a végső lépésnél a mintaillesztés helyett mintaválogatást hajtunk végre, majd a kiválasztott jellemző paraméterhalmazból a modell egy beszédkódoló eljárással beszédhangot állít elő, és így jön létre a szintetizált beszédhullám. A második fontos különbség, hogy a prozódia jellemző komponenseit (például hangmagasság, hangidőtartam) is modellezni kell a beszédszintézis esetében, mely feladatokat szintén végezhetnek rejtett Markov-modellek. A harmadik fontos különbség pedig az, hogy trifón akusztikus modellek helyett sokkal összetettebb akusztikus modellt használunk, melyben az adott hanghoz közeli és távoli hangok szegmentális és szupraszegmentális szinten is beépülnek.

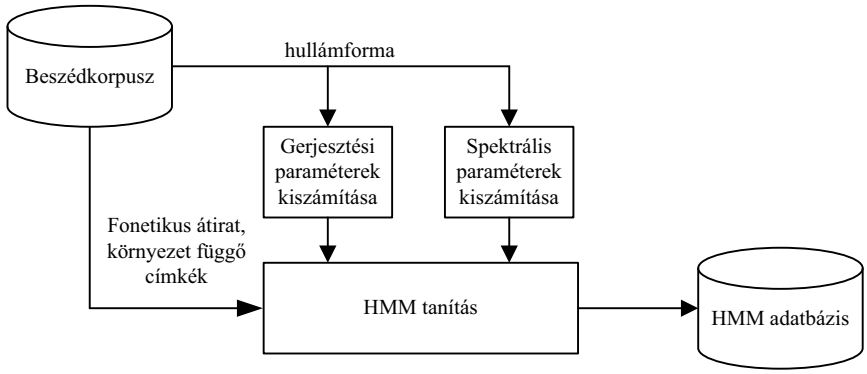
3 Rejtett Markov-modell alapú beszédszintézis

A HMM-alapú szövegfelolvasó két fő részből áll: a tanulási és a szintetizálási fázisból. A tanulás során a rejtett Markov-modelleket egy nagy, gondosan megtervezett és felcímkézett beszédadatbázis (és annak fonetikus átirata) segítségével tanítjuk be. A tanítási folyamat végére egy kisméretű HMM adatbázis áll elő, melyben a betanított beszédkorpuszra jellemző HMM paraméterek találhatóak. Ezekből válogatja majd ki a szintetizátor a beszéd előállítása során a szintetikus beszéd generálásához szükséges paramétereket. Ezen adatokból alakítja valamilyen beszédkódolási eljárással a paramétereket beszéddé.

A szintetizálási fázisban már csak a tanítás eredményét, egy néhány megabájtos adatbázist használunk. A bemeneti szöveg alapján meghatározzuk, hogy milyen hangsorozatot kell generálni és a HMM-adatbázisban tárolt paraméterekből kiválogatjuk azt a paramétersorozatot, amelyik legjobban reprezentálja az előállítani kívánt hangsorozatot. Ezekből állítjuk vissza a spektrális jellemzőket, a hangidőtartamokat, a szüneteket és az alapfrekvenciát, majd ezek alapján beszédkódoló eljárással elkészítjük a szintetizált beszéd hullámformáját.

A HMM modellek tanítására alapvetően kétfajta lehetőségünk van: beszélőfüggő modell tanítása vagy beszélőfüggetlen modell tanítása, majd az így előálló átlaghang adaptációja egy adott célszemély beszédhangjára.

Beszélőfüggő esetben a tanításhoz egy beszélő minél hosszabb hangfelvételére (legalább 1-1.5 óra), ennek fonetikus átiratára és pontos hanghatárjelölésekre van szükség. Fontos, hogy a hangfelvétel szövege fonetikusan kiegyenlített legyen. Hogy minél jobb minőségű hangot tudjunk előállítani, ügyelni kell arra, hogy a felvételek stúdió körülmények között legyenek rögzítve, továbbá hogy a fonetikus átirat és a címkézés precíz legyen. A hanghatárokat a gyakorlatban automatikus, úgynevezett kényszerített beszédfelismerési (forced alignment) módszerrel jelöljük meg. Ebből adódik bizonyos mértékű hiba. A beszélőfüggő tanítás lépéseit a 2. ábra mutatja be.



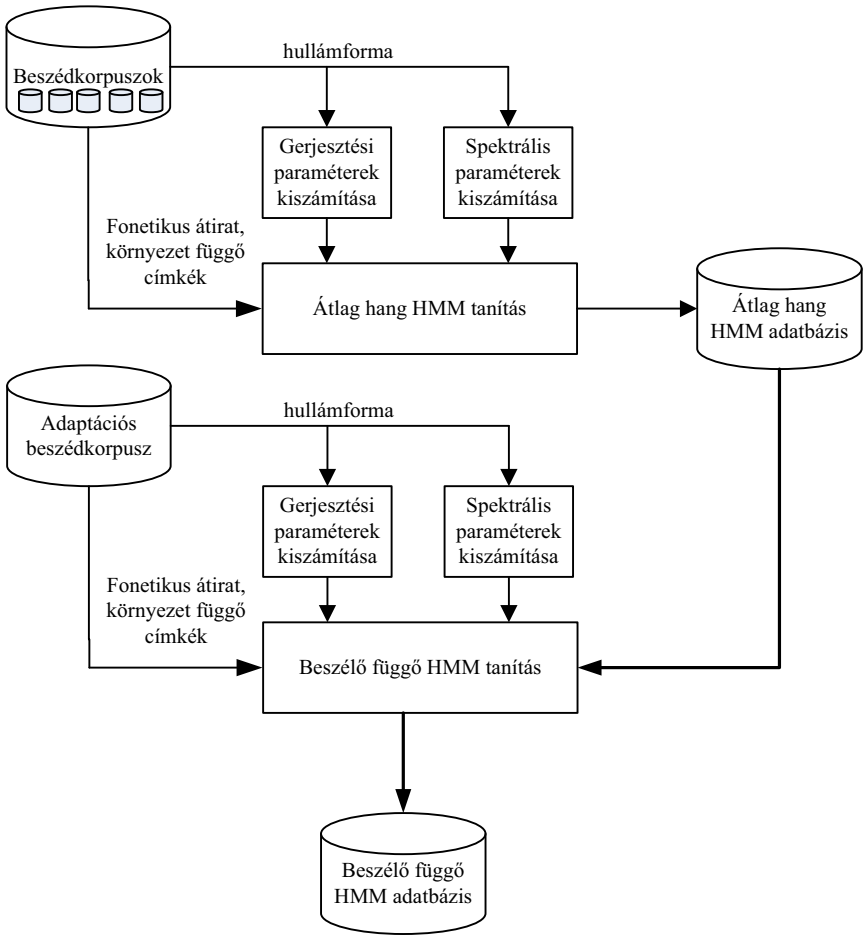
2. ábra. Beszélőfüggő HMM adatbázis tanítása.

A tanításhoz ezen túl szükségünk van az adott nyelvre jellemző környezetfüggő címkézésre és a döntési fák építéséhez egy nyelvspecifikus kérdésfájlra [9]. Ezek segítségével megkezdődhet a tanítás, mely a hosszú, több száz megabyte-ot elfoglaló

hanganyagból az adott beszélőre jellemző beszédhang paraméterek generálására alkalmas HMM adatbázist eredményez.

A HMM-alapú magyar beszédelőállításról korábban részletesen beszámoltunk [7], a továbbiakban a beszélőfüggetlen tanítás adaptációját ismertetjük.

Beszélőfüggetlen esetben először egy átlaghangot tanítunk, melyet utána egy célbeszélő hangkarakteréhez igazítunk. Ebben az esetben így áll elő a HMM adatbázis. Ezután a beszédhang előállításának módszere megegyezik a beszélőfüggő esetben használt módszerrel. A beszélőfüggetlen tanítás, majd adaptálás működési elvét a 3. ábra mutatja be.



3. ábra. Beszélőfüggetlen HMM adatbázisból kiinduló adaptált tanítás.

3.1 Beszélőfüggetlen átlaghang tanítása

A beszélőfüggetlen esetben először elő kell állítani egy ún. átlaghangot. Ennek előállításához több beszélőtől (legalább 4-5), minél hosszabb (személyenként legalább 1-1.5 óra) hangfelvételre, annak fonetikus átiratára és pontos hanghatárjelöléseire van szükség. A minél jobb minőség érdekében itt is érdemes figyelni arra, hogy a felvételek stúdió körülmények között legyenek rögzítve, illetve hogy a fonetikus átirat és a címkézés precíz legyen. Ezután automatikus módszerrel előállítjuk a beszédkorpuszhoz tartozó fonetikus átirat környezet függő címkéit, majd a HMM-eket az összes beszélő adatbázisa alapján tanítjuk be az átlaghangra, melyben jelen vannak minden egyes beszélőre az alapfrekvencia, hangidőtartam és spektrális paraméterek.

Érdekes kérdés, hogy az átlaghang tanításához férfi, női, vagy kevert hangokat használjunk. Amennyiben nagy mennyiségű férfi és női hanganyag áll rendelkezésre, a leghatékonyabb megoldást a nemfüggő átlaghang használata jelenti. A gyakorlatban azonban általában az egyik, vagy mindkét nemtől csak korlátozott mennyiségű hanganyagunk van, ezért a kevert nemű átlaghang előállítást célszerű választanunk, majd ebből adaptálni mind férfi, mind női hangra. Meg lehet csinálni, hogy ellentétes nemű átlaghangból adaptálunk női / férfihangra, azonban [10] beszámol arról, hogy ez jelentős minőség- és természetességcsökkenést okoz a végső hangnál a nemfüggő átlaghanghoz képest. [11] egy olyan eljárásról számol be, mely segítségével kevert nemű átlaghangból a nemfüggő átlaghanghoz képest minimális minőség- és természetességromlás mellett lehet női és férfihangra adaptálni.

3.2 Beszélőadaptáció

Miután elkészültek az átlaghang HMM modelljei, a célbeszélőtől származó hangfelvételekkel tudjuk a modellt az adott személy hangkarakteréhez és beszédstílusához igazítani, adaptálni. A beszélőadaptációjára alapvetően kétfajta lehetőségünk van.

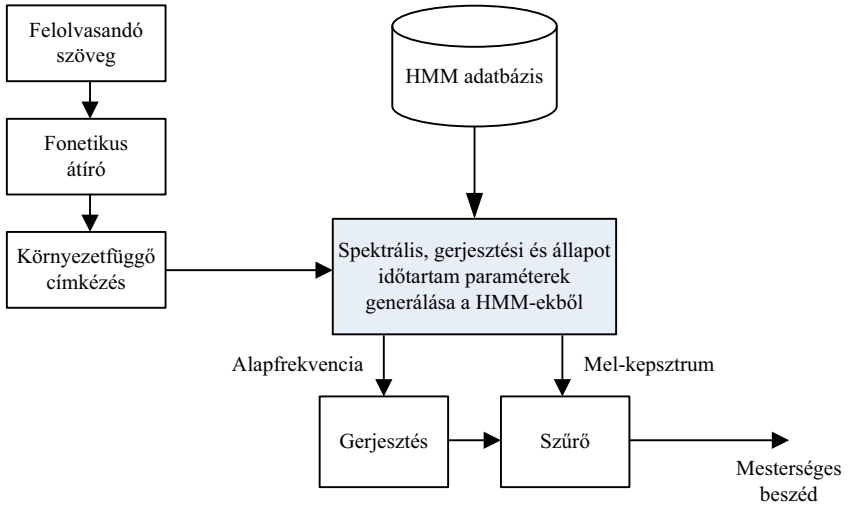
Amennyiben kevés hanganyag áll rendelkezésre a célbeszélőtől, akkor előnyös Maximum Likelihood Linear Regression (MLLR) alapú adaptációt választani [5]. [12] irodalomban ismertetett kísérlet alapján akár már öt mondat is elegendő lehet ahhoz, hogy a célszemély hangkarakterét és beszédstílusát többé-kevésbé visszaadja a mesterségesen előállított hang.

Amennyiben hosszabb adaptációs hanganyag is elérhető, akkor a Maximum A Posteriori (MAP) technikát érdemes használni [6], mely az előzőnél jobb minőségű mesterségesen generált hangot eredményez. Ennek a technológiának az új változatai, mint például a CSMAPLR (Constrained Structural Maximum A Posteriori Linear Regression) közel azonos minőséget és természetességet képviselnek, mint a beszélőfüggő tanítás esetén előállított mesterséges beszéd [13].

Természetesen mindegyik adaptációs technológia esetén szükséges az adaptációs hanganyag fonetikus átirata és a pontos hanghatárjelölések.

3.3 Beszéd előállítása

A beszéd előállítása megegyezik a beszélőfüggő esetben használt eljárással. A beszéd előállítása során a HMM által generált alapfrekvencia, hangidőtartam és spektrális paramétereket használjuk fel. A HMM-ek tanításától függően a beszéd előállítását végezheti egészen egyszerű beszédkódoló is (pl. LPC-10). A jobb minőség érdekében használhatunk ennél bonyolultabb technológiákat, mint például a MELP (Mixed Excitation Linear Prediction) kódoló. Természetesen ebben az esetben a beszédkorpuszból számolt maradék jeleket is be kell tanítanunk a HMM-ekkel.



4. ábra. A beszédhang előállítása a HMM adatbázisból.

4 Magyar nyelvű tanítás és adaptáció

A magyar nyelvű HMM-alapú beszélőadaptált szövegfelolvasó elkészítésének bizonyos lépései hasonlóak a beszélőfüggő változathoz. A döntési fák építéséhez és a környezet függő címkézéshez a korábban bemutatott eljárást használtuk [7]. Jelen cikkünkben az adaptációhoz használt adatbázisokat és az alkalmazott adaptációs technológiát ismertetjük.

4.1 A felhasznált beszédkorpuszok

Az átlaghang építéséhez négy férfi és egy női beszélőtől rögzített adatbázist használtunk. Az adatbázisokat stúdió körülmények között vettük fel, az adatbázisok szövege gondosan megtervezett, fonetikusan kiegyenlített mondatokat tartalmaz. Az átlaghang készítéséhez felhasznált adatbázisok további jellemzőit az 1. táblázat mutatja.

1. táblázat: Az átlaghang létrehozásához használt beszédkorpuszok (formátum: 44 kHz, 16 bit, mono).

Beszélő	Mondatszám	Időtartam	Méret
1. férfi beszélő	1941	170 perc	857 MB
2. férfi beszélő	1938	137 perc	694 MB
3. férfi beszélő	1944	191 perc	966 MB
4. férfi beszélő	1938	214 perc	1082 MB
1. női beszélő	1940	129 perc	652 MB

Miután készen lett az átlaghang HMM adatbázis, négy különböző beszélőtől rögzített, félig spontán hanganyagot használtunk fel közepesen zajos környezetből az adaptációhoz, melyek tulajdonságait a 2. táblázat mutatja. Mind a négy esetben publikusan elérhető parlamenti felvételeket használtunk, melyek előre megtervezettek, de spontán módon előadottak.

2. táblázat: Az adaptációhoz használt beszédkorpuszok (formátum: 44 kHz, 16 bit, mono).

Beszélő	Mondatszám	Időtartam	Méret
1. férfi beszélő	87	19 perc	94 MB
2. férfi beszélő	48	17 perc	89 MB
3. férfi beszélő	30	11 perc	58 MB
4. férfi beszélő	26	9 perc	44 MB

4.2 Az alkalmazott adaptációs technológia

A beszélőadaptáció során MLLR eljárást használtunk. Az MLLR lineáris transzformációk segítségével az átlaghang HMM modell paramétereit a cél hang 'irányába' módosítja. Az állapotkimenetek ekkor a következőképp alakulnak:

$$b_j(o_t) = N(o_t; \hat{\mu}_j; \hat{\Sigma}_j) \quad (\mathbf{1})$$

$$\hat{\mu}_j = A_{r(j)}\mu_j + b_{r(j)} \quad (\mathbf{2})$$

$$\hat{\Sigma}_j = H_{r(j)}^T \Sigma_j H_{r(j)} \quad (\mathbf{3})$$

ahol $\hat{\mu}_j$ és $\hat{\Sigma}_j$ a j-edik állapotra jellemző kimeneti sűrűségfüggvényhez tartozó várható érték vektor ill. kovariancia mátrix a lineáris transzformáció után. $A_{r(j)}$, $b_{r(j)}$ és $H_{r(j)}$ a várható érték lineáris-transzformációs mátrixa, a hozzá tartozó eltolás vektor és a kovariancia lineáris-transzformációs mátrixa az r(j)-edik regressziós osztályban. Az adott állapotokra jellemző kimeneti sűrűségfüggvényeket regressziós-fa

segítségével osztályokba soroljuk, egy adott osztályban azonos lineáris-transzformációs mátrixokat és az eltolás vektort használunk. A regressziós fa méretének az adaptációs anyag mennyiségéhez való igazításával tudjuk szabályozni az adaptáció komplexitását és általánosítható képességét. Alapvetően az MLLR két fajtáját különböztetjük meg: azonos A és H lineáris-transzformációs mátrixok esetén erőltetett MLLR-ről (Constrained MLLR, CMMLR), egyébként pedig szabad MLLR-ről (Unconstrained MLLR) beszélünk. A jelen cikkben ismertetett rendszer esetén CMLLR-t használtunk.

5 Eredmények

A rejtett Markov-modell alapú szövegfelolvasó beszélőadaptációjának megvalósításához a HTS rendszer [9] módosított, magyar nyelvű változatát vettük alapul [7]. A tanításhoz és adaptációhoz a 4.1 szakaszban ismertetett beszédkorpuszokat használtuk fel. Az összeállított rendszer minőségének szubjektív mérése céljából egy meghallgatásos tesztet állítottunk össze.

5.1 A teszt felépítése

A tesztben a korábban ismertetett adaptációs anyagok alapján négy különböző férfihangra adaptált rendszer vett részt. A teszt két részből áll. A teszt első felében a tesztalanyoknak 16 mintát (rendszerenként négyet) kellett 1-től 5-ig tartó skálán értékelniük természetesség szempontjából. Az 1 azt jelentette, hogy a hangminta zavaróan gépies hangzású, az 5 pedig azt, hogy teljesen természetes. A teszt második felében a beszélők eredeti hangkarakteréhez viszonyítva kellett a tesztalanyoknak szintén 1-től 5-ig tartó skálán megmondaniuk, hogy mennyire adja vissza a szintetizált hang az eredeti beszélő hangkarakterét. Az 1-es itt azt jelentette, hogy egyáltalán nem adja vissza, az 5-ös pedig hogy a szintetizált hangminta összetéveszthető az eredeti beszélővel. A teszt második felében minden rendszerből 5 mintát, így összesen 20 mintát hallgattak meg.

Mindkét részben a minták pszeudovéletlenszerűen lettek kiválogatva egy 40 darabos halmazból, ügyelve arra, hogy a minták előfordulási gyakorisága egyenletes eloszlást kövessen. A különböző rendszerekből kiválogatott mintákat ezután véletlen sorrendbe rendeztük. Ezen lépésekre azért volt szükség, hogy elkerüljük a memóriahatást a teszt során, tehát hogy a tesztalanyok által adott értékeket nem csak a minták tartalma, hanem a minták sorrendje is befolyásolja (pl. egy rosszabb minta után következő jobb minta sok esetben jobb pontszámot kaphat, mintha előtte is egy hasonló minőségű minta állna).

A tesztet összesen 25-en végezték el, 19 férfi és 6 nő. Internet alapú volt a teszt, az átlag életkor 35 év volt, a legfiatalabb tesztalany 21, a legidősebb 67 éves volt. 10 tesztalany beszédszakértő volt.

5.2 Az eredmények értékelése

A teszt eredményeit a 3. táblázat mutatja.

3. táblázat: A meghallgatásos teszt eredményei. Mindkét oszlopban az első érték az átlag, a második az átlagos szórás, a harmadik, zárójelben lévő érték pedig a konfidenciát jelöli =0.05 mellett.

Adaptációs korpusz	A hangminta természetessége	Hasonlóság az eredeti beszélő hangjához
1. férfi beszélő	3.2 ± 1.09 (0.2)	2.9 ± 1.08 (0.2)
2. férfi beszélő	3.1 ± 1.08 (0.2)	2.9 ± 1.05 (0.2)
3. férfi beszélő	3 ± 1.17 (0.2)	2.7 ± 1.05 (0.2)
4. férfi beszélő	3 ± 1.11 (0.2)	2.6 ± 1.06 (0.2)

Az értékekből kitűnik, hogy a hangminta természetessége a különböző beszélők esetén közel azonos, a hosszabb adaptációs anyag nem okozott szignifikáns különbséget a rövidebbhez képest. Ez azzal magyarázható, hogy mindegyik hang az átlaghangból lett adaptálva, mely már önmagában is elég információt hordoz természetes hang létrehozásához.

A teszt második felében, a hangminta összehasonlítás során azonban már meg lehet figyelni, hogy rövidebb adaptációs anyag (ld. 2. táblázat) esetén az eredeti beszélőhöz való hasonlóság csökken.

6 Jövőbeli tervek

A jövőben kísérleteket fogunk végezni azzal kapcsolatban, hogy a félig spontán, közepes minőségű adaptációs anyagokat stúdió minőségű, tervezett beszédűre cserélve hogyan változik a generált hang minősége és természetessége. A hanghatárok automatikus jelölését ellenőrizni fogjuk félautomatikus és kézi módszerekkel. Ezen túl más típusú adaptációs technológiákat is kipróbálunk (például MAP vagy CSMAPLR). Méréseket végzünk ezek minőségével kapcsolatban.

Kiemelt fontosságúnak tartjuk a beszédelőállításának folyamatát mobil platformokra optimalizálni. A fentebb ismertetett megoldás futás idejű tárhely igénye (1-2 MB) elméletileg lehetővé teszi kevés erőforrással rendelkező eszközökre való átvitelét, azonban számítás-igénye jelentős optimalizációra szorul.

7 Összefoglaló

Cikkünkben röviden áttekintettük a rejtett Markov-modell alapjait, kapcsolatát a beszédtechnológiával, és különösen a beszédszintézissel. Röviden összefoglaltuk a magyar nyelvű, beszélőfüggő HMM-alapú mesterséges beszédelőállítás elemeit, majd részletesen ismertettük a beszélőadaptációhoz szükséges lépéseket. Ezután ismertet-

tük a megvalósított rendszer szubjektív méréséhez tervezett meghallgatásos teszt felépítését és annak eredményeit. Végezetül jövőbeli terveinkre tértünk ki.

A beszédszintézis területén jelenleg az egyik leggyorsabban fejlődő terület a rejtett Markov-modell alapú beszédelőállítás. Szeretnénk a világgal lépést tartva magyar nyelven is megvalósítani a legújabb technológiákat, illetve új eredményekkel hozzájárulni a terület gyorsabb fejlődéséhez.

Hivatkozások

1. Black, A., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In Proc. ICASSP (2007), 1229-1232
2. Kim, S.-J., Kim, J.-J., Hahn, M.-S: HMM-based Korean speech synthesis system for hand-held devices. IEEE Trans. Consumer Electronics 52 (4) (2006) 1384–1390
3. N. Iwahashi, Y. Sagisaka: Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks", Speech Communications, Vol. 16, no. 2 (1995) 139–151
4. Tachibana, M., J. Yamagishi, Masuko, T., Kobayashi, T.: Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing. IEICE Trans. Inf. Syst., Vol. E88-D, no.11 (2005) 2484-2491
5. Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Adaptation of Pitch and Spectrum for HMM-Based Speech Synthesis Using MLLR. In Proc. ICASSP 2001, (1998) 805-808
6. Ogata, K., Tachibana, M., Yamagishi, J., Kobayashi, T.: Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis. In Proc. ICSLP 2006, (2006) 1328–1331
7. Tóth, B., Németh, G.: Hidden Markov model based speech synthesis system in Hungarian, Infocomm., Vol. 63, no. 7 (2008) 30–34
8. Rabiner, Lawrence R: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. (1989) 257–286
9. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K.: The HMM-based speech synthesis system version 2.0, in Proc. ISCA SSW6. (2007) 294–299
10. Isogai, J., Yamagishi, J., Kobayashi T.: Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis. In Proc. EUROSPEECH 2005 (2005), 2597–2600
11. Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T., Tokuda, K.: Improved Average-Voice-based Speech Synthesis using Gender-Mixed Modeling and A Parameter Generation Algorithm considering GV, Proc. ISCA SSW6, Aug. (2007)
12. Tamura, M., Masuko, T., Tokuda, K., Kobayashi T.: Speaker adaptation for HMM-based speech synthesis system using MLLR, Proc. ESCA/COCOSDA Workshop on Speech Synthesis (1998) 273-276
13. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J. Analysis of Speaker Adaptation Algorihms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm.' IEEE Audio, Speech, & Language Processing Vol.17 issue 1 (2009) 66-83 2009

V. Pszichológiai vonatkozású fejlesztések

Történelmi szövegek narratív pszichológiai vizsgálata a nemzeti identitás tükrében

Szalai Katalin[1], Ferenczhalmy Réka[1], Fülöp Éva[1], Vincze Orsolya PhD[1], Dr. László János[2]

[1] Pécsi Tudományegyetem Pszichológiai Intézet Doktori Iskola
7624 Pécs, Ifjúság útja 6.
[2] MTA Pszichológiai Kutatóintézet
1132 Budapest XIII. Victor Hugo u. 18-22.

Kivonat: A narratív pszichológia szerint a történelemről szóló tudásunkat, az azt megjelenítő szövegeket tekinthetjük az elbeszélések szabályait követő konstruktumoknak [24], melyekben a történészek egyes eseményeket kiemelő, másokat háttérbe szorító, az események között koherenciát teremtő munkája érvényesül. A történelmi elbeszélések, mint például az iskolai tankönyvek, nemcsak tényeket rögzítenek és továbbítanak, hanem mintát nyújtanak a nemzeti identitás jellemzőire, az elfogadott viselkedésformákra, az adott társadalomban megjelenő lelkiállapotokra, az eseményekhez való viszonyulásokra. Ezáltal a történelem mint reprezentációs forma, a nemzeti identitás kialakításában, megerősítésében is szerepet játszik. Korábban bemutatásra kerültek a narratív pszichológiai tartalomelemzéshez kifejlesztett szótár alapú lokális nyelvtanok, melyeket a NooJ integrált nyelvelemző környezetben [18] fejlesztettünk, az érzelmek, a kognitív állapotok, az intenció, az aktivitás-passzivitás stb. témakörében. Tesztelésüket különböző szociálpszichológiai jelenségek körében végeztük, illetve jelenleg ezen eszközök segítségével a nemzeti identitás jellemzőit keressük.

1 Bevezetés

A narratív pszichológia elgondolása szerint az egyéni élettörténet a személyes identitás hordozója [9] [14]; saját történetünket mondva alkotjuk meg újra és újra önmagunkat [17]. Identitásunk szempontjából viszont nemcsak az egyénileg megélt életeseményeink a fontosak, hanem a csoportunk történelmi útja is az. Önmeghatározásunknak azon részét, mely egy csoporthoz köt minket, szociális identitásnak nevezzük [21].

Ha a narratív modell keretein belül a személyes identitást megközelíthetjük az élettörténetek segítségével [3] [7] [12] [14], úgy a nemzeti identitás jegyeit a nemzeti múlt segítségével fedhetjük fel. Kutatásaink során a nemzethez mint csoporthoz kötődő identitás konstrukciójának nyelvi markerekkel kódolt mintázatait próbáljuk feltárni. Különös figyelemmel a traumatikus élményekkel vagy veszteségekkel való meg-

küzdésre, a nemzetre jellemző érzelmi állapotok feltárására, a szándékteliség, felelősségvállalás és az ágencia kérdéseire.

1.1 Történelem, identitás, narratívum

A narratív pszichológia feltevése szerint társas világunk elbeszéléseken keresztül szerveződik. Történeti tudásunk is értelmezhető a narrativitás szabályai szerint, azaz a történelem maga tekinthető szociális konstrukció eredményének, a jelentősnek ítélt múltbeli eseményekről szőtt elbeszélések láncolatának. A történész által alkotott történelmi narratívumban azon események kapnak helyet, melyek illeszkednek egy koherens identitáskonstrukcióhoz, és emlékezetre méltónak ítéltetnek.

Az emlékezetre méltó történetek a csoport kollektív emlékezetének részét képezik. A kulturális emlékezetben a régmúlt történései tárolódnak, míg a kommunikatív emlékezetben a közelmúlté [2]. A nemzet közös múltjának történetei nemzedékről nemzedékre adódnak át, biztosítva a csoport létének folyamatosságát. Azonosulási mintát kínálnak a csoporttagoknak, az egyes események társadalmilag megélhető érzelmeit, megoldási módjait nyújtják. A csoportról – jelen esetben nemzetről – szóló történetek így jelentős szerepet játszanak a csoport identitásának kialakításában, közvetítésében.

A kollektív emlékezetben őrzött történelmi reprezentációk megjelennek szépirodalmi művekben, tankönyvekben, írott sajtóban, a XX. században pedig különösen nagy jelentőségű hordozói az ún. nyilvános történelemnek a média termékei [8].

A narratív felfogás lehetővé teszi számunkra, hogy a csoport jelenlegi identitásállapotainak múltba ágyazottságát, illetve a korábbi identitásállapotok szövegek által rögzített mintázatait vizsgáljuk.

1.2 A nemzeti identitás egyéni mintázatai

A kollektív emlékezetben őrzött és nemzedékről nemzedékre áthagyományozott csoporttörténetek a nemcsak az eseményeket, de az ahhoz kötődő érzelmeket, az adott helyzetben elfogadott viselkedésmintákat, probléma-megoldási módokat is közvetítik a csoporttagok számára. Az, hogy egy csoport mennyire hatékony az őt ért külső hatásokkal való megküzdésben, a történtek feldolgozásában, feltételezésünk szerint tetten érhető a szöveg szintjén megjelenő ágenciában, a mentális állapotok mintázataiban, változásaiban.

Az ágencia fontos összetevő a nyugati kultúrák személyes és csoportidentitásának konstrukciójában. A felnőtt, érett identitás egyik szükséges eleme a megfelelő autonómia megszerzése, elérése [15]. Az autonómia jelenségén kívül az ágenciának széleskörű pszichológiai megjelenési formái lehetnek: teherbírás, fejlődés, hatalom, dominancia, kontroll, szeparáció és függetlenség. Továbbá összefüggésben áll a megküzdés jelenségével [13], a személyes hatékonysággal [4] [5], a célvezérelt cselekvéssel, illetve a célelérés hatékonyságával.

A csoportágenciáról készült kutatások az ágencia percepcióját a másokra való hatni tudás képességével, a célok teljesítésének képességével, a kollektív cselekvés képességével mérik [19]. Az ágencia jelensége alatt a csoport hatékony cselekvését értik [1] [6].

A csoportról szóló történetek tartalmazzák az eseményekhez, a saját és a másik csoportokhoz való érzelmi viszonyulást, a szereplők gondolatait, vélekedéseit. A másik tudattartalmára való következtetés a társas élet egyik szükséges feltétele. Mások mentális állapotainak figyelembevétele hozzájárul például ahhoz, hogy a megfigyelő saját viselkedését a cselekvő elvárásaihoz igazítsa, létrehozva ezáltal egy kielégítő interperszonális kapcsolat lehetőségét [16]. Ahogy a megfigyelő a másik, azaz a cselekvő mentális állapotait figyelemmel követi, felveszi a perspektíváját a hangsúly a megfigyelő önérdekéről a cselekvő érdekeinek figyelembevétele felé tolódik el.

A csoporttörténetek szempontjából a szereplők mentális állapotainak nyelvi kifejezései szerepet játszanak a csoportdinamika és a csoportközi viszonyok minőségének megjelenítésében. Míg a kognitív állapotok, a szereplők hiedelmei, vélekedései, vágyai elősegítik a szereplő nézőpontjának felvételét, ezáltal hozzájárulva az esemény megértéséhez, az érzelmi állapotok kifejezései az esemény érzelmi minőségének szabályozását implikálják. A csoportidentitás szempontjából releváns események elbeszélésében megfigyelhető mentális (kognitív és érzelmi) állapotok gyakoriságának csoportközi eloszlásából és időbeni mintázataiból az általuk közvetített identitásminőségekre, illetve többek között a történelmi sérelmeknek a nemzeti identitás szempontjából történő feldolgozására is következtethetünk. [23]

2 Az elemző eszköz

A tudományos pszichológia régóta használ tartalomelemző programokat a kvalitatív eredmények érdekében, ezek új generációja a morfoszintaktikai elemzésre is képes NooJ nyelvi fejlesztő környezet [18]. A NooJ mint eszköz fejlesztését az MTA Nyelvtudományi Intézetével közösen fejlesztjük. Maga az eszköz képes arra, hogy a különböző pszichológiai relevanciával rendelkező szavakat, kifejezéseket, szintaktikai mintázatokat megtalálja, címkézze, számszerűsítse, s ezáltal akár statisztikai elemzések tárgyává tegye. A program nemcsak puszta szólistákat képes megtalálni, ún. gráfok, lokális nyelvtanok építhetők, melyek segítségével nyelvtani szerkezeteket, szókapcsolatokat is felismer. A kutatócsoport különböző modulok mentén dolgozik a programmal, melyek többféle pszichológiai jelenség nyelvi jegyeire épülnek. Jelen vizsgálatban a mentális szótárt [22], az érzelemszótárt [11], az intencionalitás [10] és az aktivitás-passzivitás [20] szótárt alkalmaztuk.

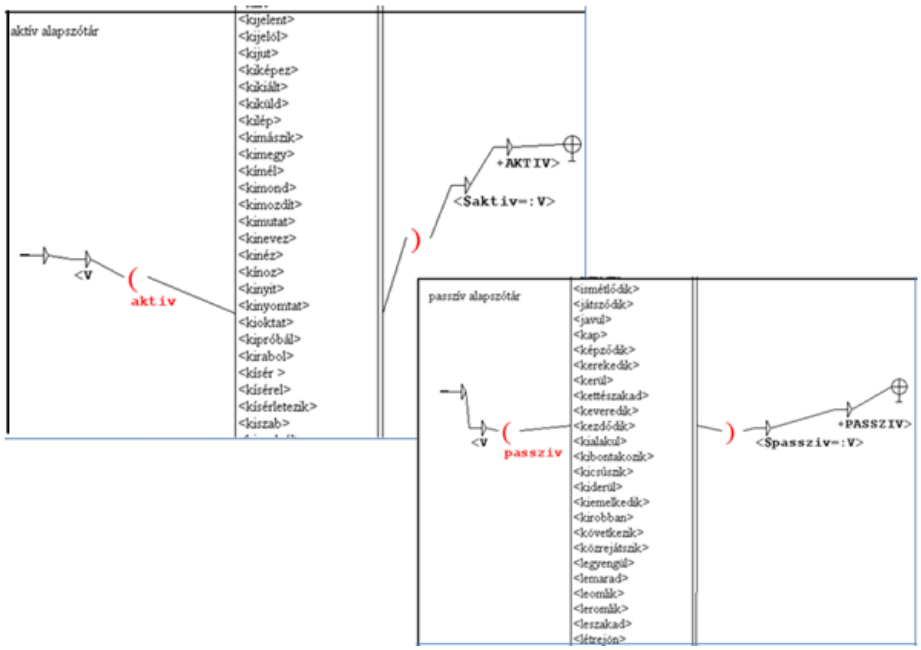
2.1 Az ágencia gráfjai

Az ágenciát jelen helyzetben az aktivitás-passzivitás, illetve az intencionalitás és kényszer nyelvi jegyeivel vizsgáltuk.

Az aktivitás illetve passzivitás kifejezéseit tartalmazó szótár összeállításához a Magyar Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által rendelkezésünkre bocsátott 10 ezer leggyakoribb igék gyűjteményét használtuk fel. Öt független bíráló segítségével osztályoztuk ezen igéket a két igekategória mentén. Aktívnak azon igéket tekintettük, amik ágense cselekvőképes, saját akaratából cselekszik, annak is tulajdonítva a történéseket - azaz belső kontrollos; cselekedeteivel hatással van kör-

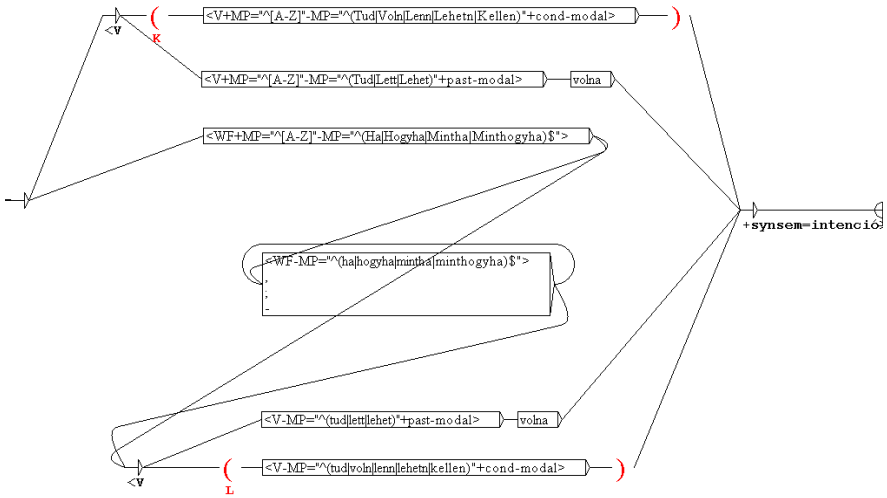
nyezetére (pl.: elér valamit, ad valakinek valamit, elfoglal valamit). Passzív igékhez tartoznak az állapotváltozás, történés igéi. Azon történések sorolhatók ide, amik a személyen kívül álló okokból - mint fizikai körülmények, transzcendens – következnek be, illetve változnak meg (pl.: valami kialakul, előfordul, valaki valamilyen helyzetbe kerül).

Az alábbi ábrán látható az aktív és a passzív alapszótár egy-egy részlete; jelenleg 941 aktív és 230 passzív igét tartalmaznak. Ezen kívül az egyes, gyakran előforduló igékre készült gráfok is a szótár részét képezik.



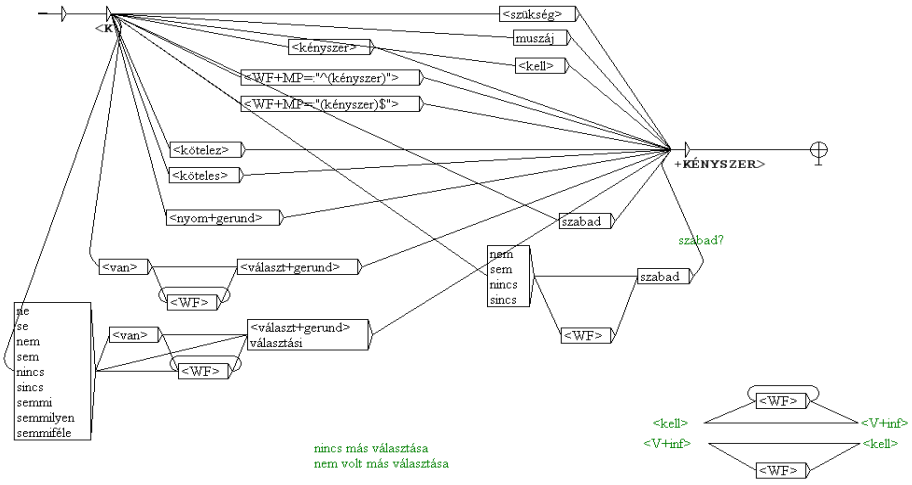
1. ábra Az aktív illetve a passzív alapszótár egy-egy részlete.

Az intenció a szöveg több szintjén jelenik meg. Önmagukban az aktív cselekvést kifejező igéket nem soroljuk ide, de ha például intencionális határozószó (pl. direkt, módszeresen, szándékosan, stb.) kapcsolódik hozzá, akkor jelöljük. Összeállítottuk az intencionális igék (pl. törekszik valamire, tervez, eldönt, akar, stb.), az intencionális főnevek (cél, terv, akarat, stb.), az intencionális melléknevek (pl. véletlen, szándékos, stb.), határozószók és névutó (végett) szótárait és lokális grammatikáit. Fontos a feltételes mód és a célhatározói alárendelő mondatszerkezet azonosítása is, melyek bizonyos esetei szintén intenciót jelenítenek meg a szövegben, ezek azonosítására és a téves találatok kiszűrésére is gráfokat hoztunk létre (2. ábra).



2. ábra A feltételes mód gráfja.

A kényszer gráfja (3. ábra) azokat az eseteket azonosítja a szövegben, amikor a cselekvés nem a cselekvő saját szándékából, hanem külső vagy belső nyomás hatására megy végbe, pl. kényszerül, muszáj, kell, nincs választása, stb.. Egyes eseteket az intencionalitás gráfok azonosítanak, melyek találatait ideszámoljuk.



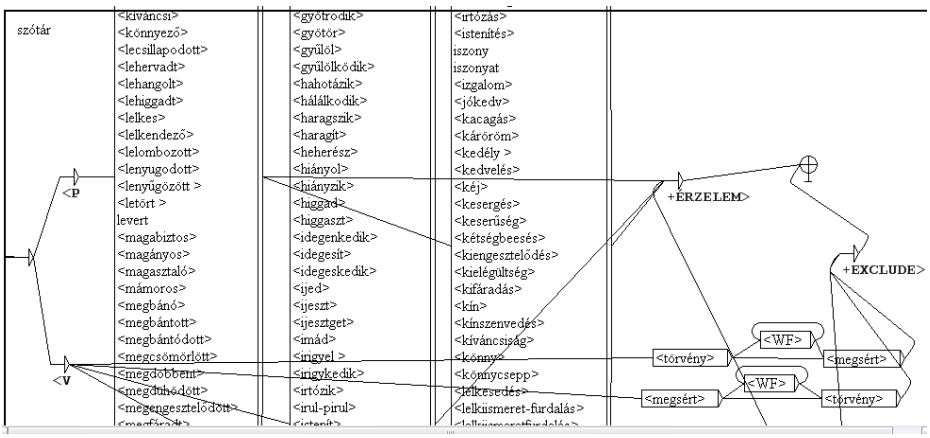
3. ábra A kényszer lokális grammatikája.

A fent látható módon kigyűjtjük az aktvitiás-passzivitás, illetve az intencionalitás – kényszer nyelvi jegyeit a szövegből, melyek szerintük az ágencia jelentését hordozzák. Arányszámot számolunk mind az aktív és passzív, mind az intenconális és kényszer nyelvi jegyeiből, majd ezeket átlagolva egy, az ágenciára jellemző arányszámot kapunk. Minél nagyobb ez az arányszám, azaz minél több aktív és intencionális kifejezést használ az elbeszélő a passzivitás és a kényszer rovására, annál nagyobb fokban jeleníti meg magát önálló akarattal bíró, céltudatos és hatékony cselekvőként a történetek menetében, annál jobban tanúskodik elbeszélése a környezetére befolyással bíró aktív ágensről. Illetve minél több passzív kifejezéssel él az elbeszélő, vagy minél több kényszernek van kitéve egy egyén vagy egy csoport, annál jobban hangsúlyozza a saját passzivitását, cselekvésképtelenségét egy adott helyzetben, s az arányszám annál kisebb.

2.2 A mentális állapotok gráfjai

A mentális állapotok vizsgálatát két modul mentén végezhetjük: a kognitív állapotok illetve az érzelmi kifejezések szótárának segítségével.

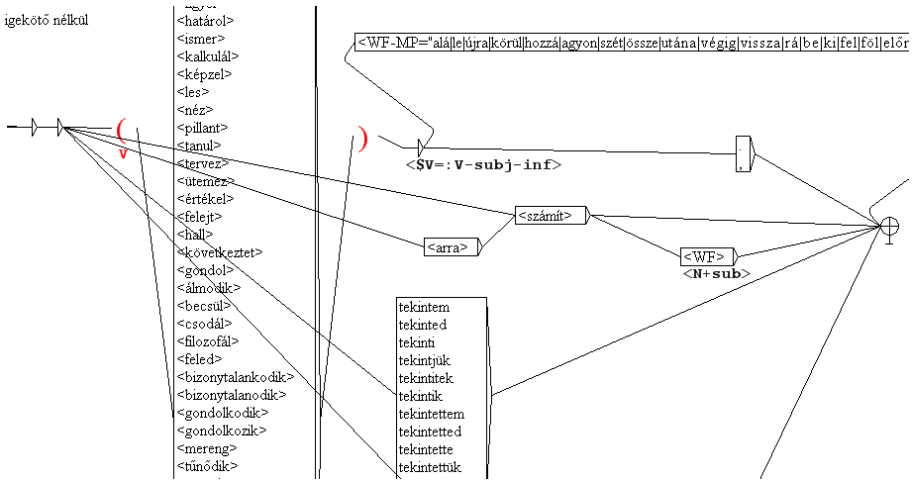
Az érzelemszavak összegyűjtésére egy korábbi munkafázisban került sor [11], melyben a magyar nyelv összes expresszív módon megjelenő érzelmi kifejezésének válogatása történt meg két független kódoló közreműködésével. Az így létrejött érzelemszótár került be azután a NooJ nyelvtechnológiai eszközbe, melyben a szavak pontos nyelvtani környezetének meghatározásával alkalmunk nyílik a magyar nyelv sajátosságait szem előtt tartó nyelvtani szerkezet keresésre. (Pl. szórendtől függetlenül megtalálhatjuk a 'valaki megsért valakit'' idiómát, és kizárhatjuk az olyan téves találatokat, mint 'valaki megsérti a törvényt'). Továbbá csoportosíthatjuk a találatainkat: jelenleg kezeljük a pozitív-negatív, humán-nem humán és az egyes történelmi pályákhoz kapcsolódó érzelmeket, melyek mind a csoportközi érzelmi helyzetek és a nemzeti identitás körvonalazásában segítenek.



4. ábra Az érzelemszótár egy részlete.

A kognitív kifejezések szótára szintén korábbi munkafázisban készült el, szintén a Magyar Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által rendelkezésünkre bocsátott 10000 leggyakoribb igei, 40000 főnévi és 15000 melléknévi lista alapján. Azon kifejezések kerültek a szótárba, amelyek episztemológiai vagy perceptuális cselekvést jelölnek. A kódolást hét független bíráló ellenőrizte.

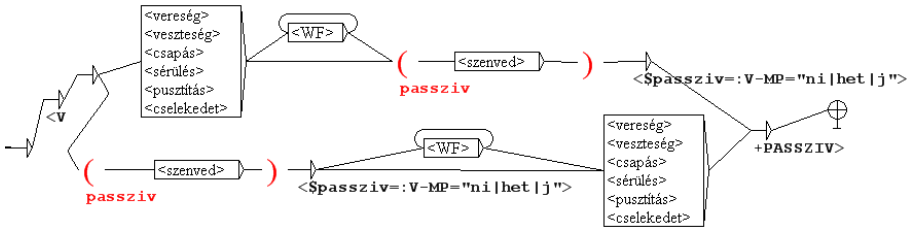
A NooJ nyelvi elemző környezetében való egyszerűbb kezelése végett a kognitív kifejezések csoportosítása morpho-szintaktikai szempontok alapján történt. Az egyik csoportot azok a *kognitív igék* alkották, amelyek szószintű kognitív jelentéssel bírnak (összesen 308 ige, pl.: általánosít, ámuldozik, analizál, asszociál). Míg a másik csoportba soroltuk azokat a kifejezéseket, amelyek csak bizonyos szókapcsolatban, vagy egy nyelvtani szerkezetben jelenítenek meg kognitív cselekvéseket. Ezeket az igéket *feltételes kognitív igéknek* neveztük el (összesen 302 ige; pl. 'általtja magát', 'felfrissíti az emlékezetét', 'átveszi valakinek a gondolatát', 'belát valamit').



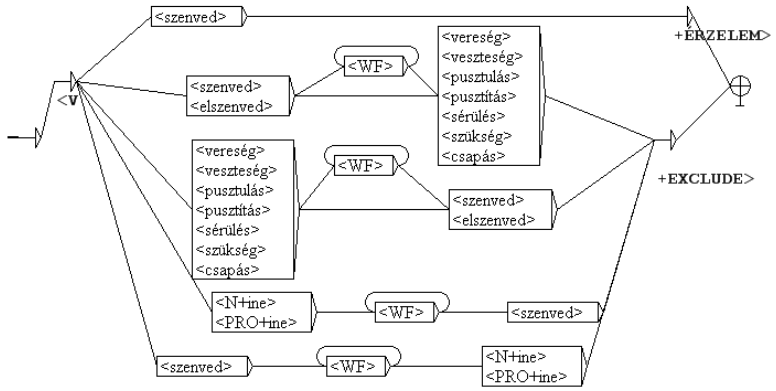
5. ábra A kognitív szótár egy részlete: az igék igekötő nélküli formáinak gráfja.

2.3 A szótárak átfedései

Az igék szövegkörnyezetüktől függően más és más jelentéssel bírhatnak, ezért előfordult, hogy ugyanaz az ige több szótárnak is a részét képezte. Többek között a 'szenved' igét kezelni kellett mind az érzelemszótáron, mind az aktivitás-passzivitás szótárán belül. Míg a 'vereséget szenved' kifejezésnek jelentése folytán PASSZIV kimenetelt kell kapnia, addig a 'valaki szenved valaki miatt' kifejezést ÉRZELEM-ként kell felismernie. Az alábbi két ábrán látható, miként oldottuk meg ezt a problémát lokális nyelvtanok segítségével.



6. ábra A 'szenved' ige gráfja az aktivitás-passzivitás szótárból.



7. ábra A 'szenved' ige gráfja az érzelemszótár részeként.

3 A vizsgálat

3.1 A vizsgálat kérdésfelvetése

A vizsgálat során a magyarok (saját csoport; in-group) és más nemzetek (out-group) megjelenítésének különbségeit kerestük. Arra voltunk kíváncsiak, hogy mely eseményeknél és milyen mértékben jelenik meg az in-group, illetve az out-group ágensként, mely eseményeknél és milyen mértékben ábrázolják a saját csoportot és más csoportokat mentális állapotok segítségével. Továbbá kerestük a magyarokra illetve más nemzetekre jellemző érzelmi mintázatokat.

3.2 A vizsgálat anyaga

Vizsgálati anyagunkban kétféle szövegkorpuszt használtunk: egyrészt több kiadótól származó általános és középiskolai tankönyvek – a magyar történelem főbb eseményeit tartalmazó – szövegrészleteit alkalmaztuk (kb. 150 ezer szó), másrészt néphistóriai szövegeket használtunk, melyek 500 fős – életkor, iskolázottság és etnikai hova-

tartozás mentén – rétegzett mintával készültek (kb. 64 ezer szó). (Ez utóbbiban az általuk legpozitívabbnak és legnegatívabbnak tartott magyar történelmi események elbeszélésére kérték a vizsgálati személyeket.) Az alábbi táblázatban láthatóak a vizsgálatban szerepelt történelmi események:

1. táblázat: A vizsgált történelmi események

Pozitív	Negatív	Pozitív és negatív
Honfoglalás	Tatárjárás	Török uralom
Államalapítás	Trianon	Habsburg uralom
Rendszerváltás	II. világháború	1956-os forradalom
	Holokauszt	

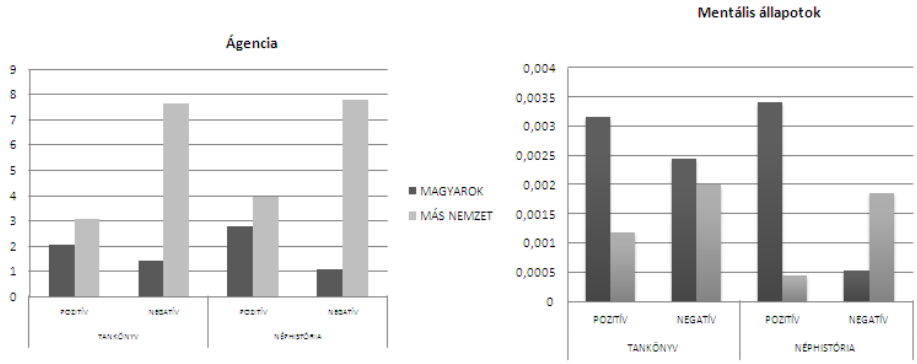
Mindegyik szövegkorpuszon lefutattuk az aktivitás, az intencionalitás, az érzelmek és a kognitív állapotok gráfját.

4 Eredmények

4.1 Eredmények az események valenciája tekintetében

A pozitív eseményeknél a saját csoport szignifikánsan több kognitív állapotot birtokol mindkét szövegcsoportban, azaz mind a történelemtankönyvekben, mind pedig a néphistóriában jóval gyakrabban ismerhetjük meg a magyar szereplők gondolatait, hiedelmeit. A negatív eseményeket a történelemtankönyvek igyekeznek több szempontból – a külső csoport nézőpontjából is - megközelíteni, serkentve a kölcsönös perspektíva-felvételt, ezáltal csökkentve a csoportközi konfliktust. A naiv történelemtörténetek negatív eseményeiben viszont szignifikánsan gyakrabban jelentek meg a külső csoport szereplőinek kognitív állapotai. (Ld.: 8 ábra)

A pozitív és a negatív események tekintetében változik a saját csoport intencionalitása és aktivitása is: a néphistória pozitív eseményeiben a magyar csoportot magasabb ágenciaértékkel ábrázolják, míg a negatív eseményeknél alacsonyabbal. Ez a tendencia pont ellenkező módon jelentkezik más nemzetek ábrázolásánál. A tankönyveknél az in-group megközelítőleg azonos ágenciaértéket mutat a pozitív és a negatív eseményeknél, viszont az out-group a néphistóriához hasonló tendenciával jelenik meg. (Ld.: 8 ábra)



8. ábra A kognitív állapotok kifejezéseinek gyakorisága illetve az ágencia arányszámai a pozitív és negatív eseményeknél a két szövegkorpuszban.

Az érzelmi kifejezések tekintetében a saját csoport esetén a pozitív néphistóriai szövegek kivételével mindenhol a negatív érzelmek dominálnak, vagyis a saját csoport akkor is negatív érzelmeket kapott többségben, ha pozitív eseményről volt szó.

4.2 Eredmények a saját csoport és más nemzetek csoportjainak tekintetében

Az ágenciaértékek szerint a legtöbb egyedi esemény tekintetében (a tankönyveknél 13-ból 11-ben, a néphistória esetében 13-ból 9-ben) az out-group jelenik meg nagyobb cselekvőképesség, szándékteliség birtokában.

Jelentős eltérések találhatók a saját, illetve a külső csoport érzelmi reakcióinak eloszlásában: a magyarok legjellemzőbb érzelmei közt a félelem, a remény és a lelkesedés tarthatók számon. A külső csoport esetében a domináló érzelmek a félelem, az öröm, a tisztelet és a bizalom. A két csoport érzelmeinek összevetésében különbséget találunk a szomorúságban, mely sokkal inkább a magyar nemzethez kötődik, emellett a remény érzése is szignifikáns mértékben inkább a saját csoport jellegzetes érzelmi élménye, eltérés található a megvetésben, amely inkább a más nemzetbeliek sajátja és általában a pozitív kapcsolati érzelmekben (szeretet, rajongás, rokonszenv, vonzás, tisztelet, bizalom) szintén a külső csoport javára.

4.3 Eredmények a kulturális és a kommunikatív emlékezet tekintetében

Az emlékezet szempontjából a kognitív állapotok eloszlása sem egyenletes. A kulturális emlékezet eseményeiben a magyar csoport kognitív túlsúlya figyelhető meg, míg a kommunikatív emlékezetben épp fordítva, más nemzetek kognitív túlsúlya jelenik meg.

Néphistóriai szövegekben sokkal több érzelem tűnik fel a kommunikatív, vagyis a közelmúlt eseményeit tartalmazó emlékezet szövegeinél, míg tankönyvekben a kulturálisnál tehát a távolabbi múlt történéseit magába foglaló emlékezetnél vannak többségben az érzelmek.

4 Megvitatás

A közelmúlt eseményei a naiv elbeszéléseknél sokkal intenzívebb érzelmi reakciókat hívnak elő, hiszen a feldolgozás alatt álló események sajátja az érzelmek aktív megosztása. A történelemkönyvek esetén ez törvényszerűen kevéssé érvényesül, hiszen ott nem a feldolgozás folyamata hangsúlyos, hanem az események közelebb hozása a befogadóhoz, így azokban a régmúlt eseményei érzelmekkel telítettebbek.

Továbbá mind a két szövegcsoportban jelentősen több kognitív állapot kapcsolódik a magyar csoporthoz a pozitív események tekintetében, illetve a kulturális emlékezet részét képező eseményeknél. Ez fontos az azonosulás elősegítése érdekében, hiszen ezen események képezik a nemzeti identitásunk alapját, és közvetítik számunkra a pozitív nemzeti érzést.

Érdekes eredmény, hogy a magyarok esetében sokkal inkább a negatív érzelmek dominálnak. Ez arra enged következtetni, hogy a magyar történelmi események nyomán létrejött az érzelmeknek egy olyan mintázata, melyben – igazodva a történelmi tapasztalatokhoz – negatív érzelmi reakciók kapcsolódnak a saját csoporthoz. Sőt, az egyes érzelmek eloszlását tekintve úgy tűnik, hogy a történelmi pálya alakulása nyomán létrejött az érzelmeknek egy jellegzetes konfigurációja, amely jellegzetesen a magyar nemzethez múltjához kötődik, hiszen egyértelmű különbségek találhatók a saját és a külső csoportnak tulajdonított érzelmek között. A történelemkönyvekben a külső csoport érzelmei a saját csoport vonatkozásában fogalmazódnak meg, vagyis számunkra az az érdekes, hogy ők tisztelnek, megvetnek-e bennünket. A saját csoport esetében pedig a félelem, a remény, a lelkesedés és a szomorúság bizonyultak tipikus érzelmi válaszoknak, amelyek mind jól illeszkednek a magyar nemzet egyedi történelmi tapasztalataihoz, illetve ahhoz az ágenciaeredmények által sugallt képhez, hogy más nemzeteket saját csoportunkhoz képest erősebbnek érzünk.

A néphistóriai szövegek a kommunikatív emlékezet részeként inkább más nemzetek perspektíváját hangsúlyozzák a kognitív állapotok találati alapján. Azt feltételezzük, hogy más nemzetek kognitív állapotainak gyakorisága ebben az esetben inkább a felelősséget jelölik, semmint a megértést. Ugyanezen eseményeknél a saját csoport kicsi ágenciaértéke – az out-group magas értéke mellett – azt mutatja, hogy a naiv elbeszélőknél a magyarság mint a nagy nemzeteknek kiszolgáltatott, kevés intencióval és cselekvőképességgel rendelkező népként reprezentálódik. Ezen eredmények viszont felvetik a saját sorsunk iránt való felelősséget vállaló – vagy felelősséget nem vállaló magatartás kérdését.

Mint látható, a történelmi szövegek az események leírása mellett a fent összefoglalt narratív eszközök segítségével jelenítik meg és közvetítik a csoportidentitást.

Hivatkozások

1. Abelson, R. P., Dasgupta, N., Park, J. & Banaji, M. R.: Perception of the collective other. Personality and Social Psychology Review, 2, (1998) 243-250
2. Assmann, J.: A kulturális emlékezet. Budapest: Atlantisz Kiadó. (1999)
3. Bamberg, M., and Andrews, M.: Introduction. In: Bamberg, M. and Andrews, M. (eds.): Considering Cuonter-Narratives: Narrating, resisting, making sense. Amsterdam: John Benjamins. (2004)
4. Bandura, A.: Perceived self-efficacy in the exercise of personal agency. The Psychologist: Bulletin of the British Psychological Society, 2, (1989) 411–424
5. Bandura, A. Self-efficacy. In V. S. Ramachaudran (Ed.), Encyclopedia of human behavior (Vol. 4, 71-81). New York: Academic Press. (1994)
6. Brewer, M. B., Hong, Y. & Li, Q.: Dynamic entitativity: Perceiving groups as actors. In V. Yzerbyt, C. Judd, & O. Corneille (Eds.), The psychology of group perception: Perceived variability, entitavity, and essentialism (25-38). New York: Psychology Press. (2004)
7. Brockmeier, J., Carbaugh, D. (Eds.): Narrative and identity: Studies in autobiography, self and culture. Amsterdam/Philadelphia: John Benjamins. (2001)
8. Gyáni, G.: Relatív történelem, Typotex Kiadó, Budapest. (2007)
9. Erikson, E.: Identity and the Life Cycle. Selected papers. (1959)
10. Ferenczhalmy R., és László J.: Az intencionalitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
11. Fülöp É., és László J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
12. Freeman, M.: Rewriting the self: History, memory, narrative. London: Routledge. (1993)
13. Lazarus, R. S.: Psychological stress and the coping process. New York: McGraw-Hill. (1966)
14. McAdams, D. P.: Power, intimacy, and the life story: Personological inquiries into identity. New York: Guilford Press. (1985)
15. McAdams,, D. P.: Coding Autobiographical Episodes for Themes of Agency and Communion. URL: http://www.sesp.northwestern.edu/docs/Agency_Communion01.pdf (2001)
16. Piaget, J.: The Moral Judgment of the Child. London: Kegan Paul, Trench, Trubner and Co. (1932)
17. Ricoeur, P.: A narratív azonosság. In: László, J. – Thomka, B. (szerk.) Narratív pszichológia, Narratívák 5. Budapest: Kijárat Kiadó. 15. (2001)
18. Silberztein, M.: NooJ Manual: a Linguistic Annotation System for Corpus Processing. (2008)
19. Spencer-Rogers, J., Hamilton, D. L., & Sherman, S. J.: The central role of entitativity in stereotypes of social categories and task groups. Journal of Personality and Social Psychology, 92, (2007) 369-388
20. Szalai K. és László J.: Az aktivitás-passzivitás modul kidolgozása Nooj tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. (2006)
21. Tajfel, H.: Human groups and social categories. Cambridge: Cambridge University Press. (1981)
22. Vincze O. és László J.: A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. (2006)

23. Vincze O. Mentális állapotok jelentősége csoporttörténetekben a saját és a külső csoport vonatkozásában.. PhD értekezés. (2009)
24. White, H. A történelem terhe, Budapest: Osiris Kiadó. (1997)

A személy- és csoportközi értékelés pszichológiai szempontú elemzése elbeszélő szövegekben

Csertő István

Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
csertopi@gmail.com

Kivonat: A Pécsi Tudományegyetem Pszichológiai Intézetének és az MTA Pszichológiai Kutatóintézetének narratív pszichológiai kutatócsoportja egy, az automatizált narratív pszichológiai tartalomelemzést lehetővé tevő módszer fejlesztésén dolgozik. A módszer az élettörténeti – önéletrajzi és csoporttörténeti – szövegek számítógéppel támogatott elemzésével olyan nyelvi markereket azonosít, amelyek szövegbeli mintázata összefüggésbe hozható különböző pszichológiai dimenziókkal, így a kapott kvantitatív adatok alapján a személyes ill. csoportidentitás állapotaira vonatkozó diagnosztikus és prediktív következtetések tehetők. A kutatócsoport több, azonos elven működő számítógépes elemzőeszközt, modult fejlesztett ki, melyek mindegyike egy-egy meghatározott pszichológiai dimenzió nyelvi markereit vizsgálja. A modulok a NooJ nyelvtechnológiai rendszerben működnek, amely lehetővé teszi a digitalizált szövegek megadott szempontok alapján történő morfológiai és szintaktikai elemzését, és erre épülve meghatározott nyelvi alakzatok azonosítását a szövegeken belül. A cikk a személy- és csoportközi értékelés moduljának elméleti hátterét és technikai megvalósítását mutatja be.

1. Narratív pszichológiai tartalomelemzés

1.1. Narratív pszichológia

Az utóbbi évtizedekben egy új megközelítés bontakozott ki a pszichológiában, a narratív megközelítés [1]. Az új paradigma szemléletében szembeállítható a korábban uralkodó klasszikus kognitív pszichológiával. A narratív megközelítés egyik képviselője, Sarbin a két irányzatot a jelenségek modellezésében és értelmezésében általuk használt alapvető analógia vagy tő-metafora alapján állítja szembe egymással [2]. A kognitív pszichológia tő-metaforája a mechanizmus: az elme működését a fizikai világ determinizmusa mintájára képzeli el, amely szigorú oksági viszonyoknak engedelmeskedik. Ebben a felfogásban az emberi lények a számítógéphez hasonlóan működnek: a körülöttük lévő világ értelmezését, a jelentésadás folyamatait passzív, nem konstruktív módon, a tudatos hozzáférés elől elrejtett információ-feldolgozó mechanizmusok segítségével valósítják meg. Ezzel szemben a narratív pszichológia tő-metaforája az elbeszélés: az észlelés, emlékezés, gondolkodás stb. folyamatait a történetszövés folyamatához hasonlítja, melynek során az emberek tapasztalataikat nar-

ratív struktúrákba szervezik. Az időben és térben érintkező eseményeket összefüggő cselekményként konstruálják meg, melyben az események közötti kapcsolatokat emberi szándékok, tervek, érzelmek és ítéletek adják, és amely mint a cselekvések értelmezési kerete meghatározza az egyes események jelentését és értelmét. Az emberek saját szándékaikat, elvárásaikat és cselekvéseiket ehhez a megkonstruált valósághoz igazítják.

Bruner, a narratív fordulat egyik vezető alakja szintén arra világít rá, hogy mindennapi gondolkodásunk narratív természetű [3]. Bruner a gondolkodásnak két alapformáját feltételezi, amelyek egyike sem vezethető vissza a másikra, ezek a paradigmatikus és az elbeszélő mód. A paradigmatikus vagy logikai-tudományos mód célja az egyes események megfigyeléséből általánosan érvényes oksági viszonyokra és igazságfeltételekre következtetni, és ezeket absztrakt fogalmi eszközökkel leképezni, oly módon, hogy eredményként a valóság objektív képét kapjuk meg. Ezzel szemben az elbeszélő mód célja és eszközei egészen más természetűek. A konkrét eseményekből nem az objektív valóságot, hanem egy hihető és értelmes történetet igyekszik kikerekíteni. Emberi szándékokkal és cselekedetekkel foglalkozik, s célja egy olyan pszichológiai realitás megteremtése, amely az eseményeket értelemmel ruházza fel.

A mindennapi emberi gondolkodás és cselekvés elbeszélő természetének feltételezéséből következik a narratív pszichológia alapvető célkitűzése: azonosítani azokat a narratív elveket és mintákat, amelyek az emberi élményeket használható tapasztalattá szervezik, és feltárni az egyének illetve csoportok által létrehozott narratívumok jellemzői és pszichológiai funkcióik közti összefüggéseket (vö.[1, 2]).

1.2. Az identitás mint narratívum

Narratív pszichológiai megközelítésben az identitás maga is egyfajta narratívum: olyan történet, melynek főhőse maga az elbeszélést létrehozó egyén illetve csoport [1, 4, 5]. Az önazonosság folytatólagos tudatát és az értékesség érzését az élettörténet koherenciája adja, az egymást követő események töretlen oksági láncolata, ellentmondás-mentessége és egy pozitív jövőbeli cél felé mutató iránya. A jól működő identitás feltétele a koherens élettörténet. Az emberek szándékaikat, elvárásaikat, terveiket a koherencia elvének megfelelően alakítják ki. Ebben az értelemben a múlt tapasztalatára épül a jelen és a jövő. Ugyanakkor az identitás mint narratívum, akár egyéni, akár csoportidentitásról van szó, soha nem egyéni teljesítmény, hanem társas konstrukció: annak a szüntelenül működő kölcsönös egyeztetési folyamatnak az eredménye, amelyben az egymás életében szerepet játszó emberek egymásról és önmagukról fenntartott történeteiket összehangolják [2, 4, 6]. Az identitás mint szelfnarratívum csak akkor működőképes, ha összhangban áll a környezetnek az egyénnel kapcsolatos vélekedéseivel, elvárásaival, céljaival.

Az élettörténet szerkezeti és tartalmi jellemzőiből következtetni lehet az egyén identitásának különböző aktuálisan érvényes minőségeire [1, 4, 5]. Gergen és Gergen [4] pl. egy egyszerű kísérlettel demonstrálta a szubjektív jóllét két életkorilag jellemző narratív mintáját az amerikaiak körében. Fiatal felnőtteket arra kértek, hogy grafikus módon, egy „életvonal” segítségével ábrázolják addig eltelt életük alakulását a jóllét szempontjából, egy idősekkel végzett korábbi interjúvizsgálat eredményeit

pedig átírták a fiatalok eredményeivel összevethető grafikonná. Két, egymással ellentétes pályát leíró görbét kaptak, amelyek egymástól eltérő további életpályák irányába mutattak. Ami itt lényeges, hogy különböző élethelyzetekben az élettörténet különböző formákban fogalmazódhat újra, eltérő hangsúlyokkal és értékelésekkel, és a narratívum tartalmi és formai sajátosságai pszichológiai implikációkat hordoznak.

1.3. Az identitás-narratívumok pszichológiai szempontú tartalomelemzése

A PTE Pszichológiai Intézetének és az MTA Pszichológiai Kutatóintézetének narratív pszichológiai kutatócsoportja egy, az automatizált narratív pszichológiai tartalomelemzést lehetővé tevő módszer fejlesztésén dolgozik. A módszer az élettörténeti – önéletrajzi és csoporttörténeti – szövegek számítógéppel támogatott elemzésével olyan nyelvi markereket azonosít, amelyek szövegbeli mintázata összefüggésbe hozható különböző pszichológiai dimenziókkal, így a kapott kvantitatív adatok alapján a személyes, ill. csoportidentitás állapotaira és folyamataira vonatkozó diagnosztikus és prediktív következtetések tehetők [1, 7].

A kutatócsoport több, azonos elven működő számítógépes elemzőeszközt, modult fejlesztett ki, melyek mindegyike egy-egy meghatározott pszichológiai dimenzió nyelvi markereit vizsgálja [7]. Az alábbiakban a személy- és csoportközi értékelés moduljának elméleti hátterét és technikai megvalósítását mutatom be (ld. még [6, 8]).

2. A személy- és csoportközi értékelés pszichológiai elemzése identitás-narratívumokban

2.1 A személy- és csoportközi értékelés szerepe a szociális identitás fenntartásában

Az értékelésnek a narratívumok megkonstruálásában betöltött központi szerepét Labov és Waletzky [9, 10] mutatta ki, akik személyes élményekről adott beszámolókat elemeztek strukturális szempontból. A szerzők a narratívumok két általános funkcióját állapították meg, amelyekkel egy kommunikatív szándékot megvalósító történetnek rendelkeznie kell. Ezen funkciók egyike az értékelés, amely végső soron egyenlő az elbeszélés mint kommunikatív aktus pragmatikai relevanciájával. Az események narrátori értékelése indokolja meg, hogy miért érdemes egyáltalán közölni a történteket, mi az elmondottak üzenetértéke. Az értékelés azért szükségszerű része a narratívumnak, mert ez mindig egy olyan cselekményt mutat be, amelyben az események elvárt, normálisnak tekintett menetétől való eltérés, valamilyen drámai fordulat következik be, és az értékelés az, ami ezt a fordulópontot megjelöli a hallgató számára.

A jelentősnek ítélt életesemények többnyire nem magányos helyzetekben, hanem társas kapcsolatok kontextusában, mások aktív részvételével zajlanak. Ennélfogva az értékelés természetszerűleg kiterjed az eseményekben érintett szereplőkre, személyekre és csoportokra is, kifejezve a narrátor hozzájuk való viszonyulását, közelségét vagy távolságát, hovatartozását, valamint a szereplők egymáshoz való viszonyáról

kialakított képét. Ezek a személy- és csoportközi értékelések, amelyek a történetekben megjelenhetnek a jutalmazás és büntetés aktusaiban, a szereplők cselekvéseinek pozitív vagy negatív interpretációiban, érzelmi reakciókban, illetve jó és rossz vonások tulajdonításában, alapvető szerepet játszanak a szociális identitás fenntartásában. A szociális identitás elmélete [11, 12] azon a tézisen alapul, hogy az egyének önazonosságukat jelentős mértékben azoktól a csoportoktól nyerik, melyeknek tartósan tagjai, és amelyek életükben meghatározó szerepet töltenek be. Egy pozitívan értékelt tagsági csoport pozitív önértékelést és a valahová tartozás biztonságát nyújtja az egyén számára. A szociális identitás azonban nem abszolút, hanem relációs kategória: a saját csoport értékét más, vele azonos típusú külső csoportoktól való pozitív megkülönböztetettsége adja. Az egyén ugyanakkor egyszerre számos csoportnak tagja, és mindig az aktuális társas szituáció határozza meg, hogy mely szociális kategória válik a megkülönböztetés alapjává. A pozitív szociális identitás igénye csoportközi összehasonlításhoz és elfogultsághoz vezet, azaz a saját csoport fel- és a külső csoport leértékeléséhez, amely megjelenhet sztereotipizálásban, diszkriminatív viselkedésben vagy agresszív versengésben. Terep- és laboratóriumi kísérletek demonstrálták, hogy a csoporthoz tartozás puszta ténye képes elindítani a csoportközi összehasonlítás és versengés folyamatait, felülírva akár a korábbi személyes barátságokat (pl. [13, 14, 15]). Attribúciós kísérletek azt igazolták, hogy a csoportközi elfogultság a viselkedésmagyarázatokban is megjelenik: a saját csoportot az egyének inkább annak sikereiért, míg a külső csoportot saját kudarcaiért teszik felelőssé [16]. Újabb vizsgálatok a stratégikus nyelvhasználatban is kimutatták a csoportközi elfogultság hatását [17].

Csoportközi kontextusban tehát a személy- és csoportközi értékelés mind viselkedéses, mind verbális formában elfogultságot mutat, melynek motivációs hátterét a pozitív szociális identitás fenntartásának igénye adja. Az értékelésbeli elfogultság a csoport jólétét fenyegető, kiélezett konfliktushelyzetekben felerősödik, megerősítve a csoportkohéziót és a kollektív azonosságtudatot.

2.2 A személy- és csoportközi értékelés narratív pszichológiai vizsgálata

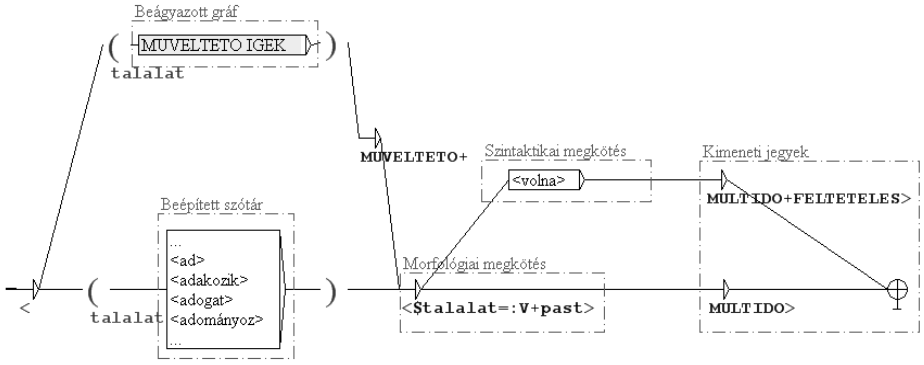
A csoportközi elfogultság létezésének ténye, a csoportidentitás dinamikájára visszavezethető volta és a társadalom életében betöltött jelentős szerepe indokolttá teszi különböző csoporttörténeti narratívumok értékelés szempontú tartalomelemzését. A tematika, keletkezési idő, forrás stb. szerint különböző narratívumokban szereplő saját és releváns külső csoportokra vonatkozó pozitív és negatív értékelések relatív gyakoriságai alapján a csoportidentitás dinamikájára vonatkozó hipotézisek fogalmazhatók meg és ellenőrizhetők. A magyar nemzeti történelem laikus és hivatásos elbeszélései kapcsán pl. vizsgálhatók a következő kérdések: Jelen van-e a történelemről szóló laikus történetekben a kérdőívvel kimutatható nacionalizmus? Megállapítható-e valamilyen összefüggés a nacionalizmus mértéke és a laikus történelemreprezentáció értékelő tartalma között? Megjelenik-e a csoportközi elfogultság az olyan hivatásos elbeszélésekben, mint a történelemkönyvi szövegek? Hogyan alakul a csoportközi elfogultság az olyan, tematika szerint különböző események narratívumaiban, mint pl. az eredetmítosz, a történelmi traumákról vagy a nemzet fénykoráról szóló történetek? Jellemezhetők-e ezek az események az értékelési aszimmetria sajátos mintázataival, az elfogultság két oldalának, a saját csoport felér-

tékelésének és a külső csoportok leértékelésének eltérő relatív hangsúlyaival? Mindezek olyan kérdések, melyek megválaszolásához közelebb vihet a narratív pszichológiai kutatócsoport által kifejlesztett számítógépes elemzőeszköz, amely a narratívumokban előforduló személy- és csoportközi értékeléseket képes azonosítani és mennyiségi adatokká átalakítani. (A nemzetitörténelem-reprezentációk más vizsgálatairól l. [1, 7])

3. A személy- és csoportközi értékelés számítógépes elemzése

3.1 Szövegelemzés a NooJ program segítségével

A narratív pszichológiai kutatócsoport számítógépes elemzőmoduljai a Max Silberztein által kifejlesztett NooJ nyelvtechnológiai rendszerben [18] szerkeszthetők és futtathatók, amely több nyelvben lehetővé teszi nagy terjedelmű digitalizált szövegkorpuszok morfológiai és szintaktikai elemzését, és erre épülve meghatározott nyelvi alakzatok azonosítását a szövegekben. A modulok az elemzési szempontokat meghatározó algoritmusokból, ún. gráfokból állnak, amelyek a NooJ grafikus kezelőfelületén szerkeszthetők. A gráfok egyrészt szótárakat, másrészt morfológiai és szintaktikai megkötéseket tartalmaznak. Az egyes gráfok akkor azonosítanak találatként egy adott szövegrészletet, ha az (1) tartalmazza a beépített szótárakban szereplő valamely elemet, és ugyanakkor (2) az azonosított szótári elem alakja és szövegkörnyezete megfelel a gráfban kódolt morfológiai és szintaktikai megkötéseknek. A kapott találatokról a NooJ-modul listát készít, amelyben az egyes találatokat az előzetesen kategorizált találati típusoknak megfelelő kimeneti jeggyel látja el. A kapott mennyiségi adatok statisztikai módszerekkel elemezhetők, ill. a találatok visszakereshetők a szövegben, ami további kvalitatív elemzést is lehetővé tesz.



1. ábra. NooJ-mintagráf. Illusztráció a NooJ rendszerben futtatható gráfok működési elvére.

A NooJ programban futtatható gráfok működési elvét és a grafikus kezelőfelület alkalmazását az 1. ábrán látható mintagráffal illusztrálom. Az ábra bal oldalán lévő nyíl szimbolizálja a szövegre alkalmazott elemző algoritmus kezdőpontját, a jobb oldalon lévő célkereszt pedig a végpontot, vagyis a keresési folyamat lezárását. Mindent, ami a két végpont között helyezkedik el, vagyis a szótárakat, a morfológiai-szintaktikai megszorításokat és az ezek között fennálló kapcsolatokat a felhasználó építi fel. A NooJ-gráf grafikus megjelenítése a keresési folyamat egymást követő lépéseit balról jobbra haladva szimbolizálja. A nyíl hegyétől induló folyamatos vonalak különböző elemzési útvonalakat jelölnek, amelyekből egyszerre több is futhat párhuzamosan egy gráfon belül.

A mintagráf alsó elemzési szálán egy beépített szótár látható, amely a keresett nyelvi alakzatokban szereplő szavakat tartalmaz. A keresett alakzatok lehetnek önmagukban az egyes szótári elemek, de a NooJ lehetővé teszi összetett szekvenciák azonosítását is. A mintagráfban egy igeszótár részletét jelöltem. A kapcsos zárójelekben a szavak szótári alakjai szerepelnek. Ezek kiegészíthetők morfológiai annotációs jegyekkel, amelyek specifikálják a szövegben keresett elemek morfológiáját. A NooJ a szótári elemek különböző toldalékolt alakjait a háttérben futó, morfológiai szempontból annotált alapszótárak alapján képes felismerni.

A mintagráfban szereplő szótártól jobbra lévő nyílhegy egy morfológiai megszorítást jelöl, amelyet a <$talalat=:V+past> parancs ad meg. A megszorítás ebben a példában arra vonatkozik, hogy a beépített szótári elemeknek csak azon alakjait azonosítsa találatként a gráf a szövegben, amelyek szófaja ige (V), igeideje pedig múlt idő (+past). A parancskódban szereplő „talalat” kifejezés rendeli hozzá a megszorítást az előtte álló szótárhoz, amelyet mint referenciát az azonos kifejezéssel ellátott kerek zárójelek jelölnek. Ezen az elemzési szálon tehát a szótárban szereplő igék múlt idejű alakjait azonosítja a gráf a szövegben.

Az azonosított elemeket a gráf a benne jelölt „MULTIDO” kimeneti jeggyel látja el. A kimeneti jegyek kétféle módon hasznosíthatóak. Egyrészt a teljes elemzett szöveg exportálható úgy, hogy a találatok a szövegen belül, eredeti helyükön jelölve vannak a kimeneti jeggyel. Ez lehetővé teszi, hogy a kutató megvizsgálja a találatok szövegen belüli elhelyezkedésének mintázatát, illetve a találatok szövegkörnyezetét. Másrészt a kimeneti jeggyel ellátott találatokból konkordancialista kérhető a NooJ-ban, amely a gyakorisági adatok statisztikai elemzését teszi lehetővé.

Az ábrán látható mintagráf alsó fő elemzési szálából kiinduló mellékszál egy egyszerű példa arra, hogyan lehet a NooJ segítségével az egyes szavak szintjén túllépve szekvenciákat is azonosítani. A mellékszál a múlt idejű ige és a közvetlenül utána álló „volna” ige együtteseit azonosítja a szövegben, amelyeket külön kimeneti jeggyel lát el („MULTIDO+FELTETELES”). Nem csupán konkrét szóalakokat, hanem egész szófajokat is meg lehet adni ilyen szintaktikai megszorításként, a megfelelő szófaji kóddal. (Pl. ige: <V>.)

A gráf kezdőpontjából kiinduló másik mellékszálon egy beágyazott gráf található, amelyet a „MUVELTETO IGEK” feliratú doboz jelöl. Az illusztráció célja szerint ez a műveltető igéket tartalmazza, amelyek kimenete a megkülönböztető „MUVELTETO+” jeggyel bővül. A beágyazott gráf a fölérendelt gráffal azonos módon működik, és ugyanúgy szerkeszthető, miután előhívtuk a grafikus kezelőfelületen. A beágyazott gráfokra, ha csupán szótárakat tartalmaznak, szintén alkalmazhatók morfológiai megszorítások a fő gráfban, ahogyan az az ábrán látható. Többszörös

beágyazás is lehetséges, vagyis a beágyazott gráfokba további beágyazott gráfok építhetők. A beágyazott gráfok két okból hasznosak. Egyrészt a kompakt ábrázolás és a hierarchikus struktúra sokkal áttekinthetőbbé és kezelhetőbbé teszi a gráfot, mint ha minden komponense egy szinten helyezkedne el. A teljes gráfstruktúra külön kezelőfelületen előhívható és szerkeszthető. Másrészt egy adott címkével ellátott beágyazott gráfot akárhányszor újra felhasználhatunk másutt a fő gráfon belül, csupán a megfelelő címkét használva, anélkül, hogy minden alkalommal újra fel kellene építenünk. Ez jelentősen növeli a munka gazdaságosságát.

3.2 A személy- és csoportközi értékelés elemzőmoduljának felépítése és működése

A személy- és csoportközi értékelés számos különböző nyelvi szerkezetben valósulhat meg a szövegben, amelyek jelentős részét képes azonosítani az e célból kifejlesztett NooJ-modul [6, 8]. Az értékelés modul a fentebb ismertetett általános elemzési elvnek megfelelően működik, vagyis a nyelvi szerkezetek azonosítása az értékelő tartalmat hordozó kulcsszavak szótárain alapul. A modul azokat a szerkezeteket azonosítja, amelyek (1) tartalmazzák a beépített szótárakban szereplő valamely kulcsszót, és (2) megfelelnek a gráfokban kódolt morfológiai és szintaktikai feltételeknek. A modul jelenleg több, hierarchikusan felépülő fő gráfból áll, amelyeket az értékelést hordozó kulcsszó szófaja alapján különítettünk el egymástól.

Az értékelő kifejezések szótárai

Az értékelést hordozó kulcsszavakat több külön szótárba rendeztük, részben szófaji alapon, részben pedig egyéb szemantikai jellemzők alapján (ld. 1. táblázat). A kulcsszavak szófaj szerint lehetnek igék, melléknevek, főnevek és határozók. Jelenleg ige-melléknév- és főnévszótáraink vannak, a határozó-szótárak a közeljövőben készülnek el. Az ige- és melléknév-szótárakat a MTA Nyelvtudományi Intézetének használati gyakoriság szerint összeállított digitális szótáraiból válogattuk ki Bigazzi Sára irányításával, aki az értékelés modult első, eredeti formájában kifejlesztette [6, 8]. A jelenleg létező főnévszótárakat Gábor Kata közreműködésével generáltuk a melléknévszótárakból, ezeket további szótárakkal fogjuk kiegészíteni a közeljövőben.

Az egyes szófajokon belül elkülönítettük a pozitív és a negatív értékelések szótárait. Az igéket ezen túlmenően további szemantikai osztályokba soroltuk az értékelő perspektíva szerint, vagyis aszerint, hogy a narrátor vagy valamelyik szereplő értékelését fejezik ki. Az értékelő igék fizikai cselekvéseket vagy érzelmi, illetve egyéb mentális állapotokat írnak le. Az érzelmi és mentális igék (pl. szeret, tisztel) jellemzően a szereplők értékeléseit tükrözik, kivéve az első személyű eseteket, amelyek a narrátor mentális állapotait írják le. Ezek a NooJ-ban megkülönböztető kimenettel elkülöníthetők. Az érzelmi állapotok azonosítására külön modul készült Fülöp Éva fejlesztésében [19], melynek az értékelés szempontjából releváns összetevőit a közeljövőben illesztjük be az értékelés modulba. A cselekvést leíró igék (pl. kritizál, bánt) egy része a szereplők, más része a narrátor értékeléseit fejezi ki. Az ún. értékelő aktusok a szereplők olyan aktusai, amelyekkel pozitív vagy negatív ítéletet fejeznek ki más szereplők irányában (pl. vádol, kritizál, méltat, dicsér). Az értékelő aktusok tehát

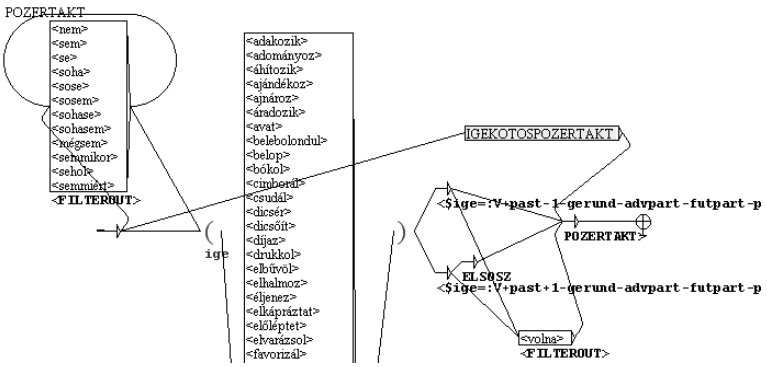
jellemzően a szereplőktől származó értékeléseket közvetítik. Itt is kezelendő kivételt képeznek az első személyű esetek. A cselekvő igék másik osztályát képezik az ún. morálisan értékelt aktusok. Ezek a szereplők olyan aktusai, amelyek maguk morálisan pozitív vagy negatív megítélés alá esnek (pl. helytáll, jóvátesz, kizsákmányol, visszaél). Ezek az igék nem megfigyelhető vagy tényszerű viselkedéseket írnak le, hanem a narrátor interpretációit közvetítik a szereplő viselkedésének értékéről (a kétféle igei leírás közti különbégről ld. [20]). A morálisan értékelt aktusok tehát a narrátornak a cselekvő ágensre vonatkozó értékeléseit közvetítik. Az értékelő perspektíva alapján történő osztályozás nem csak az igék, hanem a többi szófaj esetében is releváns. Ez a munka szintén a jövőben elvégzendő feladatok közé tartozik.

1. táblázat: Az eddig elkészült szótárak osztályozása szófaj, valencia és értékelő perspektíva szerint, példákkal.

Szófaj	Igeosztályok	Pozitív	Negatív	Perspektíva
Ige	Értékelő aktus	megdicsér	megbüntet	Szereplő
	Morálisan értékelt aktus	jeleskedik	hazudozik	Narrátor
	Érzelem és mentális áll.	szeret, tisztel	utál, lenéz	Szereplő
Mellékn.		kedves	buta	Narrátor
Főnév		ügyesség	gonoszság	Narrátor

Az értékelés modul gráfjai

A modul jelenleg három, hierarchikusan felépülő fő gráfot tartalmaz, egy igei, egy melléknévi és egy főnévi gráfot. Az igei fő gráf a pozitív és negatív igéket azonosító gráfokra oszlik, amelyek külön-külön tovább bomlanak az értékelő aktusokat és a morálisan értékelt aktusokat azonosító gráfokra. Mind a valencia, mind az értékelő perspektíva jelölve van a találatok kimeneti jegyében, valamint az első személyű alakok egy további megkülönböztető jegyet kapnak. A 2. ábra a pozitív értékelő aktusok gráfját mutatja, amely a többi, vele egy szinten elhelyezkedő gráfhoz hasonlóan épül fel. A pozitív értékelő aktusok beépített szótárára két morfológiai megszorítás vonatkozik, amely két külön elemzési szálon helyezkedik el. Az egyik elemzési szál a harmadik személyű alakokat azonosítja, a másik az első személyűeket, továbbá mindkét szál csak a kijelentő módú és múlt idejű alakokat azonosítja. A gráf az újabban kifejlesztett <FILTEROUT> funkcióval kizárja azokat az igéket, amelyeket tagadószó előz meg vagy a „volna” szó követ. A kizáró funkció úgy működik, hogy a NooJ kihagyja a találati listából azokat a találatokat, amelyek tartalmazzák a gráfban kizárandóként megjelölt szótári elemeket.



2. ábra. A pozitív értékelő aktusok gráfja. Az ebbe beágyazott gráf az elváló igekötős alakokat azonosítja.

A pozitív értékelő aktusok gráfja beágyazott formában tartalmazza az elváló igekötős alakokat azonosító algráfot, amely két további algráfra oszlik. Egyikük az ige után, másikuk az ige előtt álló elváló igekötős alakokat azonosítja. Ezekben is kizáródnak a tagadott, a „volna” szóval kombinált, illetve a nem kijelentő módú és nem múlt idejű alakok, továbbá az első személyű formák megkülönböztető kimenetet kapnak. Az értékelő szavak igekötő nélküli formában, az igekötő típusa szerint külön alszótárakban szerepelnek a gráfokban, minden alszótár a megfelelő igekötőhöz van kapcsolva, a lehetséges közbeékelődő szavakkal együtt.

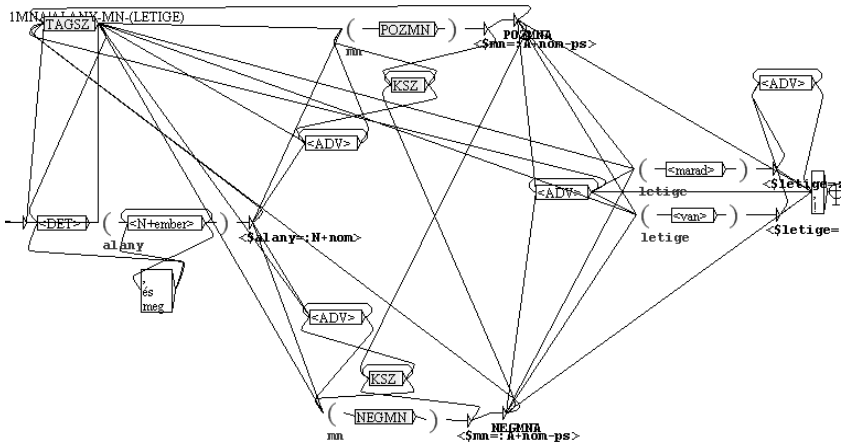
A melléknévi fő gráf két algráfra oszlik. Az egyik a melléknévi állítmányokat, a másik pedig a jelzős szerkezeteket azonosítja. Ezen a két nyelvtani szerkezeten kívül a melléknevek csak főnévi szerepben jelenhetnek meg, amely esetekre a jövőben további gráfok írandók. Az állítmányi algráf további négy algráfra oszlik a melléknév referenciája (a mondat alanya) szerint, amely a következő típusokba sorolható: 1. harmadik személyű humán referencia (pl. „A király bölcs.”), 2. harmadik személyű nem humán referencia (pl. „A trianoni békeszerződés igazságtalan volt.”), 3. első személyű referencia (pl. „Jók voltunk.”), 4. nincs referencia, beleértve a hiányzó anaforát is (pl. „Ügyes!”). A jelzős szerkezeteken belül csak a humán és nem humán referencia van elkülöníve (pl. „bölcs király”, „igazságtalan békeszerződés”). A referencia szerinti megkülönböztetés okai a következők. Az első személyű alakokat az értékelő perspektíva megállapítása miatt különítettük el, a nem humán referenciájú és a referencia nélküli szerkezeteket pedig azért, mert ezek közvetetten vonatkozhatnak a szereplőkre, tehát azonosítható a humán referencia. A gépi referenciaazonosítás fejlesztése jelenleg folyamatban van. A melléknévi állítmány négy algráfja további hat-hat algráfra oszlik, aszerint, hogy a szerkezetet alkotó alany, melléknév és esetleges létige milyen sorrendben szerepelnek a mondatban. Mivel a magyar nem kötött szórendű nyelv, ezért minden, nyelvhasználatilag lehetséges variációt figyelembe kell venni (a melléknévi fő gráf szintjeinek struktúráját l. 2. táblázat).

2. táblázat: A melléknévi fő gráf három szintjének struktúrája. * Csak a melléknévi állítmány gráfjában van.

1. szintű algráfok a nyelvi szerkezet típusa szerint	2. szintű algráfok az értékelés referenciája szerint	3. szintű algráfok* az alany-melléknév-létige (A-MN-L) szekvencia sorrendje szerint
Melléknévi állítmány Jelzős szerkezet	3. személy 3. személy nem humán 1. személy* nincs referencia*	A-MN-L A-L-MN MN-A-L MN-L-A L-A-MN L-MN-A

A 3. ábrán látható a harmadik személyű humán referenciájú melléknévi állítmány egyik specifikus variációját, az alany-melléknév-létige sorrendű szekvenciát kezelő algráf. Az algráf a mondat alanyát egy humán referenciájú főneveket tartalmazó, szemantikailag annotált szótár alapján azonosítja, amelyet Gábor Kata szerkesztett a narratív kutatócsoport számára. Az alany azonosítása után az elemzés két szálon folytatódik tovább, melyek a pozitív, illetve negatív mellékneveket azonosítják. A beépített melléknévszótárakra egy morfológiai megkötés vonatkozik, amely szerint csak az alanyesetben és birtokjelek nélkül álló melléknevek számítanak találatnak. A gráf a több elemből álló felsorolásokat is kezelni tudja egy, a melléknévből induló, önmagába záródó elemzési körrel, amely figyelembe veszi a felsorolt melléknevek közti vesszőket, kötőszavakat és határozókat (Pl. „Julcsa okos, kedves és nagyon csinos."). A pozitív és negatív melléknevek elemzési szála összekapcsolódik egymással oly módon, hogy lehetségessé válik az előjel szempontjából vegyes melléknévi állítmányok azonosítása is (pl. „Géza okos, de hanyag."). A kimeneti jegyek úgy vannak elhelyezve a gráfban, hogy a felsorolásban szereplő minden egyes melléknév a valenciájának megfelelő külön kimeneti jegyet kap, tehát a kimenetben annyi pozitív és negatív jegy jelenik meg, ahány melléknév a találatban szerepel. Ez azért fontos, mert minden egyes értékelő melléknév külön értékelésnek tekinthető, és így külön-külön beleszámít a kapott gyakorisági adatokba. A gráf az elemzés következő lépésében azonosítja a melléknevek után esetlegesen megjelenő létigét és határozókat, és végül a tagmondatot lezáró írásjelet. Ez utóbbi azért lényeges, mert a tagmondatvégi írásjel hiányában előfordulhat, hogy a melléknév nem a gráf által azonosított főnévre, hanem egy másikra vonatkozik, amely a mondatban a találat után áll (pl. „*Péter kedves* emberekkel találkozott."). A melléknévi állítmányok esetében a tagadásnak több lehetséges variációja létezik, a gráf ezek mindegyikét kizárja. (Pl. „*Nem* a fiú okos.", „A fiú *nem* okos.", „A fiú okos *nem volt*, …")

A jelzős szerkezetek fő gráfja a melléknévi állítmányok gráfjaihoz hasonlóan épül fel, azzal a különbséggel, hogy nem veszi figyelembe az azonosított szekvencia előtt és után álló elemeket.



3. ábra. Az alany-melléknév-létige sorrendű szerkezeteket kezelő egyik algráf.

A főnévi fő gráf két algráfból áll. Az egyik az értékelő főnevet tartalmazó birtokos szerkezeteket azonosítja (pl. „a király bölcsessége”), a másik az értékelő főnév + ige kombinációkat, amelyekben az ige rendeli hozzá az értékelő főnevet annak referenciájához (pl. „a hős *bátorságot tanúsít*”). Történelemkönyvi szövegek próbaelemzése alapján ez a két nyelvi szerkezet az, amely az algoritmizálható formák közül a leggyakrabban előfordul. Mindkét gráf további algráfokra bomlik, a melléknévi gráfokkal megegyező módon: egyrészt a négy lehetséges referenciatípust külön algráfok kezelik mindkét esetben, másrészt a „főnév + ige” négy algráfján belül az alany-főnév-ige szekvenciák különböző sorrendű variációit további külön gráfok azonosítják. A birtokos szerkezetek és a főnév + ige szerkezetek azonosítása az igei és melléknévi gráfoknál már bemutatott eszközökkel, ezeknek a keresett főneves szerkezetekre való specifikus alkalmazásával történik.

3.3 Az automatikus elemzés korlátai: explicit és implicit értékelések

Az automatizált tartalomelemzés szótárakra épülő módszere, bármennyire komplex is, nem vállalkozhat kimerítő hermeneutikai szövegfeldolgozásra, ami azt jelenti, hogy nem képes minden olyan tartalmat feltárni az elemzett szövegekben, amely értékelő jelentést hordoz. Vannak olyan eseményleírások, amelyek nem tartalmaznak a kontextustól viszonylag független értékelő jelentéssel bíró kulcsszót, implicit módon mégis értékelést fejeznek ki. A következő szövegrészlet erre mutat egy példát: “*Azt ígérte, eljön, de nem jött el. Azóta színét se láttam, és még arra sem volt képes, hogy felhívjon.*” Itt a narrátor egyértelműen értékítéletet közöl távolmaradó társával kapcsolatban, amely nagyjából lefordítható a „felelőtlen” kifejezésre. Ebben az esetben azonban nincs olyan kontextusfüggetlen nyelvi marker, amely alapján az értékelés automatikus elemzéssel azonosítható lenne. Ebből következően az értékelés modul csak az explicit értékeléseket képes azonosítani. A mindennapi kommunikációs tapasztalatainkból eredő intuíciónk alapján azonban azt feltételezhetjük, hogy ha a

narrátor kommunikatív szándéka az, hogy valakit egy esemény kapcsán értékeljen, akkor ezt valahol a szövegben explicit módon is megteszi (Pl. a fentebbi példamondatot követheti egy olyasféle megállapítás, hogy „kiborít ez az ember", vagy „szörnyen utálom az ilyet".) Ezek az explicit kifejezésformák pedig azonosíthatók (vagy a jövőben azzá válnak) az értékelés modullal. A szövegek gépi és manuális elemzéseinek találati eredményeit összevetve a számítógépes modul hatékonysága ellenőrizhető és fejleszthető.

Hivatkozások

1. László J: A történetek tudománya. Bevezetés a narratív pszichológiába. Bp.: ÚMK. (2005)
2. Sarbin, T. R. Az elbeszélés mint a lélektan tő-metaforája. In László J. és Thomka B. (Szerk.), Narratívák 5. Narratív pszichológia. Bp.: Kijárat Kiadó. (2001) 59-76
3. Bruner, J.: A gondolkodás két formája. In László J. és Thomka B. (Szerk.), Narratívák 5. Narratív pszichológia. Bp.: Kijárat Kiadó. (2001) 15-27
4. Gergen, K. J. & Gergen, M. M.: Narrative and the self as relationship. In L. Berkowitz (Ed.), Advances in experimental social psychology 21. California: Academic Press. (1988) 17-56
5. McAdams, D. P.: A történet jelentése az irodalomban és az életben. In László J. és Thomka B. (Szerk.), Narratívák 5. Narratív pszichológia. Bp.: Kijárat Kiadó. (2001) 157-175
6. Bigazzi, S. és Nencini, A.: How evaluations construct identities: the psycholinguistic model of evaluation. In Vincze O. és Bigazzi S. (Szerk.), Élmény, történet – a történetek élménye. Tanulmányok László János 60. születésnapjára. Bp.: ÚMK. (2008) 91-105
7. Ehmann B. és Garami V.: Az énbevonódás nyelvi markerei történelmi eseményekről szóló laikus elbeszélésekben. In Vincze O. és Bigazzi S. (Szerk.), Élmény, történet – a történetek élménye. Tanulmányok László János 60. születésnapjára. Bp.: ÚMK. (2008) 41-51
8. Bigazzi S., Csertő I. és Nencini A. A személy- és csoportközi értékelés pszicholingvisztikája. In IV. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2006. dec. 7-8. (2006) 267-277
9. Labov, W. & Waletzky, J.: Narrative analysis: Oral versions of personal experience. In J. Heim (Ed.), Essays on the verbal and visual arts Seattle: American Ethnological Society. (1967) 12-44
10. Labov, W.: The transformation of experience in narrative syntax. In W. Labov, Language in the inner city Oxford: Blackwell. (1972) 354-396
11. Tajfel, H.: Human groups and social categories: Studies in social psychology. Cambridge: Cambridge University Press. (1981)
12 Tajfel, H., & Turner, J. C.: The social identity theory of intergroup behavior. In: S. Worchel & W. Austin (Eds.), The Psychology of Intergroup Relations (2nd ed.). Chicago: Nelson-Hall. (1986)
13 Sherif, M., Harvey, O. J., White, J., Hood, W., & Sherif, C.: Intergroup Conflict and Cooperation: The Robber's Cave Experiment. Norman: University of Oklahoma, Institute of Social Relations. (1961)
14. Sherif, M.: In Common Predicament: Social Psychology of Intergroup Conflict and Cooperation. Boston: Houghton Mifflin. (1966)
15. Tajfel, H. Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations. New York, NY: Academic Press. (1978)
16. Pettigrew, F. T.: The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. Personality and Social Psychology Bulletin, 5(4), (1979) 461-476

17. Maass, A., Salvi, D., Arcuri, L. & Semin, G.: Language use in intergroup contexts: the linguistic intergroup bias. Journal of Personality and Social Psychology, 57(6), (1989) 981-993
18. www.nooj4nlp.net
19. Fülöp É. és László J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével. In IV. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2006. dec. 7-8. (2006) 296-304
20. Semin, G. R. & Fiedler, K.: The linguistic category model, its bases, applications and range. In W. Stroebe & M. Hewstone (Eds.), European Review of Social Psychology 2. Chichester: Wiley. (1991) 1-30

Technológiai fejlesztések a NooJ pszichológiai alkalmazásában

Vincze Orsolya[1], Gábor Kata[2], Ehmann Bea[3], László János[4]

[1] PTE Pszichológia Intézet
orsolyavincze@hotmail.com
[2] MTA Nyelvtudományi Intézet
gkata@nytud.hu
[3] MTA Pszichológia Intézet
ehmann@mtapi.hu
[4]MTA Pszichológia Intézet
laszlo@mtapi.hu

Kivonat: A NooJ nyelvi fejlesztő környezete egy jól kezelhető, dinamikus felületet nyújt az automatizált narratív pszichológiai szövegelemzésben. Az előadás több éves pszichológiai módszertani fejlesztés legújabb eredményeit kívánja bemutatni, különös tekintettel a NooJ nyelvi fejlesztő környezetében kialakított protézisnyelvtanra [1], amely a pszichológiailag releváns kifejezéseket (mentális állapotok, aktív-passzív igék, közelítést-távolítást jelző igék…stb) szemantikai és nyelvtani szerepük alapján összekapcsolja. Ezt megelőzően a nyers szöveg nyelvi elemzését a MorphoLogic Moose szintaktikai elemzőprogramja [2] végzi, ami előkészíti a protézisnyelvtan számára a szövegeket: a szöveget bekezdésekre, mondatokra, tokenekre bontja, elvégzi a szavak morfológiai elemzését, valamint nem csupán beazonosítja az NP és VP csoportokat, de össze is illeszti őket. Kiosztja a nyelvtani szerepeket a főnévi csoportokra és a tematikus szerepeket a vonzatokra. Ez utóbbi esetben a tematikus szerepek kiosztásához a Moose rendszer vonzatkeret-leíró formalizmusát kibővítettük *theta* jeggyel.

1 Bevezetés

A PTE Pszichológia Intézet és az MTA Pszichológiai Intézet kutatóiból álló narratív kutatócsoport hazai és külföldi nyelvtudományi, informatikai és pszichológiai kutatócsoportokkal együttműködve az elmúlt öt évben jelentős nemzetközi áttöréssel járó kutató-fejlesztő munkát végzett. A kutatások eredményeként megszületett és nemzetközi elfogadást nyert a tudományos narratív pszichológia. Az új tudományos paradigma lényege, hogy az emberek természetes közegben zajló, hétköznapi viselkedéséből és kommunikációjából tudományos eszközökkel képes személységükre, lelki állapotaikra és társas beállítódásaikra vonatkozó következtetéseket levonni. Ez úgy történik, hogy a személyes élettörténeti eseményekre, illetve a társadalmi csoportok, például a nemzetek történetére vonatkozó elbeszélések nyelvi és kompozíciós tulajdonságait tudományos eszközökkel megfeleltetjük az identitásképzés pszichológiai

folyamatainak. A nyelvi mintákat nyelvtechnológiai eszközökkel számítógépes programokká fejlesztjük, és ezekkel a programokkal elemezzük a természetes szövegeket. Ez képessé tesz arra, hogy a lelki állapotokról és tartós beállítódásokról diagnosztikus és a társas alkalmazkodás különböző formáit előre jelző eredményeket kapjunk. A tudományos narratív pszichológia fogalmai és eljárásai, amellett, hogy a személyiség és a társas élet pszichológiai folyamatainak komplex megközelítését teszik lehetővé, különösen előnyösnek bizonyultak olyan problémák vizsgálatában, ahol jelen idejű kutatásokra nincs lehetőség, például történeti szövegek esetében, illetve ahol a kérdőíves vagy teszteljárások alkalmazásának lehetősége behatárolt, például addiktológiai betegek esetében. Az alkalmazási lehetőségek köre kiterjed az űrkutatás területére is, mivel a narratív pszichológiai diagnosztikus eljárások alkalmasnak tűnnek a hosszabb űrutazáson részvevő személyek pszichológiai állapotának monitorozására is.

Jelen dolgozat célja, hogy áttekintést nyújtson az automatikus narratív pszichológiai eljárás újabb technikai fejlesztéseiről.

2 Narratív pszichológiai modulok

A kutatócsoportunk által kidolgozott automatikus tartalomelemző eljárás pszichológiailag releváns nyelvi változók köré csoportosuló modulokba rendeződik, mint például az aktivitás-passzivitás [3], érzelem [4], kognitív [5], értékelés [6], intencionalitás [7], idői modulok [8], pszichológiai perspektíva [9].

A pszichológiai modulok több almodulból tevődnek össze, amelyek az elemzés szintjén a pszichológiai jelentés és a technikai kivitelezés tekintetében is különböző komplexitásúak. Ugyanakkor a tartalomelemző algoritmusok működése bizonyos tekintetben azonos: szó- és mondatszintű elemzést végeznek. Ezeken belül azonban eltérések mutatkozhatnak az egyes modulok között a tekintetben, hogy milyen morfológiai vagy szintaktikai megszorításokat alkalmaznak.

2.1 NooJ nyelvi fejlesztő környezet alkalmazása az automatikus pszichológiai tartalomelemzésben

Az egyes modulok automatikus tartalomelemző algoritmusai a NooJ nyelvi fejlesztő környezetében kerültek kidolgozásra [10], ami dinamikus felületet biztosít, lehetővé téve a szoftver biztonságos és rugalmas kezelését nem nyelvészek számára is.

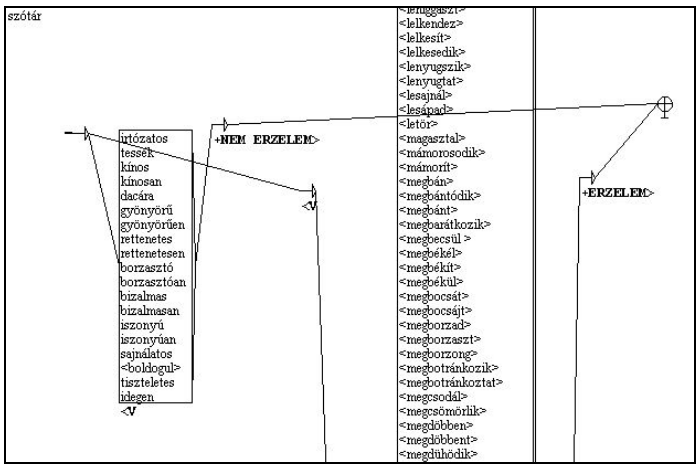
A szoftver központi eleme a szótár, aminek szókincsét egyfelől a magyar írott nyelv általános szókincsét reprezentáló szövegkorpuszokból (Magyar Nemzeti Szövegtár [11], Szeged Korpusz [12]), másfelől specifikus pszichológiai szövegekből álló korpuszból nyertük ki. Ez utóbbiban megtalálhatóak klinikai pszichológiai populációkkal (depressziós, borderline, droghasználó, krízisben lévő betegekkel) készített mélyinterjúk, többgenerációs traumatizált családinterjúk, normál populációkkal (teljesítmény-, veszteség-, párkapcsolati interjúk) felvett féligstruktúrált interjúk, valamint nemzeti és etnikai vonatkozású szövegkorpuszok. Az általános korpuszokból a magyar nyelvben használatos gyakori szóalakok morfoszintaktikailag elemzett formái

kerültek be az általunk használt szótárba, amit a speciális pszichológiai szövegkorpusz gyakran előforduló szavaival egészítettünk ki.

A szoftver motorja véges állapotú technológián alapul, grafikus felülete lehetővé teszi a nyelvtanok gráfként való megjelenítését és szerkesztését. Ezáltal olyan környezetet biztosít, melyben egységesen kezelhetők a nyelvi elemzés különböző szintjei (inflexiós és derivációs morfológia, szintaktikai elemző és transzformációs szabályok). Az automatizált narratív pszichológiai elemzés megközelítésében a gráfokban megjelenő lokális nyelvtanok olyan algoritmusoknak tekinthetők, amelyek pszichológiailag releváns kifejezések beazonosítását végzik.

Ennek megfelelően első lépésben minden modul esetében megtörtént az adott modul tematikájába illeszkedő szavak szótári leválogatása a Magyar Nemzeti Szövegtár leggyakoribb 10 000 igéje, határozói és névutói alapján[1]. Bizonyos modulok esetében további jelentésteli dimenziók is bevezetésre kerültek, mint csoportosító változók: például az érzelmi állapotok „pszichológiai annotációja” [4] során, a valencia mellett, a primer és a társas érzelmek elkülönítése is csoportosító szempontként jelent meg.

A legtöbb modul esetében a gráfok két típusba sorolhatóak: szólistás és szintaktikai gráfok. Mivel az automatizált pszichológiai tartalomelemzés gyakorisági adatokkal dolgozik, a szólistás gráfok készítése értelmes technológiai eljárásnak bizonyul. Ilyenkor a gráfban csupán az adott pszichológiai jelentéskategóriába illeszkedő szavak listája kerül be, minimális szintaktikai megszorítással vagy anélkül (1. ábra).

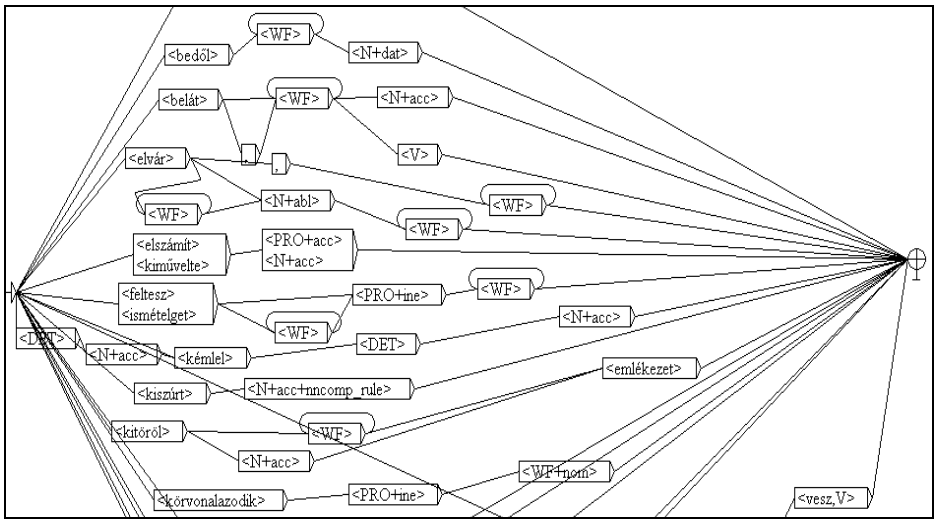


1. ábra. Érzelem modul algráfja.

A szintaktikai gráfok készítése azoknál a kifejezéseknél fordul elő, ahol szintaktikai függőségi viszonyok figyelembevétele szükséges az adott pszichológiai jelentés megragadásához. Például a „bedől” ige csak abban az esetben minősül kognitív kife-

[1] A szótárak fejlesztését a PTE Pszichológia Intézet és az MTA Pszichológiai Kutatóintézet, a szótárak morfológiai annotációját az Szegedi Tudományegyetm és az MTA Nyelvtudományi Intézete végezte.

jezésnek, ha részeshatározós esetű főnév követi <N+dat>. Míg a „belát” igénél az azt követő főnév tárgyas vonzata a feltétel <N+acc> (2. ábra)



2. ábra. Szintaktikai szekvenciára épülő elemző algoritmus.

A lokális nyelvtanok találatait a program konkordanciába rendezi, az adott találat kívánt számú karakterkörnyezetével együtt. Mivel a lokális nyelvtanok lezárása egy szemantikai indexszel történik, a program arra is lehetőséget biztosít, hogy a találatokat szemantikai indexükkel együtt a teljes szövegkörnyezetben lássuk (3. ábra).

```
A nemzet és a császár kibékítésében nagy része volt Erzsébet császárnénak is.
Az ifjú fejedelemasszony <V POZITÍVÉRZELEM">együtt érzett az</V> elnyomott
magyar nemzettel. <V NEGATÍVÉRZELEM"">Fájt</V> neki a nemzet <V
NEGATÍVÉRZELEM"><N>szenvedése</N></V> és <V POZITÍVÉRZELEM""><N>öröme</N> telt
benne, ha javíthatott a sorsán. Állandóan <V "KOGNITIV"><V>érdeklődött</V>
a</V> viszonyai iránt, <V "KOGNITIV">megtanulta</V> nyelvét, történelmét és
szószólója volt császári férje oldalán. A császár <V „KOGNITIV">ismerte e</V>
<N>vonzalmát</N> és felesége kedvéért sokszor gyakorolt kegyelmet, hogy <V
POZITÍVÉRZELEM""><N>örömet szerezzen</N> a császárnénak. De a nemzet is <V
„KOGNITIV">tudta, hogy Erzsébet császárné a pártfogója s azért nagy <V
POZITÍVÉRZELEM""><N>örömmel</N> fogadta a hírt, hogy az országgyűlésre ő is
lejön császári urával.
```

3. ábra. Szemantikai indexek megjelenítése a szövegben.

2 Technikai fejlesztések

A modulok technikai fejlesztését több tényező is lehetővé tette. A Szegedi Tudományegyetemnek köszönhetően az elemzések alapjául szolgáló szótár szemantikai adatbázis információval bővült. Az MTA Nyelvtudományi Intézetben elkészült a nyelvtani, valamint a tematikus szerepek beazonosítására szolgáló lokális nyelvtan, amihez a szövegeinket a MorphoLogic Moose szintaktikai elemzőprogramja [11] készíti elő.

2.1 A szótár szemantikai bővítése

Az alapszótárban a főnevek pszichológiailag releváns szemantikai jegyekkel bővültek. A Szegedi Tudományegyetem által elkészített főnévi adatbázis 20788 főnévi lemmához társít szemantikai információt, melyek különböző szociális kapcsolatokat (rokon, egyéb társadalmi kapcsolat, szűk családi kapcsolat), csoportok jellegét (etnikai, vallási) és egyéb, a tartalomelemzés szempontjai szerint releváns jellemzőket kódolnak (1. táblázat).

1. táblázat: Szemantikai jegyek példája.

szó	Ember	nem	foglalkozás	kapcsolat	csoport	etnikai
betörő	X	xy				
házasságtörő	x	xy	x	x		
jégtörő						
szentségtörő	x	xy				
kitörő						

2.2 Tematikus szerepek beazonosítása

Bármilyen jellegű pszichológiai szövegelemzésben elengedhetetlenül fontos a nyelvtani és a tematikus szerepek beazonosítása. Mivel erre egyenlőre a NooJ szoftver nem képes, egy segédprogram beiktatása vált szükségessé.

A Moose szintaktikai elemzőprogram a nyers szöveg nyelvi elemzése során a szöveget bekezdésekre, mondatokra és tokenekre bontja, elvégzi a szavak morfológiai elemzését, valamint beazonosítja a főnévi (NP) és igei (VP) csoportokat. Az igei csoportok beazonosításánál a program a vonzatkeret-adatbázis segítségével az igéhez sorolható vonzat és szabad határozó NP-ket is beazonosítja.

A tematikus szerepek kiosztásához a MetaMorpho rendszer vonzatkeret-leíró formalizmusát kibővítettük egy új jeggyel (*theta*). A theta jegy a vonzathoz rendelt meghatározott tematikus szerep. Lévén, hogy a pszichológia tartalomelemzésben a tematikus szerepek azonosítása különösen fontos az értelmezés szempontjából, ezért minden modul esetében kiválogattuk a vonzatos igéket és egyszerű példamondatokon keresztül 2640 vonzatkeret-leírást készítettünk, amelyekkel végül kibővült a MetaMorpho rendszer vonzatkeret-leíró formalizmusa. Az automatikus ellenőrzés és

a felmerült hibák javítását tartalmazó validációs ciklus után összesen jelenleg 2322 tematikus szereppel annotált vonzatkeret áll rendelkezésre a rendszerben (2. táblázat).

2. táblázat: Annotált vonzatkeretek tematikus szerepeloszlásai.

Összes vonzatkeret:	2322
Th-jeggyel annotált vonzat összesen:	3174
AG (ágens) jeggyel annotált vonzat:	1447
PAT (páciens) jeggyel annotált vonzat:	749
EXP (experiens) jeggyel annotált vonzat:	646
STI (stimulus) jeggyel annotált vonzat:	270
BEN (beneficiens) jeggyel annotált vonzat:	55
REC (recipiens) jeggyel annotált vonzat:	5
SRC (forrás) jeggyel annotált vonzat:	1
INS (instrumentum) jeggyel annotált vonzat:	1
GOAL (cél) jeggyel annotált vonzat:	0

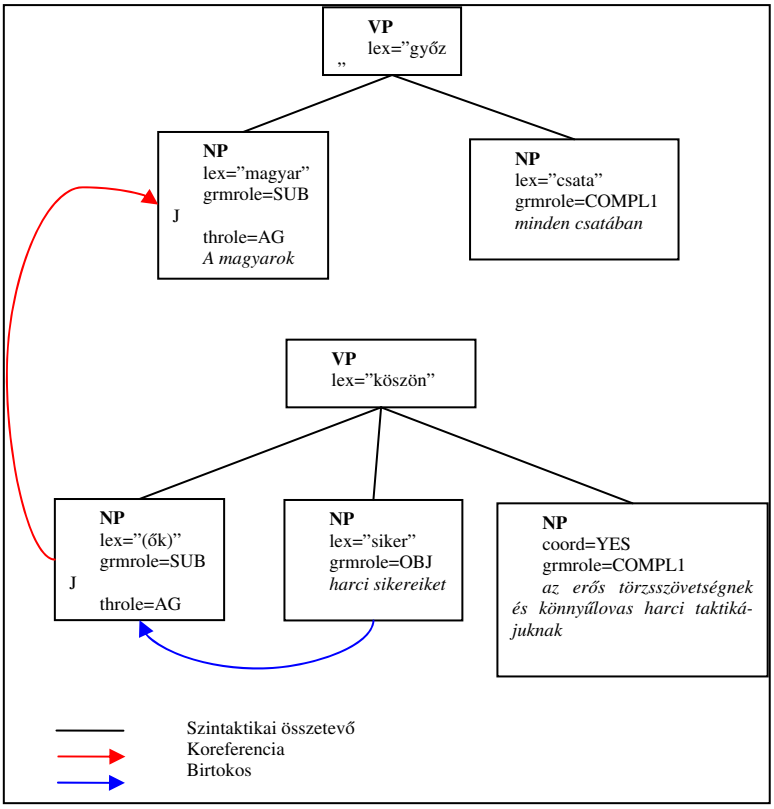
2.3 Szövegbeli utalások feloldása

A szövegekben előforduló utalások természetes jelenségek, ami nem okoz különösebb nehézséget az olvasó számára a szöveg követésében. A tartalomelemzés során az NP-k közötti utalás, azaz amikor a főnévi csoportok egy része nem közvetlenül utal a való világ entitásaira, hanem a szövegben korábban bevezetett ilyen kifejezésre hivatkozik, nem elhanyagolható mennyiségű találati hibát okoz.

A technikai fejlesztések során kétféle, főnévi csoportok közötti utalástípussal foglalkoztunk: a) koreferencia, b) elvált birtokos. Ezek feloldására a Moose szintaktikai elemzőprogram olyan szabályalapú algoritmusokat alkalmaz, amelyek behelyettesítik a hivatkozott kifejezések szótári alakját az utaló kifejezésekbe, ezáltal a NooJ alkalmazásban egyszerű lexikális alakok keresésére nyílik lehetőség.

A Moose szintaktikai elemzőprogram hat különböző NP-koreferencia feloldását végzi el: egyszerű ismétlés, tulajdonnév-variánsok, szinonimák, hipernima, névmási és zérónévmási anafora. Továbbá beazonosítja az összetartozó birtokosoknak és birtokoknak megfelelő kifejezések közötti viszonyokat a szövegben, különös tekintettel azokra az esetekre, ahol a birtokosnak és a birtoknak megfelelő NP-k nem közvetlenül követik egymást.

A nyelvi elemzés során tehát, amit a Moose szintaktikai elemzőprogram végez, megtörténik a nyelvtani és a tematikus szerepek beazonosítása, valamint a hivatkozások feloldása (4. ábra).

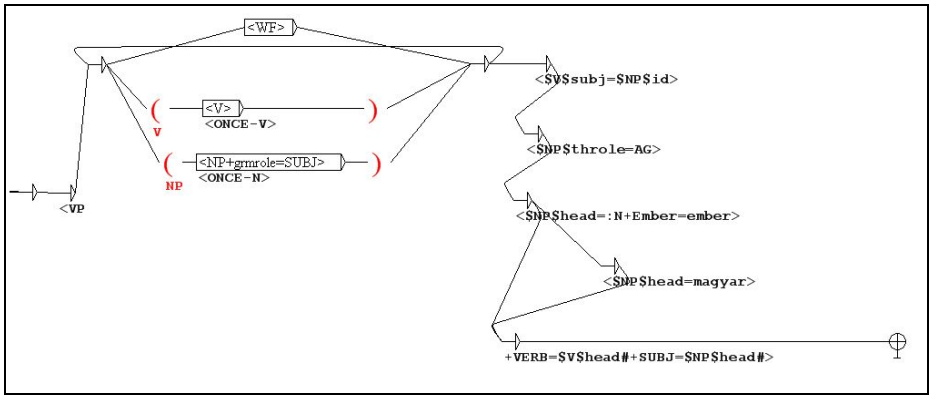


4. ábra. A nyelvi elemzés folyamata.

2.4 Protézisnyelvtan a NooJban

A Moose szintaktikai elemzőprogram által biztosított nyelvtani elemzés a nyers szöveget olyan XML struktúrában jeleníti meg, amiben a dependenciaviszonyokat a szövegszavakhoz társított attribútumok értékei kódolják. Az így előállt szöveg képezi a NooJ bemenetét, ahol a pszichológiai mintázatok beazonosítása történik. Ahhoz, hogy az egyes pszichológiai modulokhoz tartozó korábban kidolgozott lokális nyelvtanok az elemzett mondat szóelemeinek teljes dependenciaviszonyát lefedjék, szükség volt egy ún. *protézisnyelvtan* kidolgozására [1] (5. ábra). A protézisnyelvtan jelentősége, hogy szabad szórendű nyelvekben az összetevők közötti függőségi viszonyok és egyeztetési jelenségek kezelését, illetve a lexikai és a függőségi tulajdonságok szerinti lekérdezést teszi lehetővé. A NooJ-ban ennek technikai hátterét a szoftver új funkciói (a felismert elemek változókban való tárolása, lexikai megszorítások) valósítják meg, melyek így a NooJ-t a véges automatákénál nagyobb leíró kapacitással ruházzák fel.

A protézisnyelvtan lényege, hogy először rekurzívan begyűjti és változókban tárolja a mondat állítmányát és a névszói csoportokat, majd ún. lexikai[2] megkötések segítségével ellenőrzi, hogy ezek rendelkeznek-e bizonyos tulajdonságokkal. A pszichológiai elemzések általános céljával összhangban itt az ige és vonzatai közti szintaktikai és szemantikai viszony beazonosítása történik, azaz a vonzatok grammatikai és tematikus szerepe szerint szűrjük a találatokat.



5. ábra. Protézisnyelvtan.

Az elemzés során a gráf kigyűjti a szöveg mondataiból azokat a találatokat, melyekben az ige alanyi szerepű vonzata ágens tematikus szereppel rendelkezik (5. ábra alapján). Mivel a keresett elemek, vagyis az ige és bővítményei tetszőleges sorrendben követhetik egymást, valamint egyéb elemek is közéjük ékelődhetnek, ezért felismerésükhöz olyan gráfot kell készítenünk, mely egy rekurzív 'hurokban' tartalmazza mind az igét (<V>), mind jelen példában az alanyt (<NP+grmrole=SUBJ>, alanyi szerepű NP), melyek tetszőleges sorrendben követik egymást, és közéjük ékelődve tetszőleges egyéb elemeket (<WF>, word form: tetszőleges szóalak) is megenged. A gráf bal oldali része ezt a hurkot ábrázolja. A tetszőleges szóalakokon (<WF>) kívül a többi felismert elemet piros zárójelekkel jelölt $NP és $V változókban tároljuk, ez teszi lehetővé, hogy a gráf jobb oldalán a lexikai megszorításokban hivatkozhassunk rájuk.

A lexikai megszorítások szerkezete és a rendelkezésre álló jegykészlet
A grammatikai funkció szerinti szűréshez az alábbi jegykészlet használható:

NP+grmrole= COMPL (vonzat), MOD (szabad határozó), OBJ (tárgy), SUBJ (alany), UNKNOWN (egyéb, fel nem ismert)

Nem elég azonban a főnév funkcióját ellenőrizni, külön megszorítással kell megbizonyosodnunk arról is, hogy az adott grammatikai szerepet az adott ige bővítmé-

[2] A 'lexikai' ebben a kontextusban úgy értendő, hogy nem a szövegben, hanem a hozzá tartozó annotációs szerkezetben kódolt információról van szó, ám ez lehet szintaktikai természetű információ is.

nyeként tölti be (vagyis az összetett mondatokban sem keverednek össze a különböző igék bővítménykeretei). Ehhez az XML struktúrában szereplő azonosító (id) attribútumok értéket kell összehasonlítani:

<Vsubj=NPid>	*alany*
<Vobj=NPid>	*tárgy*
<Vcompl1=NPid>	*egyéb bővítmény*

A tematikus szerepek szerinti kereséshez az alábbi jegykészlet áll rendelkezésre:

NP+throle=AG (ágens), PAT (páciens), REC (recipiens), STI (stimulus), EXP (experiens), SRC (forrás), GOAL (cél), INS (eszköz), BEN (beneficiens), UNKNOWN (egyéb, fel nem ismert)

A tematikus szerep annotációját szintén a Moose szintaktikai elemző helyezi el a szövegben, ami az alábbinak megfelelő lekérdezést tesz lehetővé:

<NPthrole=AG>

A találatok tovább szűrhetők lexikai megszorítások hozzáadásával, illetve a pszichológiai modulok kombinálásával. Így például a cselekvő alanyú igék közül kiszűrhetjük azokat, melyeknek alanya egy etnikai csoportot jelölő főnév. Ezeket tovább csoportosíthatjuk az etnikumok szűrésével (pl. magyar cselekvők vs. egyéb népcsoportok). Ennek megfelelően a névszói bővítmény (fejének) szemantikai és/vagy lexikális tulajdonságaira vonatkozó megszorításokat a protézisnyelvtan alábbi csomópontjaiban adhatjuk meg:

szemantikus tulajdonságok:
<NPhead=:N+Ember=ember>
<NPhead=:N+Nem=Y>
<NPhead=:N+etnikai=N>

lexikális tulajdonságok:
<NPhead=magyar>
<NPhead=fejedelem>

2.5 A nyelvtechnológiai változtatások bevezetése a pszichológiai modulokba

Az újonnan alkalmazott Moose szintaktikai elemzőprogram, valamint az erre illeszkedő NooJban kifejlesztett protézisnyelvtan valamennyi, már kifejlesztett pszichológiai modult érintett: szükségesség tette az eddig használt lokális nyelvtanok egy részének átírását. Azokban az esetekben, ahol a pszichológiai modulok lokális nyelvtanai a szólistás algoritmust követik, a protézisnyelvtanban az NP és VP csoportok egyszerű konkretizálással szűkíthetők a pszichológiailag releváns NP és VP csoportokra. Azonban a szintaktikai algoritmust követő lokális nyelvtanokat, amelyek nem

szószintű, hanem szó feletti találatot adnak, nem lehet egy az egyben illeszteni a protézisnyelvtan VP/NP csoportjával. A probléma megoldása különösen lényeges a pszichológiai jelentés megragadása szempontjából, hiszen a találatok nem elhanyagolható részét képezik az ilyen, szintaktikai szekvenciákra épülő jelentések.

Hivatkozások

1. Váradi T, Gábor K.: A magyar Intex fejlesztéséről. In III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004) 3-10
2. Prószéky G., László T., Ugray, G.: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta (2004) 138-142
3. Szalai K., László J.: Az aktivitás-passzivitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
4. Fülöp É., és László J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
5. Vincze O. és László J.: A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
6. Bigazzi S., Csertő I., Nencini, A.: A személy- és csoportközi értekelés pszicholingvisztikája. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
7. Ferenczhalmy R., László J.: Az intencionalitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
8. Ehmann B., Garami V., Szabó J.,: NooJ fejlesztések a szubjektív időélmény tartalomelemzéses vizsgálatára. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
9. Pólya, T., Ferenczhalmy R., Fülöp É., Vincze O.: A pszichológiai perspektíva előfordulása történelem tankönyvi szövegekben V. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2007)
10. Silberstein, M.: NooJ manual. Paris:Université de Franche-Comté (2005)
11. Váradi, T.: The Hungarian National Corpus. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria, (2002) 385-389
12. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005), 409-412

A NooJ alapú narratív pszichológiai tartalomelemzés alkalmazása pszichológiai állapotváltozások monitorozására űranalóg szimulációs kísérletben

Ehmann Bea[1], Balázs László[2], Fülöp Éva[1], Hargitai Rita[3], László János[1,3]

[1] MTA Pszichológiai Kutatóintézet, Szociálpszichológiai Osztály, Budapest
ehmannb@mtapi.hu
[2] MTA Pszichológiai Kutatóintézet, Űrkutató Csoport, Budapest
balazs@cogpsyphy.hu
[1] MTA Pszichológiai Kutatóintézet, Szociálpszichológiai Osztály, Budapest
fulop81@gmail.com
[3] PTE Pszichológiai Intézet, Pécs
hargitairita@freemail.hu
[1] MTA Pszichológiai Kutatóintézet, Szociálpszichológiai Osztály, Budapest
[3] PTE Pszichológiai Intézet, Pécs
laszlo@mtapi.hu

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

A szerző nem járult hozzá a cikk elekronikus formában történő közzétételéhez.

Versenyképességi kulturális orientációk azonosítása vezetői narrációkból

Mikulás Gábor

GM Consulting
mikulasg@gmconsulting.hu

1 Bevezetés

Az információs szolgáltatásokban hitelességi szempont, hogy a döntéshozatali folyamatot több forrásból, illetve szempont szerint összeállított információs csomag támogassa, hiszen így javul a döntés megalapozottsága. Források lehetnek az ügyben érintettek korábbi sajtómegjelenései, üzleti tranzakciós adatai, vonatkozó céginformációk, továbbá az illetékesek nyilatkozatai is, melyeket például telefonos vagy személyes kapcsolatfelvétel útján lehet „beszerezni". Ezek gyakran a legértékesebbek is, hiszen olyan adatok merülhetnek fel, melyek mások által is elérhető, nyomtatott formában (még) nem léteznek. További jellemzője ezeknek az információknak, hogy „gyenge jelek" [1], azaz a változások előjeleit a versenykörnyezetben nem „harsányan" (pl. értékesítési diagram), hanem szerényen (pl. munkatársak nyilatkozatai) jelzik. Ezek a narrációk (szöveges közlések) – amennyiben originálisak, azaz pr-es szakember, újságíró vagy más közvetítő által nem „manipuláltak" – nemcsak a közölt tényadatok, hanem azon túlmutató, szélesebb közönség által kevéssé azonosítható szövegmélyi információt is tartalmaznak, melyek szövegelemzés segítségével azonosíthatók, értékes, unikális, és gyakran prediktív információt kínálva a megrendelőnek. Az információszolgáltatás tehát szélesedhet, értékesebbé válhat általuk [2]. Ezekhez az információkhoz nyújt segítséget például a kultúrakutatás eszközrendszere.

A Sapir—Whorf-hipotézisből [3] kiindulva – mely a tudat és a nyelv közötti kapcsolat meghatározó voltát emeli ki – természetesnek tűnik a narráció tudatos vagy tudat alatti identitásmegerősítő funkciója. „... a jellegzetes nyelvi elemek ismételt kimondása megerősíti, és újra létrehozza a csoport értékeit, valamint az egyén státusát és szerepeit. Ezeknek az eszközöknek a segítségével fenntartják a csoport belső koherenciáját, és világosan meghatározzák annak határait (a kívülállók nem használják a jellegzetes formákat)." [4] Ezért nem is meglepő, hogy narratív pszichológia elméleti keretében az emberi tapasztalat – benne a kulturális orientáció – narratív formákba szerveződik. Egy szociális reprezentáció – például szervezeti kulturális konstrum – megosztása és megvitatása így tartalmának narratív szerveződésén keresztül történik.

2 Kultúrakutatás és versenyképesség

„A [szervezeti] kultúra a közösség tagjainak közös tapasztalatokból származó és generációkon keresztül átöröklődő, a közösség valamennyi tagja által osztott motivációinak, értékeinek, meggyőződéseinek, identitásainak és a lényeges események kö-

zös értelmezéseinek vagy jelentéseinek összessége." [5] Az utóbbi harminc évben többféle iskola kidolgozta módszertanát a kulturális orientációk feltérképezésére. A nemzetközi GLOBE-projekt (Global Leadership and Organizational Behavior Effectiveness)[1] Szervezetek középvezetőinek kérdőíves felmérése által határoz meg különböző szinteket kilenc kulturális dimenzióban [6], nemzetek és szervezetek szintjén egyaránt. Ezeket a dimenziókat – kulturális orientációkat – az 1. táblázat tartalmazza.

1. táblázat: Kulturális orientációk a GLOBE-kutatásban.

hatalmi távolság	„Annak mértéke, hogy egy szervezet vagy a társadalom tagjai mennyire várják el és fogadják el a hatalom egyenlőtlen eloszlását, hogy a hatalom a szervezet vagy kormány magasabb szintjére rétegződjön, és oda koncentrálódjon."
bizonytalanságkerülés	„Annak a mértéke, hogy egy szervezet vagy a társadalom tagjai kialakult társas normákra, rituálékra, és bürokratikus gyakorlatra támaszkodva mennyire törekszenek a bizonytalanság elkerülésére, mérsékelve ezzel a jövőbeli események előrejelezhetetlenségét."
intézményi kollektivizmus	„Annak a mértéke, hogy a szervezetek és társadalom intézményi normái és gyakorlata mennyire bátorítják és jutalmazzák az erőforrások kollektív elosztását és a kollektív cselekvést."
csoportkollektivizmus	„Annak a mértéke, hogy az egyének szervezetükben vagy családjukban mennyire juttatják kifejezésre büszkeségüket, lojalitásukat és összetartozás-érzésüket."
nemi egyenlőség	„Annak a mértéke, hogy a társadalom vagy egy szervezet mennyire minimalizálja a nemi szerepek közti különbségeket elősegítve ezzel a nemek közötti egyenlőséget."
rámenősség / asszertivitás	„Annak mértéke, hogy az egyének társas kapcsolataikban mennyire határozottak (asszertívek), szembenállóak (konfrontatívak) és agresszívek szervezeteikben vagy a társadalomban."
teljesítményorientáció	„Annak a mértéke, hogy egy szervezet vagy a társadalom mennyire bátorítja a csoporttagokat a teljesítmény növelésére és a kiválóságra, és mennyire jutalmazza őket ezért."
jövőorientáció	„Annak mértéke, hogy egy szervezet vagy a társadalom tagjai milyen mértékben adják a fejüket olyan magatartásformákra, mint a tervezés, a jövőbe való befektetés, a javak egyéni vagy kollektív felélésének elhalasztása."
humánorientáció	„Annak a mértéke, hogy a szervezetek vagy a társadalom tagjai mennyire bátorítanak és jutalmaznak másokat arra, hogy igazságosak, méltányosak, önzetlenek, barátságosak, nagylelkűek, gondoskodók és kedvesek legyenek."

A módszertan kérdőíves felmérésében külön rákérdez a válaszadó által tapasztalt (leíró), illetve a szerinte kívánatos (normatív) állapotra, mindezeket országos és szervezeti szintre vonatkoztatva egyaránt.

[1] A projekt weboldala: http://www.thunderbird.edu/sites/globe/

A kilenc orientáció közül Bakacsi [7] kutatásában hatot jelölt meg, melyek prediktív módon határozzák meg adott országok versenyképességét. (Versenyképesség: a World Competitive Yearbook definíciója szerint: „A versenyképesség elemzi, hogy a nemzetek és a vállalatok hogyan menedzselik kompetenciáik összességét annak érdekében, hogy jólétet és profitot érjenek el.” [8]) Bakacsi kutatása szerint

- pozitívan korrelál a várható versenyképességgel a bizonytalanságkerülés, az intézményi kollektivizmus, a teljesítményorientáció és a jövőorientáció leíró, társadalmi szintű értékeivel
- negatívan korrelálva jelzi előre a versenyképességet hatalmi távolsági index és a csoportkollektivizmus (büszkeség a saját csoportra) leíró, társadalmi szintű értékeivel.

Ez a tapasztalat alapvetően összecseng más módszertanok – pl. Hofstede, Trompenaars és Hampden-Turner felméréseivel is. (Megjegyzendő, hogy a versenyképességnek mindig lehetnek a benchmarktól eltérő útjai is.)

A kutatásban a vonatkozó irodalmak alapján feltételeztük, hogy a vezető és a szervezet kultúrája hosszabb távon azonos, vagy erősen megközelítik egymást. Feltételeztük továbbá azt is, hogy országos vagy társadalmi szinten azonosak a versenyképességi orientációk, mint szervezeti szinten.

A kutatás egyrészt arra kereste a választ, hogy a versenyképesség orientációi mennyiben mutathatók ki felsővezetői narrációk tartalomelemzésén keresztül. A másik cél: a vizsgált szervezetek versenyképességi szempontú jellemzése a kapott eredmények tükrében.

3 Módszertan

3.1 GLOBE-kérdőív

A kiválasztott négy, szolgáltatásaiban az információfeldolgozás valamely formáját végző szervezet közül kettő magánvállalkozás (magyar tulajdonú regionális tanácsadócég illetve alapítványi tulajdonú kiadóvállalat) illetve egy állami nagyszervezet, valamint annak regionális egysége volt. A GLOBE-kérdőív a középvezetők töltötték ki. A kérdőívek feldolgozását a hazai GLOBE-központ munkatársai végezték.

3.2 Interjú készítése, a szöveg feldolgozása

Az interjúk a csúcsvezetőkkel készültek. Feltétel volt, hogy a csúcsvezető legalább öt évet töltsön el az adott szervezet kultúrájában. A félig strukturált interjúban a csúcsvezetők négy kérdésre válaszolva a cégtörténetről, saját karrierjükről, a cégtervekről és egy-egy személyes sikerről, kihívásról beszéltek, mely fejenként 3000-től 4000 szóig terjedő korpuszt eredményezett.

A kapott korpuszelemzésre történő előkészítése a tagolást jelentette, melynek alapja: egy gondolat – egy egység. Azon belül, ha ugyanannak a gondolatnak más szempontja kerül előtérbe, az újabb egység. Szintén új egység, ha a gondolaton belül más szemszögébe helyezkedik a beszélő. A tagolás befolyásolhatja az orientációs változók számát. Időnként előfordult, hogy a közbeékelés esetén a közbeékelt szöveget a feldolgozás során a befogadó szöveg utánra helyeztem (így a befogadó szöveg orientációja nem kétszer, hanem egyszer jelöltetett). E módszer természetesen nem minden esetben adhat egyértelmű tagolást.

Bár a tartalomelemzésnek, diskurzuskutatásnak kiterjedt irodalma van, az irodalomkutatás során nem találtam kultúrakutatási célra kész hazai, illetve külföldi módszertant, emiatt és a korlátozott kapacitás miatt a kutatás – a módszertani kísérletezést vállalva – négy szolgáltató szervezetre korlátozódott. A saját módszertani alkalmazás első lépése az irodalomkutatás során talált – a kultúrakutatás szempontjából – részmegoldások [8] áttekintése volt. E tapasztalatok hasznosnak bizonyultak a szövegfeldolgozás utáni kiegészítő szövegjellemzés során.

A korpusz feldolgozása két módszerrel történt: tartalomelemzéssel és motivációkutatással.

3.3 Tartalomelemzés

Kérdése: mekkora a Bakacsi által meghatározott versenyképességi orientációk szerinti tartalmi motívumok számaránya a szövegekben. Az interjúszövegek leírt változatait a tartalomelemzés során egy e célra készített néhány oldalas instrukció alapján két független kódoló kódolta, a vitás eseteket egy harmadik személy bírálta felül. Feldolgozásra olyan kódok kerültek, melyeket két kódoló egybehangzóan vett találatnak, azaz a hat GLOBE-orientáció legalább egyikének pozitív vagy negatív narratív kódjaként.

3.4 Motivációkutatás

Kapitány Ágnes és Kapitány Gábor által felvázolt motivációrendszer és a versenyképességi orientációk megfeleléseinek számaránya a szövegekben. A motivációelemzéshez a beszéd szerkesztési sajátosságait kell megfigyelni és jelenlétükből, azok mértékéből kapcsolódásaiból következtethetünk arra, hogy a beszélőt milyen hajtóerők mozgatják [10]. A motivációk területei:

kapcsolatteremtés
környezeti hatások
ismeretek rendezése
tekintély- és mintakövetés
feladatvégzés szükséglete
az „erkölcs” szükséglete
birtoklás
a „dominancia” szükséglete
szabadság, személyre szabott életmód
életcéligény.

A szerzőpáros minden motivációt három részre bont: belső késztetések, célok és megfeleléskészségek. Megállapításuk szerint akkor van bennünk viszonylagos harmónia, ha a három motivációfajta egyensúlyban, nagyjából egyenlő arányban van jelen személyiségünkben [11]. A felsorolt motivációk nem feleltethetők meg egyértelműen a kulturális orientációknak, viszont megállapíthatók szorosabb rokonvonások közöttük. Pl.: a dominancia szükséglete a hatalmi távolsági indexszel.

A kódolást a motivációkutatással ellentétben egy kódoló végezte, ezért a keletkezett adatok csak tájékoztató jellegűek.

3.5 Az adatok feldolgozása; trianguláció

A kutatás tehát háromfelől közelített a versenyképességi orientációkhoz. Mindezt kiegészítette a versenyképességi orientációkkal kapcsolatos irodalomkutatás a pszichológiai, szociológiai és vezetéstudományi irodalomban.

A tartalomelemzés, a motivációkutatás és a GLOBE-eredmények közötti korrelációkat az SPSS elemzőszoftver mutatta ki. A korrelációs tábla input-adatai:

1. GLOBE: normatív szervezeti, normatív országos, leíró szervezeti, leíró országos dimenziókban a hat versenyképességi orientáció (hatalmi távolság, bizonytalanságkerülés, csoportkollektivizmus, intézményi kollektivizmus, teljesítményorientáció, jövőorientáció)
2. Tartalomelemzés: a fentebb leírt hat versenyképességi orientáció megjelenése a cégtörténet, életút, cégtervek és a siker, kihívás narratív egységekben
3. Motivációkutatás: a fentebb leírt hat versenyképességi orientáció megjelenése a cégtörténet, életút, cégtervek és a siker, kihívás narratív egységekben.

A korrelációvizsgálat során a GLOBE—tartalomelemzés, GLOBE—motivációkutatás és a tartalomelemzés—motivációkutatás kerültek sorra.

A vizsgált szervezetek kis száma miatt a kapott eredmények óvatosan kezelendők, ugyanakkor több, szakirodalmakból már ismert mintázat is kirajzolódott az eredményekből. Ilyen például az ingajelenség [10], mely az országos szinten a normatív és a leíró értékek közötti különbséget jelzi, azaz az emberek által követendőnek tartott illetve ténylegesen követett gyakorlat közötti űrre utal.

4 Eredmények

A vizsgált hat versenyképességi orientációból háromban sikerült következtetések levonására alkalmas mintákat találni. Ezek az orientációk: hatalmi távolság, csoportkollektivizmus és intézményi kollektivizmus.

4.1 Hatalmi távolság

2. táblázat: A hatalmi távolság korrelációi (a negatív korrelációk szürke háttérrel kiemelve).

	Tartalomelemzés				Motivációkutatás			
	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.
Cégtörténet	**0,744**	**0,738**	**0,860**	**-0,826**	0,499	0,684	**0,822**	**-0,809**
Életút	**0,773**	**0,795**	**0,803**	**-0,859**	-0,038	**0,752**	0,333	**-0,832**
Cégtervek	**0,786**	0,633	0,272	-0,545	0,644	0,424	0,010	-0,293
Siker, kihívás	**0,866**	-0,034	0,461	0,061	0,228	0,782	0,574	**-0,882**

A mintákat elemezve, jelzéseiket összegezve azt feltételezhetjük, hogy a hatalmi távolságra utaló narratív jelek esetén a nyilatkozó

- hatalmi távolságot tart kívánatosnak szervezeti és országos szinten
- szervezeti szinten hatalmi távolságot tapasztal
- országos szinten nem tapasztal hatalmi távolságot.

Ez az eredmény – a Bakacsi-féle kutatásban meghatározott versenyképességi mintázattal összevetve – magasabb szintű versenyképességre utal.

4.2 Csoportkollektivizmus

3. táblázat: A csoportkollektivizmus korrelációi.

	Tartalomelemzés				Motivációkutatás			
	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.
Cégtörténet	0,365	**0,736**	0,604	**-0,835**	0,583	0,225	**0,722**	-0,295
Életút	-0,206	0,174	-0,215	-0,188	**-0,698**	0,461	-0,580	-0,380
Cégtervek	-0,028	**-,972(*)**	-0,387	**,993(**)**	0,519	-0,512	0,390	0,468
Siker, kihívás	0,162	0,565	0,312	-0,643	0,175	0,003	0,105	-0,081

A jövőre vonatkozó cégtervek esetén valószínűsíthető az ingajelenség (a normatív országos és a leíró országos érték ellentétes irányban „leng ki"). A múltra vonatkozó interjútémák esetében ez csak alacsony szintű korrelációk mintázatával mutatkozik meg.

A mintákat elemezve, jelzéseiket összegezve azt feltételezhetjük, hogy a (cég)tervekben csoportkollektivizmusra utaló narratív jelek használója országos szinten magas csoportkollektivizmust észlel (ld. leíró országos korreláció), de alacsony értéket tart kívánatosnak (ld. normatív országos korreláció).

Ez az eredmény – a Bakacsi-féle kutatásban meghatározott versenyképességi mintázattal összevetve – alacsonyabb szintű versenyképességre utal.

4.3 Intézményi kollektivizmus

4. táblázat: Az intézményi kollektivizmus korrelációi (a negatív korrelációk szürke háttérrel kiemelve).

	Tartalomelemzés				Motivációkutatás			
	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.	Norm. szerv.	Norm. orsz.	Leíró szerv.	Leíró orsz.
Cégtörténet	**-,970(*)**	**-0,941**	**-0,945**	0,561	-0,551	-0,467	-0,079	0,560
Életút	-0,594	-0,653	**-0,864**	0,271	**-,966(*)**	**-0,939**	-0,674	**0,792**
Cégtervek	**0,943**	**0,855**	**0,795**	-0,518	**0,741**	**0,859**	0,477	**-,951(*)**
Siker, kihívás	-0,512	-0,542	**-0,866**	0,083	-0,605	**-0,710**	**-0,764**	0,475

A mintákat elemezve, jelzéseiket összegezve az alábbi mintázatot feltételezhetünk:

5. táblázat: Az intézményi kollektivizmus narratív és kulturális korrelációja.

	szervezetben a nyilatkozó…	országosan a nyilatkozó…
A múltra vonatkozó narrációban az intézményi kollektivizmus markerei	alacsony intézményi kollektivizmust észlel és kíván	alacsony intézményi kollektivizmust kíván, miközben magasat észlel
A jövőre vonatkozó narrációban az intézményi kollektivizmus markerei	magas intézményi kollektivizmust észlel és kíván	alacsony intézmény kollektivizmust észlel, és magasat kíván

Ez az eredmény – a Bakacsi-féle kutatásban meghatározott versenyképességi mintázattal összevetve – magasabb szintű versenyképességre utal.

4.4 A tartalomelemzés és a motivációkutatás alkalmazott eljárásainak kontrollja

A 6. táblázat azt mutatja, hogy a tartalomelemzés és motivációkutatás módszereivel mely és kérdéscsoportokban sikerült az orientációkat egymással szignifikáns módon kimutatni.

6. táblázat: A tartalomelemzés és a motivációkutatás közötti korrelációk.

	hatalmi távolság	bizonyta-lanságke rülés	csoport-kollekti-vizmus	intézm. kollekti-vizmus	teljesít-ményorien táció	jövő-orien-táció	*Σ*
Cégtörténet (múlt)	**0,942**	-0,159	0,190	0,370	0,558	0,166	*2,067*
Életút (múlt)	0,604	-0,191	**0,805**	0,378	-0,368	0,050	*1,278*
Cégtervek (jövő)	**0,948**	0,053	0,359	0,491	-0,172	-0,249	*1,430*
Siker, kihívás (múlt)	-0,288	0,126	**0,814**	**0,887**	-0,672	-0,306	*0,560*
Σ	*2,206*	*-0,172*	*2,169*	*2,125*	*-0,654*	*-0,338*	*2,067*

A két módszertan közötti korreláció feltárása a hatalmi távolság, valamint a csoport- és az intézményi kollektivizmus esetében sikerült leginkább, legkevésbé a teljesítményorientációban. A kérdéscsoportokat tekintve leginkább a cégtörténetben, legkevésbé a szabadabb tartalmú siker, kihívás témájában sikerült az orientációkat egyöntetűen azonosítani.

5 Összefoglalás

A kutatás eddigi eredményei megmutatták, hogy a tartalomelemzés és a motivációkutatás egyaránt alkalmas versenyképességi kulturális orientációk azonosítására. A módszertanok egymástól függetlenül és együtt is használhatók. Ez utóbbi esetben van mód a módszerek és a kapott eredmények kontrolljára. Lehetőség kínálkozik továbbá más, nyelvészeti, szociálpszichológiai, közgazdasági és menedzsmentkutatások eredményeinek bevonására is, melyek megerősíthetik az e közleményben is vázolt eredményeket. Mindez a folyamatban lévő projektben meg is történik. E különböző módszertanok egybevágó eredményei felhasználhatók az egyes vizsgált intézmények versenyképességének prediktív szempontú jellemzésére is.

E vezetéstudományi PhD-kutatás eredményeinek értékelése még folyamatban van. A projekt honlapja: http://www.gmconsulting.hu/inf/cikkek/312/index.php

Hivatkozások

1. Zanassi, A.: Text mining: the new competitive intelligence frontier : real cases in industrial, banking and telecom / SMEs world. VSST2001 Conference Proceedings, Barcelona, Oct. 17, (2001); G. S. Day, H. J. Schoemaker, P.: Tartsuk szemmel a perifériát. Harvard Businessmanager (2006. május) 74-85
2. vö.: Wormell, I.: Adding values to the retrieved information. FID Review 1 (1999) 4/5 83-90
3. bővebben pl.: Róka J.: Kulturális változatok a nyelvhasználatban. In: Róka J., Hochel, S. (szerk.): Interkulturális és nemzetközi kommunikáció a globalizálódó világban. Budapesti Kommunikációs és Üzleti Főiskola, Budapest (2009) 147-50
4. Flower, R.: Hatalom. In: Síklaki I. (szerk.): Szóbeli befolyásolás, II. Nyelv és szituáció. Typotex, Budapest (2008) 236
5. House et al.: Culture, leadership, and Organizations. The GLOBE study of 62 societies (Vol. 1.) Sage, Thousand Oaks, CA (2004) 15

6. House, R,, Javidan, M., Hanges, P., Dorfman, P.: Understanding cultures and implicit leadership theories across the globe: an introduction to project GLOBE. Journal of World Business 37 (2002) 3-10
7. Bakacsi Gy.: Kultúra és gazda(g)ság – A gazdasági fejlődés és fejlettség és a GLOBE kultúraváltozóinak összefüggései. Vezetéstudomány 38 (2007) Különszám 35-45
8. Garelli, S.: Competitiveness of nations: The fundamentals: World competitiveness Yearbook. IMD, Lausanne (2005)
9. pl.: Ehmann B.: A szöveg mélyén: a pszichológiai tartalomelemzés. Új Mandátum, Budapest (2002), László J.: Történetek tudománya : bevezetés a narratológiába. Új Mandátum, Budapest (2005), p. 239, Flower R.: Hatalom. In: Síklaki I. (szerk.) Szóbeli befolyásolás. II. Nyelv és szituáció. Typotex, Budapest (2008), Pennebaker, J. W.: What our words can say about us: Toward a broader language psychology. Psychological Science Agenda 15 (2002) 8-9.
http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/PsychSciAgenda.pdf
10. Kapitány Á., Kapitány G.: Hogyan beszélnek vágyaink és törekvéseink. Szorobán, Budapest (1993)
11. Bakacsi Gy.: The Pendulum Effect: Culture, Transition, Learning. In: Makó Cs., Warhurst, Ch. (eds.): The management and organisation of firm in the global context. Institute of Management Education, University of Gödöllő, Department of Management and Organisation, Budapest University of Economic Sciences, Budapest (1999) 111-118

VI. Gépi tanulás

Gépi tanulási módszerek ómagyar kori szövegek normalizálására

Oravecz Csaba, Sass Bálint, Simon Eszter

MTA Nyelvtudományi Intézet
e-mail:{oravecz,sass.balint,eszter}@nytud.hu

Kivonat A nyelvemlékek számítógéppel segített feldolgozása és elemzése számos problémát felvet, a nyelvtörténeti kérdésektől az egészen konkrét technológiai nehézségekig. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási „forgatókönyv" egyik gyakori közös átalakító lépése a szokásos betűhű átírásban kiadott szövegek mai modern helyesírású változatának előállítása. Ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, ezért az azokban sikerrel alkalmazott zajos csatorna modellt adaptáljuk és vizsgáljuk ennek eredményességét a transzliterációs feladatban.

Kulcsszavak: gépi tanulás, zajos csatorna modell, nyelvtörténet, normalizálás, transzliteráció

1. Bevezetés

A Nyelvtudományi Intézetben április óta folyik egy projekt, melynek a célja egy elektronikus nyelvtörténeti adatbázis létrehozása. Az adatbázis tartalmazza az összes ómagyar szövegemléket, a középmagyar korból pedig különféle szempontok szerinti arányosan válogatást úgy, hogy minden nyelvjárás, műfaj, regiszter súlyának megfelelően legyen képviselve benne. Ehhez első lépésben össze kell gyűjteni az összes elektronikus formában elérhető szöveget, majd egységes formátumra hozni őket. A szövegemlékek eredeti, betűhű változatukban és egy ún. *normalizált változatban* is elérhetők, kereshetők lesznek. Ez a normalizálási lépés a szövegfeldolgozási munkafolyamatnak az a lépése, amikor az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. [14]). A folyamat számítógépes modellezésének célja az, hogy választ kapjunk arra a nagyon fontos gyakorlati kérdésre, hogy a rendkívül időigényes manuális átírási munka kiváltható-e gépi eljárással, így a szükséges emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára. Mivel ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, így feltétlen érdemesnek tűnik az azokban sikerrel alkalmazott módszerek adaptálása és eredményességének vizsgálata.

A dolgozat központi kérdése annak meghatározása, hogy az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe, és melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását eredményezi. Ennek érdekében szükség van az adott modellben használt jegyeket tartalmazó specifikusan annotált tanító szövegekre, melyekből jelenleg korlátozott mennyiség áll a rendelkezésünkre — lévén a normalizálás nyelvtörténeti szakértelmet kívánó, időigényes munka. További nehézséget jelent, hogy az egyes nyelvemlékek írásmódja, a bennük előforduló speciális ómagyar karakterek halmaza is meglehetősen különbözik egymástól. A „könyvméretű magyar írásosságot" a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. A 14-16. században a helyesírás még egyáltalán nem volt egységesítve, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. Ezért nehéz egyértelmű konverziós szabályokat meghatározni, valamint emiatt kritikus kérdés az, hogy a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre. Mindezek miatt célszerű a problémát valamilyen valószínűségi alapú paradigma keretei között vizsgálni, egyik legkézenfekvőbb erre Shannon zajos csatorna modellje [16].

Esetünkben a normalizálás tulajdonképpen egybeesik azzal a fogalommal, amit a nyelvtörténészek értelmezésnek hívnak. Az értelmezés hagyományosan a régi nyelvi adatoknak mai magyar nyelvre való „fordítását" jelenti. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még annyira sem várható el. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. ÓMS: Vylag uilaga [világ világa]), vagy kettős hangértéke van egy-egy betűnek (pl. MK: zerzete zerent [szerzete szerint]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is (pl. az u,v,w több évszázadon át jelölhette az u,ú,ü,ű,v hangok bármelyikét).

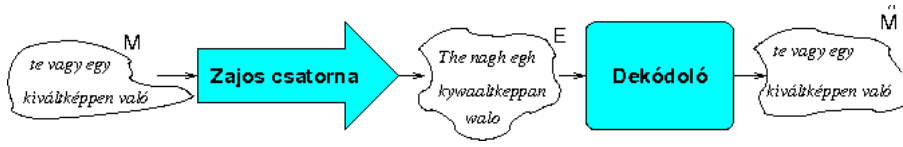
A dolgozat a következőképpen épül fel. A 2. rész rövid leírást ad az eddigi rokonítható kezdeményezésekről. A 3. rész az eljárás elméleti alapjait tárgyalja, míg a 4. részben a modell tanításának folyamatát mutatjuk be. Az 5. rész a modell alkalmazásáról és a lehetséges kiértékelési módszerről ad leírást. Rövid összefoglalás zárja a dolgozatot a 6. részben.

2. Kitekintés

A kitűzött feladat egyrészt lényegében tekinthető két reprezentáció közötti fordítási feladatnak, így közvetlenül rokonítható azokkal a megközelítésekkel, ahol a szövegnormalizáláshoz komplex gépi fordítási modelleket használnak [15,11,1]. További kapcsolódó problémakör a graféma-fonéma konverzió, ahol [12] korai valószínűségi modelljére támaszkodik a legtöbb megoldási javaslat. [6] tartalmaz

részletes összehasonlítást, ahol kimutatja, hogy a gépi tanulási módszereket használó modellek jobb eredményeket adnak, mint a kézzel írt szabályokon alapulók. Számos analógiás, továbbá rejtett Markov-modellen alapuló eljárást is eredményesen alkalmaztak [2,17]. Az általunk használt módszer előzménye [9] helyesírás-ellenőrzésre kidolgozott eljárása, illetve ennek továbbfejlesztett csatorna-modellt alkalmazó változatai [3,18].[1] A következő fejezet ezt modellt ismerteti részletesen. A fentiektől eltérő paradigmájú, szabály alapú megközelítésre példa [10].

3. Zajos csatorna alapú szövegnormalizáló modell



1. ábra. Szövegnormalizálás zajos csatorna modellben.

Az 1. ábrán látható modellben az eredeti szöveget úgy tekintjük, mint a normalizált változat egy zajos kommunikációs csatornán átment „eltorzított" változatát. Jelölje M a modern helyesírású normalizált szövegváltozat pl. egy (rész)mondatnyi sztringjét, E pedig ennek eredeti betűhű átiratát. A dekódoló feladata annak az M karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális,

$$\hat{M} = \operatorname*{argmax}_M P(M|E) \tag{1}$$

illetve a szokványos átalakítással:

$$\hat{M} = \operatorname*{argmax}_M \frac{P(E|M)P(M)}{P(E)} = \operatorname*{argmax}_M P(E|M)P(M) \tag{2}$$

A feladat tehát egyrészt a $P(E|M)$ transzliterációs modell-eloszlás (csatornamodell) és a $P(M)$ normalizált szövegmodell-eloszlás (forrásmodell) meghatározása.

Forrásmodellként a normalizált szövegből készült karakter N-gram modelleket használhatunk, ahol vizsgálható a módszer pontossága N függvényében. Mivel a normalizált szöveg alapvetően mai magyar nyelvű anyag, a forrásmodell felépítésében nagy mennyiségű adat hozzáférhető és használható, így N a szómodelleknél megszokott 3-nál nagyobb is lehet. A transzliterációs modell paramétereinek meghatározására többféle lehetőség kínálkozik, melyeknek előfeltétele olyan tanító korpusz, amely $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmaz.[2] Az

[1] Természetesen számos további gépi tanulási paradigma is alkalmazható a feladat megoldására, a döntési fáktól a log-lineáris osztályozókig.

[2] $i < j,\ k < l$ karakterek közötti pozíciókat jelölő indexek, $j = i+1,\ l = k+1$ esetben karakter→karakter megfeleltetést kapunk.

1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni. A modell paramétereit a tanító korpuszból becsüljük, míg a lehetséges modern szövegváltozatok halmazát a megfeleltetésekből generáljuk. Az alkalmazott eljárás hasonló [3] gépelési hibákat javító módszeréhez, melynek alapján a transzliterációs modell formálisan az alábbi módon írható le.

Legyen $\mathrm{Part}(M)$ a modern nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza (hasonlóan $\mathrm{Part}(T)$ az eredeti alakra). Egy adott $R \in \mathrm{Part}(M)$ partícióra, ahol R $|R| = j$ darab szegmentumból áll, legyen R_i az i-edik szegmentum. Ekkor ($|T| = |R|$ esetén, ahol $T \in \mathrm{Part}(E)$)

$$P(E|M) = \sum_{R \in \mathrm{Part}(M)} P(R|M) \sum_{T \in \mathrm{Part}(E)} \prod_{i=1}^{j} P(T_i|R_i) \tag{3}$$

Egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának. Csupán a legjobb particionálást tekintve (3) az alábbira egyszerűsödik:

$$P(E|M) = \max_{R \in \mathrm{Part}(M), T \in \mathrm{Part}(E)} P(R|M) \prod_{i=1}^{j} P(T_i|R_i) \tag{4}$$

[3] modelljéhez hasonlóan $P(R|M)$ meghatározásával egyelőre mi sem foglalkozunk, vagyis ezt a tényezőt nem vesszük figyelembe (illetve a partíciók felett jobb híján jelenleg egyenletes eloszlást feltételezünk).

4. A modell tanítása

4.1. A transzliterációs modell tanító korpuszának előállítása

A tanító korpusz két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából állt elő. A Müncheni emlék [7] a 16. század elejéről származó, sajátos nyelvemlékünk. Sajátossága abban rejlik, hogy egyszerre tartalmaz egyházi és világi szövegeket, valamint latin és német nyelvű részleteket is (ezeket a normalizálás és a tanító korpusz építése során kihagytuk). A Szabács viadala [8] a 15. század második felében keletkezett, eredeti magyar nyelvű vers. A legrégebbi ránk maradt históriás ének, a Mátyás király egyik haditettét elbeszélő 150 sor egy hosszabb költeménye része lehetett. A két nyelvemlék tokenszáma (a nem magyar nyelvű részek elhagyásával) összesen 1525.

A betűhű lejegyzés normalizálásánál két alapvető szempontot tartottunk szem előtt: az egységességet, és ugyanakkor az eredetihez való hűséget legalábbis a morfoszintaktikai reprezentáció szintjén. A normalizált alaknak alkalmasnak kell lennie arra, hogy automatikus morfológiai elemzést végezzünk rajta, ezért az erre a reprezentációs szintre való leképezésnél azokat a helyesírási és hangtani különbségeket neutralizáltuk, amelyek az egyébként azonos szóalakokat (ugyanazon lexikai szó ugyanazon morfoszintaktikai jegyekkel bíró előfordulásait) az

eredeti szövegekben véletlenszerű módon megkülönbözteti. Hogy a normalizálást a lehető legegyszerűbb legyen megvalósítani, az automatikus elemzéshez használandó morfológiai elemző elkészítése minél kevesebb adaptációs munkát igényeljen, és minél kevesebb bizonytalansági tényező legyen a leképezés során, a normalizált alakok formáját úgy határoztuk meg, hogy azok a lehető legnagyobb mértékben kövessék a mai magyarban érvényes helyesírási konvenciókat.

A korpusz alapesetben mintegy 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 megfeleltetést tartalmaz, továbbá nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezéseket is, ahol a leképezés megfelelő oldalán üres szimbólum áll. A kiinduló leképezéseket kiterjesztjük olyan továbbiakkal, ahol a két oldalhoz konkatenáljuk adott N szomszédos leképezésből származó szimbólumokat. Körülbelül 7000 kiterjesztés adódik az eredeti megfeleltetésekhez. Az üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek. Példaképpen legyen $N = 3, M = \text{te}, E = \text{the}$, ekkor az alábbi kiinduló leképezések kerülnek a tanítókorpuszba:

$$\begin{aligned} \text{t} &\rightarrow \text{t} \\ \epsilon &\rightarrow \text{h} \\ \text{e} &\rightarrow \text{e} \end{aligned}$$

melyekből továbbá az alábbi helyettesítések generálódnak:

$$\begin{aligned} \text{t} &\rightarrow \text{th} \\ \text{e} &\rightarrow \text{he} \\ \text{te} &\rightarrow \text{the} \end{aligned}$$

A tanítókorpusz manuális előállítását gépi eszközökkel támogattuk. Automatikusan előállítottunk egy olyan változatot, ahol a régi szöveg karakterszinten közelítőleg párhuzamosítva volt a modern szöveggel. Ezt már csak javítani kellett kézzel, így nagy mértékben csökkent a manuális munkaigény. A Prószéky-kóddal kódolt régi szövegek esetében természetesen egy karakternek vettük a különféle Prószéky-kódokat (pl. 'y2', 's43'). A kimenet pontosságának javítása érdekében a következő heurisztikákat alkalmaztuk:

- ha a Prószéky-kód betűje egyezett a mai betűvel, elfogadtuk jó illeszkedésnek
- ha a jelen karakterpár nem egyezett, de a következő igen, akkor elfogadtuk ezt az eltérést az illeszkedésben
- ezt kiterjesztettük két egymás utáni nem egyező karakterpár esetére is
- ha a jelen karakterpár nem egyezett, de vagy a régi vagy a mai szövegben alkalmazott egy elcsúsztatással egyezést találtunk, akkor megfelelően beillesztettünk egy $\epsilon \rightarrow k$ vagy $k \rightarrow \epsilon$ illeszkedést, és csak az egyik szövegben léptünk tovább egy karakterrel.

Ezután az egyes helyettesítések valószínűsége a következőképpen számítható:

$$P(\alpha \rightarrow \beta) = \frac{C(\alpha \rightarrow \beta)}{C(\alpha)} \tag{5}$$

$C(\alpha \rightarrow \beta)$ a tanítókorpuszban látott $\alpha \rightarrow \beta$ helyettesítések, $C(\alpha)$ pedig az α sztring előfordulásainak száma.

4.2. A forrásmodell

A forrásmodell mintegy 10 millió szóból, 65 millió karakterből készült az MNSZ egyik alkorpuszából. Ilyen mennyiségben karakter alapú modelleknél különösebb jelentősége a szöveg regiszterének nincsen, ez a modell paramétereit lényegesen nem befolyásolja. Ugyancsak kevésbé sarkalatos kérdés ilyenkor az alkalmazott simító eljárás. A modell építésénél a CMU nyelvmodell készletet használtuk [5], és az alapbeállítású Good-Turing simítást alkalmaztuk (más eljárás kiválasztása nem változtatott az eredményen, így maradtunk az alapbeállításnál).

5. A modell alkalmazása

Adott E eredeti sztring esetén az $\mathrm{argmax}_M\, P(E|M)P(M)$ értéket kell kiszámítanunk. Ennek általunk alkalmazott (jelenleg teljesen nem optimalizált) módja a következő. Az eredeti szöveg minden partíciójából a transzliterációs modell helyettesítéseiből a lehetséges modern változatokat legeneráljuk, melyekhez a modell hozzárendeli a valószínűségüket is. Ennek alapján kapunk egy rangsort a kapott változatokra, amit aztán a nyelvmodell segítségével újrarendezünk, így alakul ki a az eljárás végleges kimenete.

5.1. Kiértékelés

A projekt kezdeti szakaszában egyelőre csak előzetes eredmények állnak rendelkezésre. Ennek illusztrációja a 2. ábrában látható.

fwl (fül)=>		ygen (igen)=>	
-8,80780895229285	föl	-10,8729908279143	igén
-10,7227286786192	fel	-11,3178857141749	igen
-11,0558158154337	fül	-11,5989613202567	igény
-11,2756412387919	fől	-13,4229320257043	igyen
-12,4574295350367	fol	-14,3578433608162	igin
-12,790296695296	ful	-14,478835649955	igyén
-13,519092302452	fely		
honneg (honnét)=>		sabach (szabács)=>	
-19,1117218113907	honneg	-17,2582527599661	szabács
-19,5230300429664	honnég	-18,1187648297282	sabács
-20,8376176340216	honnét	-18,6771909747334	szabacs
-21,8538140705439	honyneg	-19,1848409742852	sábacs
-22,2098585020436	honynég	-19,5520665992527	szabach
-22,5639991398073	hónneg	-19,9685260661797	szabách

2. ábra. Legjobb n listák különböző bemenetekre.

Az alkalmas kiértékelési módszer legjobb n-es listák vizsgálata, és ezekben a pontosság vizsgálata (a fedés ebben az esetben nem hordoz újabb információt). A

módszer valós használhatósága abban mutatkozik meg, hogy a manuális annotáció redukálható a felkínált alakok közötti választásra, ami jelentősen felgyorsítja a szövegnormalizálás elkerülhetetlen kézi ellenőrzését. Kézenfekvő, hogy az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n-gram nyelvmodellel, és a kimenet szűrhető, illetve átrangsorolható morfológiai elemzés segítségével.

6. Összefoglalás és további feladatok

A dolgozatban megmutattuk, hogy egyszerű sztochasztikus modellek miként alkalmazhatók két reprezentációs szint közötti fordítási feladatra. A további kutatásban számos újabb, a 2. részben említett gépi tanulási módszer alkalmazására van lehetőség [4,13,17], melyek kiértékelése megalapozottan kimutathatja, hogy a vizsgált modellek között melyik a leghatékonyabb, ezzel együtt pedig választ adhat arra a nagyon fontos gyakorlati kérdésre, hogy a manuális átírás hatékonyan kiváltható-e gépi eljárással, így a szükséges emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára, illetve minimális kézi ellenőrzésre. Az itt használt megközelítés is számos részletében finomítható, így a szóhatárok kezelésére, illetve a lehetséges partíciók feletti eloszlásra is kidolgozható modell, és természetesen a jelenlegi implementáció hatékonysága is nagy mértékben növelhető.

Hivatkozások

1. Aw, A.T., Zhang, M., Xiao, J., Su, J. A phrase-based statistical model for SMS text normalization. In: Proceedings of the COLING/ACL. Sydney, Australia. Association for Computational Linguistics (2006) 33–40
2. Bellegarda, J. R. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. Speech Communication 46(2)(2005) 140–152
3. Brill, E., Moore, R. C. An Improved Error Model for Noisy Channel Spelling Correction. In: ACL-00, Hong Kong (2000) 286–293
4. Chen, S. F. Conditional and Joint Models for Grapheme-to-Phoneme Conversion. In: EUROSPEECH-03 (2003)
5. Clarkson, P. R., Rosenfeld, R. Statistical language modeling using the CMU-Cambridge toolkit. In: EUROSPEECH-97, 1. kötet (1997) 2707–2710
6. Damper, R. I., Marchand, Y., Adamson, M. J., Gustafson, K. Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches. Computer Speech and Language 13(2)(1999) 155–176
7. Haader, L. A Müncheni emlék. *Magyar Nyelv* (101)(2005) 161–178
8. Imre, S. *A Szabács Viadala.* Akadémiai Kiadó, Budapest (1958)
9. Kernighan, M. D., Church, K. W., Gale, W. A. A Spelling Correction Program Base on a Noisy Channel Model. In: COLING-90, II. kötet. Helsinki (1990) 205–211
10. Kiss, G., Kiss, M., Pajzs, J. Normalisation of Hungarian Archaic Texts. In: Proceedings of COMPLEX 2001. University of Birmingham, (2001) 83–94
11. Kobus, C., Yvon, F., Damnati, G. Normalizing SMS: are two metaphors better than one? In: Proceedings of the 22nd International Conference on Computational Linguistics, 1. kötet. Manchester, United Kingdom. Association for Computational Linguistics (2008) 441–448

12. Lucassen, J., Mercer, Robert L. An information theoretic approach to the automatic determination of phonemic baseforms. In: ICASSP-84, 9. kötet (1984) 304–307
13. Marchand, Y., Damper, R. I. A multi-strategy approach to improving pronunciation by analogy. Computational Linguistics 26(2)(2000) 195–219
14. McEnery, T., Hardie, A. Lancaster Newsbooks Corpus (2003) http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm
15. Raghunathan, K., Krawczyk, S. Investigating SMS Text Normalization using Statistical Machine Translation. Stanford University (2009)
16. Shannon, C. E. A Mathematical Theory of Communication. Bell System Technical Journal 27(3)(1948) 379–423
17. Taylor, P. Hidden Markov Models for Grapheme to Phoneme Conversion. In: INTERSPEECH-05. Lisbon, Portugal (2005) 1973–1976
18. Toutanova, K., Moore, R. C. Pronunciation Modeling for Improved Spelling Correction. In: ACL-02. Philadelphia, PA. (2002) 144–151

Vektoralapú felügyelet nélküli jelentés-egyértelműsítés nagyméretű tanuló korpuszok esetében

Papp Gyula

Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
Interdiszciplináris Műszaki Tudományok Doktori Iskola
1083 Budapest, Práter utca 50/a
gyupa@digitus.itk.ppke.hu

Kivonat A cikk felügyelet nélküli jelentés-egyértelműsítési (JEÉ) algoritmusok egy lehetséges javítását mutatja be. A módosítás kulcsa a módszerek által alkalmazott tanítóhalmazok méretének megnövelése. Végeztünk egy kísérletet, amely során többféle vektoralapú felügyelet nélküli JEÉ algoritmust teszteltünk a SenseClusters ([1]) programcsomag segítségével. A módszerek kiértékeléséhez egy szabványos adathalmazt, a Senseval-3 JEÉ verseny ([2]) angol főneveit használtuk. A Senseval-3 tanulóadatok mellé a British National Corpus-ból gyűjtöttünk környezeteket annak érdekében, hogy növeljük az algoritmusok tanulóadathalmazainak méretét. A Senseval-2 verseny főnevein végzett paraméterhangolás után az eredmények javulást mutatnak a bővített méretű tanulóhalmazok alkalmazása esetén. Az így kapott rendszer versenyképes a legjobb felügyelet nélküli JEÉ rendszerekkel, például a HyperLex ([5]) algoritmussal.

Kulcsszavak: jelentés-egyértelműsítés, felügyelet nélküli jelentés-egyértelműsítés, környezet-reprezentáció

1. Bevezetés

A jelentés-egyértelműsítés (JEÉ) a nyelvtechnológia egyik legkutatottabb területe. A feladatnak két fő megoldási módszere van: a felügyelt és a felügyeleti nélküli gépi tanulás.

Felügyelt tanulás esetén kézzel címkézett szövegre van szükség ahhoz, hogy a vizsgált többjelentésű szó (a továbbiakban célszó) aktuális jelentését el tudjuk dönteni egy bizonyos környezetben. A felügyelet nélküli módszerek viszont nem igényelnek címkézett tanuló mintákat (környezeteket). Sőt, a célszó aktuális jelentését sem egy előre megadott jelentéslistából választják ki, mint a felügyelt tanulást alkalmazó módszerek. A felügyelet nélküli rendszerek célja a célszó különböző használati eseteinek elkülönítése. (Felügyelet nélküli esetben a szójelentés helyett a használati eset kifejezés használatos.) Klaszterezési algoritmusokat használnak a hasonló tanító minták csoportosításához. Szerencsés esetben az így kialakított klaszterek a célszó egyes használati eseteit reprezentálják.

Korábban még nem látott példa esetén az aktuális környezet reprezentációjához leginkább hasonló klasztert tekintjük az algoritmus által vélt használati esetnek.

A felügyelet nélküli JEÉ egyik előnye a felügyelt módszerekkel szemben az, hogy nem igényel a célszó jelentéseivel címkézett tanítókorpuszt. Mivel olyan címkézetlen környezetből, amely tartalmazza a célszót, tetszőlegesen sok gyűjthető, felvetődött, hogy érdemes lehet megnövelni a felügyelet nélküli algoritmus tanító mintáinak a számát. Feltételeztük, hogy több tanító példára jobb eredményt adnak a felügyelet nélküli JEÉ algoritmusok.

Sok felügyelet nélküli JEÉ módszer vektoralapú reprezentációt használ, amit [4] vezetett be. Emellett gráfalapú módszerekkel is jó eredmények születtek, például mind a HyperLex algoritmus ([5]), mind az optimalizált változata ([3]) nagyon jól teljesít.

Ennek a cikknek a célja annak bemutatása, hogy a környezetek számának növelése hogyan befolyásolja a felügyelet nélküli JEÉ algoritmusok teljesítményét. Elvégeztünk egy kísérletet több vektoralapú módszeren annak vizsgálatára, hogy több tanító környezet esetén javul-e a klaszterezési teljesítmény. A következő fejezet bemutatja a vektoralapú JEÉ módszerek lényegét. Ezt követően bemutatjuk a kísérlet során vizsgált programcsomagot. A 4. fejezet tárgyalja magát a kísérletet. Az elért eredményeket foglalja össze az 5. fejezet. Végül egy rövid összefoglalással zárul a cikk.

2. Vektoralapú felügyelet nélküli JEÉ

A vektoralapú JEÉ algoritmusoknak minden egyes célszóhoz szükségük van egy-egy tanítókorpuszra. Minden egyes korpusz olyan környezetekből áll, amelyek tartalmazzák az aktuális célszót. A környezet egy rövid szövegegység, általában egy mondat, egy bekezdés vagy egy k méretű szóablak, középen a célszóval. A korpuszok formátumára nincs semmilyen megkötés; a módszerek nem igényelnek semmilyen címkézést, így egyszerű szöveg is lehet a korpuszok tartalma.

Ez a fejezet a vektoralapú JEÉ algoritmusok általános működését mutatja be egy adott célszóhoz tartozó tanító korpusz esetén.

2.1. Jegykiválasztás

A vektoralapú módszerek első lépésben jegyeket választatnak ki a korpuszból. (A jegykiválasztás történhet más adatból is, azonban nem ez a szokás.) A jegyek általában szavak, bigramok vagy szó-együttelőfordulások. Egyszavas jegyek lehetnek például a tanítókorpuszban bizonyos számnál gyakrabban előforduló szavak. A bigramok egy kis (2-4 méretű) szóablakban gyakran együtt előforduló rendezett szópárok. Az együtt-előfordulások jellemzően nagyobb szóablakban (például azonos mondatban vagy bekezdésben) gyakran előforduló rendezetlen szópárok. A jegyek kiválasztására a minimális gyakoriságon kívül statisztikai tesztek is alkalmazhatók.

A jegykiválasztás a JEÉ algoritmusok egyik legfontosabb lépése, mert ezek adják majd a környezeteket reprezentáló vektorok dimenzióit.

2.2. Környezet-reprezentáció

A jegykiválasztás eredménye a célszó használati esetei szempontjából releváns-nak vélt jegyek halmaza. A környezet-reprezentációs lépés során az algoritmusok minden egyes környezethez egy-egy vektort rendelnek. Léteznek első-, ill. másodrendű környezetvektorok.

Az elsőrendű környezetreprezentációs vektorok úgy állnak elő, hogy a vektor i-edik eleme az előző lépés során gyűjtött i-edik jegy előfordulási száma az adott környezetben.

A másodrendű környezetvektorok ([4] vezette be őket) bonyolultabb módszerrel számíthatók ki. Bigram és együtt-előfordulási jegyek esetén értelmezzük őket. Első lépésben egy mátrixot készítünk, amelynek sorai a jegyek első, oszlopai pedig a második szavai. A mátrix celláit a sorhoz, ill. az oszlophoz tartozó szavaknak a korpuszbeli együtt-előfordulási száma alapján töltjük ki. Az aktuális környezetet úgy reprezentáljuk, hogy a környezet azon szavait, amelyek szerepelnek a mártix sorcímkéi között, helyettesítjük a sorukkal. Az így kapott vektorok átlaga lesz a környezetreprezentáció.

[6] összehasonlította az első, ill. másodrendű környezet-reprezentációkat. Megmutatta, hogy nagy mennyiségű szöveg esetén az elsőrendű, míg kis méretű korpusz esetén a másodrendű ábrázolás esetén jobb a JEÉ algoritmusok teljesítménye.

2.3. Dimenziószám-csökkentés

A környezet-reprezentációs vektorok általában elég ritkák, azaz viszonylag sok 0 elem található bennük. Néhány esetben a dimenziószámuk is túl nagy a klaszterezési algoritmusok számára. Emiatt javasolta [4] a szinguláris érték-felbontást (SVD), amellyel a másodrendű jegyek mátrixának méretét lehet csökkenteni, mindezt simítással egybekötve. Elsőrendű jegyek esetén a jegyvektorokból mint sorokból előállított mátrixra is alkalmazható az SVD transzformáció. Az SVD javíthatja a JEÉ rendszerek hatékonyságát. (Egyéb módszerek is alkalmazhatók a dimenziószám-csökkentésére.)

2.4. Klaszterezés

Miután rendelkezésre állnak a környezet-reprezentációs vektorok, már tetszőleges klaszterező algoritmus használható a célszó használati eseteinek elkülönítésére. Általában a klaszterező algoritmus bemenő paraméterként igényli a kialakítandó klaszterek számát. [7] javasolt néhány függvényt a megfelelő klaszterszám előre történő meghatározására.

2.5. Kiértékelés

Több módszer is létezik felügyelet nélküli JEÉ rendszerek kiértékelésére; [3] foglalja össze ezeket. Egy lehetőség az algoritmus eredményét „kézzel” elemezni. Másik alternatíva lehet JEÉ rendszer teljesítményét egy alkalmazásban mérni.

Esetleg a célszó jelentéseivel címkézett korpusz is használható a kiértékelésre. Végül azt is megtehetjük, hogy a tökéletesnek vélt klaszterekkel hasonlítjuk össze az algoritmus eredményét.

3. A vizsgált jelentés-egyértelműsítő rendszer

A kísérletünk során alkalmazott JEÉ rendszer az ingyenesen elérhető SenseClusters programcsomagból ([1]) és egy saját fejlesztésű kiértékelő modulból állt. Ez a fejezet röviden bemutatja a rendszer moduljait és a számukra szükséges bemenő paramétereket.

3.1. A SenseClusters moduljai

A SenseClusters bemenetként a célszót tartalmazó bekezdéseket vár. A jegykiválasztó modulja segítségével lehetőség van egyszavas, bigram, ill. együttelőfordulási jegyek gyűjtésére. Egyaránt lehetséges minimális előfordulási gyakoriságot, valamint valamilyen statisztikai mértéket megadni a jegyek elfogadásához.

A jegykiválasztás után a környezetreprezentációs modul hajtódik végre. A modul egyszavas jegyek esetén elsőrendű, bigram és együtt-előfordulási jegyek estén pedig mind első-, mind másodrendű környezet-reprezentációra képes.

A dimenziószám-csökkentő modul SVD transzformáció segítségével próbálja a reprezentációs vektorokat simítani. (Ennek a modulnak a végrehajtása opcionális.)

A SenseClusters a CLUTO programot ([8]) alkalmazza klaszterezésre. A CLUTO egyaránt támogat agglomeratív, particionális és hibrid klaszterezési algoritmusokat. Ezek a módszerek bemenő paraméterként igénylik az előállítandó klaszterek számát. Ennek meghatározásában a SenseClusters PK1, PK2, PK3 ([7]), ill. GS ([9]) mértékei nyújtanak segítséget.

3.2. A kiértékelő modul

Annak érdekében, hogy a kísérletünk eredményei más hasonló munkákéval összehasonlíthatóak legyenek, [3]-hoz hasonlóan a célszó jelentéseivel címkézett környezeteken végeztük a kiértékelést. Ezeket a környezeteket felosztottuk tanító és kiértékelő részre. A kiértékelési folyamat egy felügyelt tanulási feladat: a címkézett tanulókörnyezeteken tanuljuk meg a klaszter-jelentés hozzárendeléseket, a hatékonyságot pedig a címkézett tesztkörnyezeteken mérjük.

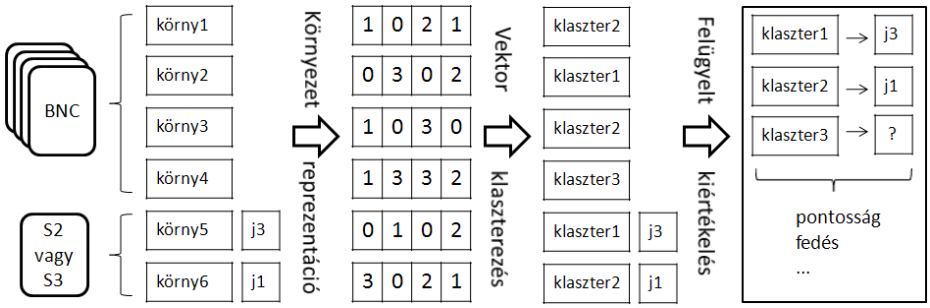
Habár a SenseClusters is nyújt kiértékelő szolgáltatásokat, ezeket nem lehetett a fent leírt módon alkalmazni, úgyhogy egy saját kiértékelő modul elkészítésére volt szükség.

Ugyan ez a kiértékelési módszer a rendszert felügyelet nélküli és felügyelt tanulás keverékévé teszi, fontos megjegyezni, hogy a klaszterek előállítása teljesen felügyelet nélkül zajlik, csupán a klaszter-jelentés párok kialakítása történik felügyelt módon.

3.3. Paraméterek

A kísérlet során az egész SenseClustert egy felügyelet nélküli JEÉ rendszernek tekintettük. A 3.1-es alfejezetben bemutatott különböző funkcióit (például hogy első- vagy másodrendű reprezentációt alkalmazunk) a rendszer szabad paramétereinek tekintettük. Ezeket próbáltuk hangolni.

Természetesen a paraméter-optimalizálást más adathalmazon kell végezni, mint amin a rendszer teljesítményét mérjük. A kiértékelést az elkülönített adatokon az optimalizált paraméterekkel végeztük.



1. ábra. A kísérlet folyamata

4. A kísérlet

A kísérletet egy szabványos adathalmazon végeztük el azért, hogy más rendszerekével összehasonlítható legyen az eredmény. A Senseval-3 jelentés-egyértelműsítő verseny 20 angol főnevét, ill. az ezekhez tartozó korpuszokat választottuk ki erre a célra. A kísérlet során a Senseval-3 tanító adathalmaz - tesztadathalmaz elkülönítését alkalmaztuk.

A paraméterek hangolására a Senseval-2 verseny 20 angol főnevét választottuk ki. A hozzájuk tartozó korpuszokat használtuk az optimalizálás során végzett kiértékeléshez.

4.1. A kísérlet menete

A kísérlet folyamata az 1. ábrán látható. Első lépésként az egyes célszavakhoz állítottuk elő a korpuszokat. Minden egyes célszóhoz a British National Corpusból (BNC) gyűjtöttünk olyan bekezdéseket, amelyek az adott többjelentésű szót tartalmazzák. A Senseval adathalmaz környezeteit, amelyek a célszavak jelentéseivel címkézettek, hozzáadtuk a megfelelő szóhoz tartozó, BNC-ből gyűjtött korpuszhoz. (A Senseval adathalmaz Senseval-2-t jelent paraméter optimalizálás, és Senseval-3-at kiértékelés esetén.) Ezzel a módszerrel minden egyes célszóhoz 2000-3000 környezetet sikerült gyűjteni.

Ezután a SenseClusters moduljait hajtottuk végre a kigyűjtött környezeteken. Végül a saját kiértékelő modul segítségével mértük az egyes módszerek hatékonyságát.

Az egész kísérletet megismételtük csupán a Senseval adatokon. Az így kapott eredményeket a bővített adathalmazon kapottakkal összevetve tudtunk következtetést levonni a környezetek számának szerepéről.

4.2. Az optimális paraméterek

A 4. fejezet bevezetésében már szerepelt, hogy a paraméterek hangolása a Senseval-2 adatokon történt. Abban az esetben, amikor a SenseClusters-t a bővített adathalmazon futtattuk, a legjobb eredményt elsőrendű együtt-előfordulási jegyek segítségével értük el. Ez összhangban van [6] következtetéseivel, melyeket a 2.2. alfejezetben említettünk. Egy particionális klaszterezési módszer, az ún. ,,Repeated Bisection" algoritmus bizonyult a legjobbnak. Az SVD transzformáció alkalmazása nem javított az eredményeken.

Az optimális paraméterhalmaz nagyon hasonló volt abban az esetben, amikor csak a Senseval-2 adatokon végeztük a kísérletet, mindössze a jegyek típusa bizonyult más esetben optimálisnak: az egyszavas jegyek nyújtották a legjobb teljesítményt.

5. Az eredmények

Az 1. táblázaton szerepelnek a kísérlet eredményei. A leggyakoribb jelentés heurisztika jelentette a baseline módszert. Az optimalizált paraméterekkel elindított algoritmus a BNC környezetekkel kiegészített adatokon futtatva kis mértékkel jobbnak bizonyult az alapadatokon futtatott esetnél. Mindkét verzió lényegesen felülmúlja a baseline algoritmust.

Az elért eredmények versenyképesek a többi Senseval-3 adatokon kiértékelt felügyelet nélküli JEÉ rendszerek eredményeivel. A 2. táblázaton látható, hogy egyedül az optimalizált HyperLex algoritmus ([3]) teljesített jobban. Az SCBNC, ill. az SCS3 nevek jelölik az általunk alkalmazott rendszereket. (A többi rendszer hatékonyságát [3] mérte meg.)

Habár ezek a rendszerek ugyanazon az adathalmazon lettek kiértékelve, mégis nehéz őket összehasonlítani a különböző alkalmazott tanítási módszerek miatt. Némelyik algoritmusnak szüksége van a leggyakoribb jelentés ismeretére (ezeket MFS-Sc jelöli, ha a leggyakoribb jelentést a SemCor, MFS-S3, ha a Senseval-3 adatok alapján számítja a módszer), némelyek a Senseval-3 tanítópéldák 10%-át használják a klaszter-jelentés hozzárendelés tanulására (10%-S3TR), mások erre a teljes tanítóhalmazt igénybe veszik (S3TR) [3].

6. Összefoglalás

Ez a cikk bemutatott egy lehetséges módszert felügyelet nélküli JEÉ algoritmusok teljesítményének növelésére. Ehhez mindössze olyan környezetekre volt

1. táblázat. A kiértékelés eredménye a Senseval-3 főneveken. Az első oszlopban szerepelnek a vizsgált célszavak. Emellett állnak a leggyakoribb jelentések arányai. Az utolsó két oszlop mutatja a kísérlet eredményét az alapadatok, valamint a BNC környezetekkel kiterjesztett korpuszok esetén. A táblázatban feltüntetett számok a pontossági értékek. (A fedés megegyezik a pontossággal.)

Szó	MFS	SCS3	SCBNC
argument	51.4	48.6	51.4
arm	82.0	85.0	85.7
atmosphere	66.7	72.8	71.6
audience	67.0	70.0	76.0
bank	67.4	72.7	72.0
degree	60.9	67.2	68.8
difference	40.4	48.2	43.0
difficulty	17.4	47.8	26.1
disc	38.0	71.0	66.0
image	36.5	60.8	60.8
interest	41.9	59.1	66.7
judgment	28.1	40.6	40.6
organization	73.2	73.2	69.6
paper	25.6	44.4	52.1
party	62.1	64.7	65.5
performance	32.2	42.5	46.0
plan	82.1	78.6	77.4
shelter	44.9	42.9	48.0
sort	65.6	65.6	65.6
source	65.6	50.0	50.0
Átlag:	54.5	61.9	62.9
(Senseval-2 adatokon)	51.9	59.0	59.8

2. táblázat. A Senseval-3 angol főnevein kiértékelt felügyelet nélküli JEÉ rendszerek összehasonlítása.

Rendszer	Típus	Pontosság	Coverage
HyperLex	S3TR	64.6	1.0
SCBNC	S3TR	62.9	1.0
SCS3	S3TR	62.0	1.0
Cymfony	10%-S3TR	57.9	1.0
Prob0	MFS-S3	55.0	0.98
MFS	-	54.5	1.0
Ciaosenso	MFS-Sc	53.95	0.90
clr04	MFS-Sc	48.86	1.0
duluth-senserelate	-	47.48	1.0

szükség, amelyek tartalmazták az éppen vizsgált célszót. Ezekkel a környezetekkel kiegészítve a tanuló korpuszt némileg javult az algoritmusok hatékonysága. A módszer hátránya, hogy a tanulási folyamat idejét megnöveli.

A paraméterek optimalizálása után kapott vektoralapú JEÉ algoritmus versenyképes a jelenlegi legjobb hasonló rendszerekkel, azonban az összehasonlítás elég nehéz feladat a különböző módon végzett klaszter-jelentés hozzárendelések miatt.

Hivatkozások

1. Purandare, A., Pedersen, T.: SenseClusters - finding clusters that represent word senses. In: Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04). San Jose, USA (2004) 1030–3031
2. Mihalcea, R., Chklovski, T., Kilgarriff, A.: The Senseval-3 English lexical sample task. In: Senseval-3 proceedings (2004) 25–28
3. Agirre, E., Martínez, D., de Lacalle, O.L., Soroa, A.: Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In: Proc. of the TextGraphs Workshop: Graph-based algorithms for Natural Language Processing. New York, USA (2006) 89–96
4. Schütze, H.: Automatic word sense discrimination. Computational Linguistics, **24**(1) (1998) 97–123
5. Véronis, J.: HyperLex: lexical cartography for information retrieval. Computer Speech & Language, **18**(3) (2004) 223–252
6. Purandare, A., Pedersen, T.: Word sense discrimination by clustering contexts in vector and similarity spaces. In: Proc. of the Eighth Conference on Computational Natural Language Learning (CoNLL). Boston (2004) 41–48
7. Pedersen, T., Kulkarni, A.: Selecting the „right" number of senses based on clustering criterion functions. In: Proc. of the Posters and Demo Program of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics. Trento (2006) 111–114
8. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: Proc. of the 11th Conference of Information and Knowledge Management (CIKM). McLean, USA (2002) 515–524
9. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the Gap statistic. Journal of the Royal Statistics Society (Series B) **63**(2) (2001) 411–423

Magyar igei vonzatkeretek gépi tanulása

Babarczy Anna, Serény András, Simon Eszter

BME GTK Kognitív Tudományi Tanszék,
1111 Budapest, Stoczek utca 2.
e-mail: {babarczy,esimon}@cogsci.bme.hu, andras.sereny@gmail.com

Kivonat A lexikális információ gépi tanulását lehetővé tévő módszerek a számítógépes nyelvészet fontos részterületét alkotják, mert számos természetes nyelvi kétértelműség csak lexikális tudás birtokában oldható fel. Igék esetén ilyen lexikális tulajdonság az is, hogy az ige milyen vonzatkeretekben szerepelhet, azaz milyen kategóriájú bővítményekkel együtt jelenik meg a mondatban. Cikkünkben az igei vonzatkeretek gépi tanulásának más nyelvekre jól működő megközelítéseit, statiszitkai módszereit alkalmazzuk magyar nyelvre. Ezzel párhuzamosan kutatásunknak célja az is, hogy valamilyen módon modellezzük az emberi nyelvelsajátítást, legalábbis a vonzatkeretek elsajátítását; a gépi tanulási görbéket gyereknyelvi adatokból számított tanulási görbékkel vetjük össze.

Kulcsszavak: vonzatkeret-elsajátítás, pszicholingvisztika

1. A lexikális tudás kérdése

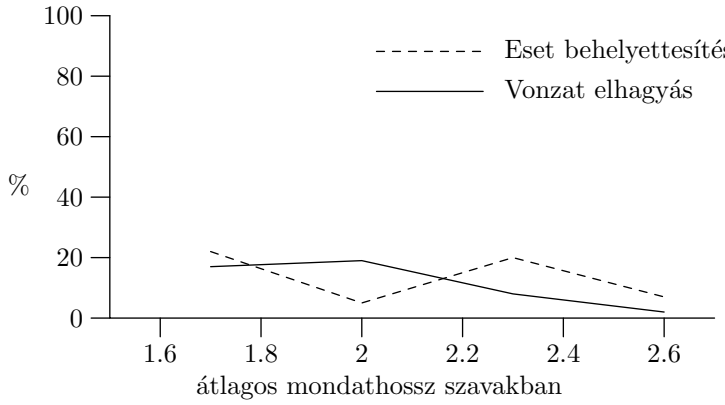
Lexikális tudás elsajátítása alatt a szavak és ezek idioszinkratikus (nem általános elvekből következő) tulajdonságainak elsajátítását értjük, beleértve szemantikai és szintaktikai tulajdonságokat. A predikatív nyelvi elemek – köztük az igék – lexikális tulajdonságai közé tartozik a vonzatszerkezetük, azaz hogy milyen kategóriájú, illetve morfoszintaktikai szerkezetű bővítményekkel jelenhetnek meg a mondatban. Ez a tudás nem csak a mondatalkotás, hanem a mondatfeldolgozás szempontjából is elengedhetetlen. Például az *elad* és a *megsimogat* igék vonzatkeretének ismeretében tudjuk azt, hogy míg az alábbi (1) mondat kétértelmű (Lili szomszédja lehet a cselekvés célpont argumentuma vagy a kutya eredeti gazdája), a (2) alatt szereplő mondat nem az (Lili szomszédja itt nem lehet argumentum).

(1) Marci eladta Lili szomszédjának a kutyáját.

(2) Marci megsimogatta Lili szomszédjának a kutyáját.

A lexikális tudás elsajátításának mechanizmusai két szempontból is érdekes kutatási téma. Egyrészt a pszicholingvisztikában fontos kérdés a nyelvi tudás ezen alapelemének fejlődése, másrészt a számítógépes nyelvfeldolgozás területén a gépi elemző rendszerek egyik fő problémája. Kutatásunk a gyereknyelv empirikus tapasztalataiból kiindulva próbálja a gépi nyelvfeldolgozás módszereit

fejleszteni, míg a másik irányban a számítógépes modellek működésén keresztül igyekszünk fényt deríteni az empirikus tapasztalatok mögött rejlő emberi tanulási mechanizmusokra. A korai automatikus lexikonépítési kísérletekben nem számítógépes célokra készült szótárak elektronikus változatát használták nyersanyagként. Az automatikus módszerek közül ez a megközelítés áll legközelebb a kézi előállításhoz, éppen emiatt rendelkezik a nem automatikus módszer fő hátrányaival: nem elég rugalmas, és nem teszi lehetővé az automatikus bővítést, ezáltal nem vihető át más területre. A szótár használatánál robusztusabb megközelítést jelent az igei vonzatkeret-információ automatikus kinyerése nagyméretű korpuszokból. A gyereknyelvi adatok is arra utalnak, hogy az anyanyelv elsajátításakor a mentális lexikon nem az egyes igék vonzatszerkezetének egyenkénti memorizálásával épül, hanem a gyerekek, az input statisztikai tulajdonságait felhasználva, mintákat vonnak ki abból. Ez a tanulás egyes szakaszaiban hibákhoz vezethet. Amint az 1. ábrából kiderül, a gyereknyelvben előforduló vonzatkeretek nem mindig felelnek meg a célnyelvtan által elfogadott vonzatkereteknek.

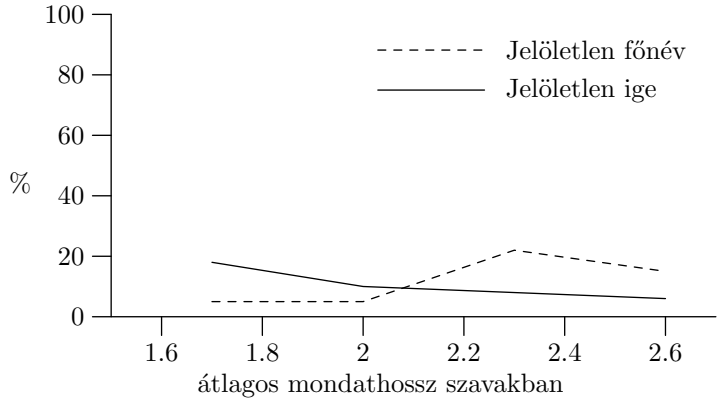


1. ábra. Helytelen nem alanyi esetragok és elhagyott kötelező vonzatok aránya a korai magyar gyereknyelvben. Három gyerek spontán nyelvi produkciójának súlyozatlan átlaga. Korpuszméret: 18 644 szó

A feladatot úgy fogalmazhatjuk meg, hogy ha adott egy F vonzatkeretkészlet és egy V igekészlet, az inputban megjelenő mondatok alapján döntsük el minden $(f, v) \in F \times V$ párról, hogy a nyelvtan szerint f lehet-e v vonzatkerete. A tanulás eredményeként megengedett ige–vonzatkeret párok alkotják a tanuló lexikonját. A gyereknyelv esetében az input a gyerek nyelvi környezetét jelenti, a számítógépes modell pedig digitális korpuszokból tanul. A továbbiakban igei vonzatkeret alatt egyszerűen azt az információt értjük, hogy az ige bővítményei a mondatban milyen (felszíni) esetben vannak, mivel a magyar nyelvben a vonzatok szintaktikai, illetve tematikai szerepét elsősorban az esetrag jelöli.

A fenti leírás feltételezi, hogy a gyerek számára is adott egy vonzatkeretkészlet és egy igekészlet, és a feladata hasonlóképpen az, hogy az igékhez a megfelelő von-

zatkereteket rendelje. Ezt a feltételezést az a megfigyelés támasztja alá, hogy a korai gyereknyelvet egyszavas mondatok jellemzik, igék és főnevek egyaránt, melyeket tekinthetünk predikátumok és argumentumok egyszerű megjelenítésének. A magyarban (és más gazdag morfológiájú nyelvekben) a korai gyereknyelv mondatai jellemzően ragozott szavakból állnak: az igék inflexiókkal, a főnevek pedig esetragokkal jelennek meg. Természetes gyereknyelvi korpuszelemzéseink megerősítették ezt a megfigyelést: a 2. ábrán látható, hogy viszonylag kevés inflexióelhagyási hiba fordul elő a magyar gyereknyelvben azelőtt is, hogy az átlagos mondathossz elérné a két szót (a jóval gyakoribb morfofonológiai hibákat és ragbehelyettesítéseket itt figyelmen kívül hagyjuk). Feltesszük tehát, hogy a gyerek



2. ábra. A jelöletlen (esetraggal nem ellátott, nem alanyi szerepű) főnevek és a jelöletlen (személyraggal nem ellátott, nem egyesszám harmadik személyű alanyú) igék aránya a korai magyar gyereknyelvben. Három gyerek spontán nyelvi produkciójának súlyozatlan átlaga. Korpuszméret: 18 644 szó.

számára adott a világ eseményeinek és az azt leíró nyelvnek predikátumokba és a hozzájuk tartozó argumentumokba való szerveződése. A fenti adatokra támaszkodva feltesszük továbbá, hogy a gyerek számára ismert az esetragozás mechanizmusa. Ezek a nyelv általános törvényszerűségeiből következő tudások, melyek eredetével kutatásunk nem foglalkozott.

2. A gépi modellek

2.1. Alapelvek

Kutatásunk fő irányvonala az argumentumstruktúrák elsajátításának számítógépes modellezése volt. A vonzatkeretek gépi tanulására első megközelítésként Brent [1] statisztikai módszerének gazdag morfológiájú nyelvekre adaptált változatát alkalmaztuk. Bár Brent módszere – a számítógépes nyelvészet fejlődési

mondat	KR annotáció
Én	NOUN<PERS<1>>
ma	ADV
már	ADV
nyertem	VERB<PAST><PERS<1>>
négy	NUM
forintot.	NOUN <CAS<ACC>>

1. táblázat. Mondat morfológiai annotációja a KR-kód felhasználásával.

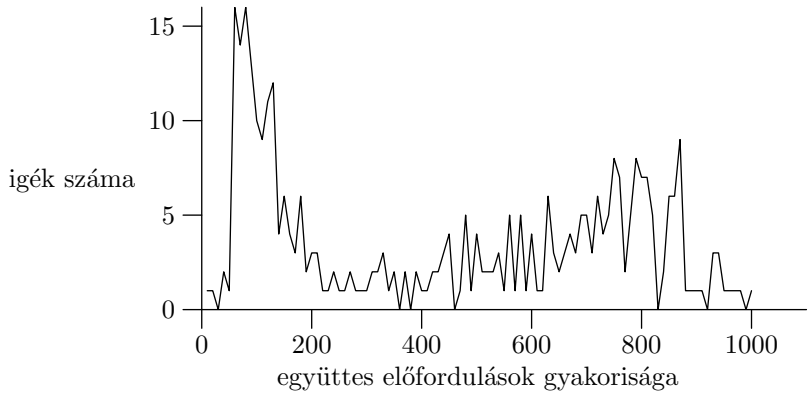
ütemét tekintve – elég réginek nevezhető, magyar vonzatkeretek azonosítására (tudomásunk szerint) ez az első alkalmazása. A magyar nyelvvel foglalkozó munkák közül a miénkhez hasonló tárgyú [6], de ez az idiomatikus, nem kompozicionális, rögzített lemmával előforduló igei szerkezetek kigyűjtését tűzi ki célul. Röviden, Brent eljárásának az a feltételezés az alapja, hogy minden vonzatkerethez tartoznak ún. jegyek. Egy jegy olyan mintázat vagy formai sajátosság, amelynek megjelenése egy mondatban valószínűsíti, hogy a mondatban előfordul a jegyhez tartozó igei vonzatkeret. Például a „tárgyas ige" vonzatkerethez tartozhat a következő jegy: a mondatban pontosan egy ige van, és van benne tárgyesetű névszó. Az általunk használt jegyrendszer egyszerű reguláris kifejezésekből áll, melyek a KR morfológiai annotációs kód [10] elemeire illeszkednek: egy jegy illeszkedik egy mondatra, ha a megfelelő reguláris kifejezés illeszkedik a mondathoz tartozó morfológiaiannotáció-sztringre. Az 1. táblázatban egy példát láthatunk. A magyar ditranzitív vonzatkeret például a következő kódnak felel meg:

(CAS<ACC>.* CAS<DAT>) |(CAS<DAT>.* CAS<ACC>)

A számítógépes modellben felhasznált jegyeket a gyereknyelvi korpuszban konzisztensen előforduló, a felnőtt nyelvtan szabályainak megfelelő argumentumszerkezetek részletei adják. Minden jegyhez tartozik egy hibavalószínűség, ez annak a valószínűsége, hogy a jegy ugyan megjelenik egy mondatban, de a jegyhez tartozó vonzatkeret mégsem tartozik az adott predikátum megengedett vonzatkeretei közé.

2.2. Hibavalószínűségek

A hibavalószínűségek (ε) meghatározása különböző módszerekkel történhet. Elméleti szempontból az a módszer tűnt az emberi nyelvelsajátítás legjobb megközelítésének, amely a vonzatkeretek disztribúciójára épül. (Amint a 3. alfejezetben látni fogjuk, végül nem ez a módszer bizonyult a legsikeresebbnek.) Vesszük a korpuszban egyenként legalább N-szer előforduló igék első N előfordulását, és kiszámoljuk, hogy egy f vonzatkerethez tartozó jeggyel hány ige szerepel egy adott $1 \leq i \leq N$ gyakorisággal. A 3. ábrán a magyar tranzitív keretet jelölő CAS<ACC> jegyre vonatkozó statisztika látható. (Részletesebb leírásához lásd [8].) Azt az i_0 gyakoriságot keressük, amelyre igaz, hogy (ebben az esetben)



3. ábra. A tranzitív keretet jelző CAS<ACC>jegy előfordulási valószínűsége a korpuszban szereplő igékkel.

az intranzitív igék többsége i_0 vagy annál kisebb gyakorisággal fordul elő az adott jeggyel, míg a valódi tranzitív igék többsége i_0 vagy annál nagyobb gyakorisággal fordul elő a jeggyel. A megfelelő gyakorisági érték esetén a fenti grafikon bal oldalán egy (ferde) binomiális alakzat jelenik meg. Ebből becsülhetjük meg i_0 értékét, majd az ε hibavalószínűséget. A hibavalószínűségek ismeretében egy statisztikai modellel döntünk arról, hogy egy ige megjelenhet-e egy adott vonzatkerettel. Három különböző statisztikai modellt próbáltunk ki: binomiális modell, likelihood hányados modell és relatív gyakoriságok.

2.3. Binomiálishipotézis-próba

Ebben a modellben a nyelvtan kiinduló állapotában minden ige–vonzatkeret párra az áll, hogy egy adott ige nem jelenhet meg egy adott vonzatkerettel, és a nyelvtan csak megfelelő pozitív input hatására módosul (konzervatív tanulás). Az automatikus vonzatkeret-kinyerés feladatának megoldásához először is definiálnunk kell azokat a számszerűsíthető tulajdonságokat, melyek a keresett lexikai információra jellemzőek. A legtöbb módszer az ige és a vonzatjelölt együttes előfordulási statisztikáiból indul ki.

Tehát minden f vonzatkerethez hozzárendelünk egy jegykészletet

$$f \mapsto \{c_1^f, c_2^f, \ldots c_{n_f}^f\}$$

és egy e_f hibavalószínűséget, ahol a hibavalószínűség

$$e_f = P(c_i^f \text{ megjelenik } S\text{-ben} \mid v\text{-nek } f \text{ nem vonzatkerete})$$

Miután minden keresendő vonzatkerethez rögzítettük jegyek egy halmazát, a következő egyszerű statisztikai modellel döntünk arról, hogy egy ige megjelenhet-

e egy adott vonzatkerettel:

$$p_e = P(C(v,f) \geq m \mid v\text{-nek } f \text{ nem vonzatkerete}) = \sum_{r=m}^{n} \binom{n}{r} \varepsilon_f^r (1-\varepsilon_f)^{n-r}.$$

Veszünk egy v igét és egy f vonzatkeretet. Nullhipotézisünk, hogy a nyelvtan szerint az ige nem jelenhet meg ezzel a vonzatkerettel. A korpuszban megszámoljuk, hogy az ige hányszor fordul elő összesen (n), és hányszor fordul elő a vonzatkerethez tartozó jegyekkel ($C(v,f)$). Ha az ige viszonylag sokszor fordul elő a vonzatkerethez tartozó jegyek valamelyikével (p_e kisebb, mint egy előre meghatározott érték), akkor ez arra utal, hogy nullhipotézisünk hibás, a nyelvtan megengedi ezt az ige–vonzatkeret párt. Pontosabban, az ige minden előfordulásakor véletlen kísérlet eredményének tekintjük, hogy egy jegy megjelenik-e, vagy nem. A jegy megjelenésének valószínűsége (a nullhipotézis mellett) éppen a jegyhez tartozó hibavalószínűség. A kísérletek eredményei egymástól függetlenek.

2.4. Likelihood hányados próba

A gyereknyelvi elemzésekből tudjuk azonban, hogy a vonzatkeretek elsajátítása során túláltalánosításra utaló tanulási mintákat figyelhetünk meg, vagyis az első modell szigorúan konzervatív tanulási algoritmusa valószínűleg nem felel meg a pszicholingvisztikai tényeknek (a modell eredményeit a cikk 3. alfejezetében ismertetjük). Míg az első néhány életévben a gyerek nyelvi produkciójában az ige–vonzatkeret párok száma folyamatosan emelkedik, a helyes argumentumstruktúrák aránya egyes tanulási fázisokban akár csökkenhet is (U-alakú tanulási görbe). Az előbbi mérőszámot a számítógépes nyelvészet felidézés (recall) fogalmának, az utóbbit pedig a pontosság (precision) fogalmának feleltethetjük meg. Célunk a gyereknyelv és a modell felidézési és pontossági görbéinek egymáshoz való közelítése.

Második modellünkkel olyan statisztikai módszert implementáltunk, amely azt teszteli, hogy egy adott v ige megjelenése és egy adott f vonzatkerethez tartozó jegy megjelenése egy mondatban független eseményeknek tekinthetők-e, azaz hogy együttes előfordulásuk mennyire véletlenszerű. Ha a két esemény nem független, f v vonzatkeretének tekinthető. A likelihood hányados logaritmusa

$$\lambda = l(\frac{k_1+k_2}{n_1+n_2}, k_1, n_1) + l(\frac{k_1+k_2}{n_1+n_2}, k_2, n_2) - l(\frac{k_1}{n_1}, k_1, n_1) - l(\frac{k_2}{n_2}, k_2, n_2),$$

ahol k_1, n_1, k_2, n_2 rendre v és f jegyének együttes előfordulásának számát, a korpuszban szereplő igék számát, f jegyének más igékkel való előfordulásának számát és a v igével nem azonos igék számát jelöli, valamint $l(q,n,k) = k \log q + (n-k)\log(1-q)$. Ismert, hogy λ eloszlásban tart egy χ^2 eloszláshoz, tehát λ értékeit a χ^2 eloszlás kritikus értékeihez hasonlítva adott szignifikanciájú próbához jutunk. (A modell részletesebb leírását lásd [8].)

Mivel ez a modell egy adott vonzatkeret más igékkel való előfordulási gyakoriságát érzékenyebben veszi figyelembe, mint az előző modell hibavalószínűségi

paramétere, elméletben közelebb áll az emberi nyelvelsajátítás esetében feltételezett általánosító majd a hibás általánosításokat „visszatanuló” tanulási mechanizmushoz.

2.5. Relatív gyakoriságok

Harmadik modellünk a [5] által baseline-nak javasolt eljárást valósítja meg. Ez az egyszerű módszer azokat az ige–vonzatkeret párokat fogadja el, ahol a vonzatkerethez tartozó jegyek és az ige együttes előfordulási gyakoriságának az ige előfordulási gyakoriságához viszonyított aránya meghalad egy küszöbértéket. A küszöbértéket empirikus úton határozzuk meg.

3. Eredmények

A három modellt a magyar Webkorpusz [4] egy 800 ezer mondatos darabján és a Szeged Korpuszon [2] teszteltük. A Webkorpusz morfológiai annotációját és egyértelműsítését a Hunpos szófaji egyértelműsítővel [3] végeztük. A morfológiai elemzés a KR annotációs nyelvtant használja (ennek részletes leírását lásd [9], [10]). Néhány eredmény látható a 2. táblázatban (az eredmények részleteit lásd [8]). Összességében azt állapíthatjuk meg, hogy mindhárom modell teljesítménye

Módszer	Korpusz	Igék száma	Pontosság	Felidézés	F-mérték
Binomiális	Webkorpusz	1000	70%	67%	68%
Binomiális	Szeged Korpusz	1000	63%	50%	56%
Relatív gyakoriság	Webkorpusz	1000	90%	67%	76%
Likelihood próba	Webkorpusz	1000	25%	79%	39%
Likelihood próba	Szeged Korpusz	1000	35%	56%	43%
Binomiális	Webkorpusz	200	64%	94%	76%
Binomiális	Szeged Korpusz	200	75%	70%	72%

2. táblázat. A három modell teljesítménye a három leggyakoribb vonzatkeret elsajátításában.

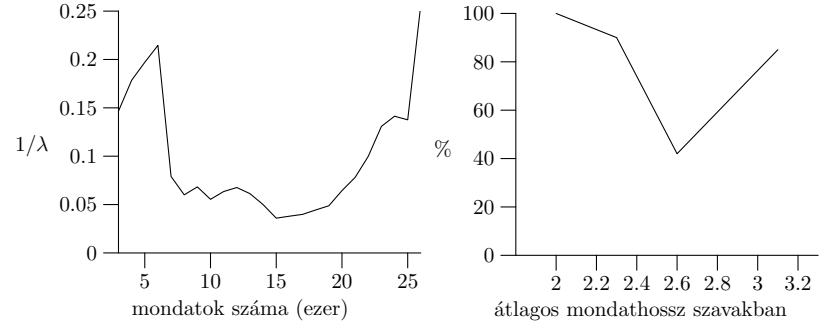
jelentősen javul, ha csak a három leggyakoribb vonzatkeretet vesszük figyelembe. A Brent-féle binomiális módszeren alapuló kísérletet több hibavalószínűségi értékkel is elvégeztük, a táblázatokban elsősorban a 2.2. alfejezetben ismertetett módon előre megbecsült hibavalószínűségi értékekkel számolva kapott értékeket tüntettük fel. Az eredmények alapján azt látjuk, hogy ha emeljük a hibavalószínűség értékét, akkor a pontosság megnő, a felidézés értéke viszont csökken. Az F-mérték számításakor persze kiegyensúlyozódnak ezek az értékek, de alacsonyabb hibavalószínűségnél összességében jobb teljesítményt kapunk. A likelihood hányados próba a binomiális módszernél a gépi nyelvfeldolgozás szempontjából kissé gyengébb eredményeket hozott, de a tanulási görbe arra enged

következtetni, hogy több tanító adaton (nagyobb korpuszon) a jelenleginél jobban teljesítene. A pszicholingvisztikai párhuzamot tekintve a felidézés magas értéke a pontosság alacsony értékével párosítva a gyereknyelv fejlődésének azt a szakaszát idézi, amikor a kezdeti konzervatív tanulási stratégiát felváltja az általánosító stratégia. Meglepő módon, a gépi tanulás szempontjából a relativ gyakoriságon alapuló döntés adta a legjobb eredményt.

Módszer	Korpusz	Igék száma	Pontosság	Felidézés	F-mérték
Binomiális	Webkorpusz	100	61%	71%	64%
Binomiális, $\varepsilon = 0,5$	Webkorpusz	100	94%	34%	51%
Relatív gyakoriság	Webkorpusz	100	77%	56%	65%

3. táblázat. A modellek teljesítménye 43 magyar vonzatkeret elsajátításában.

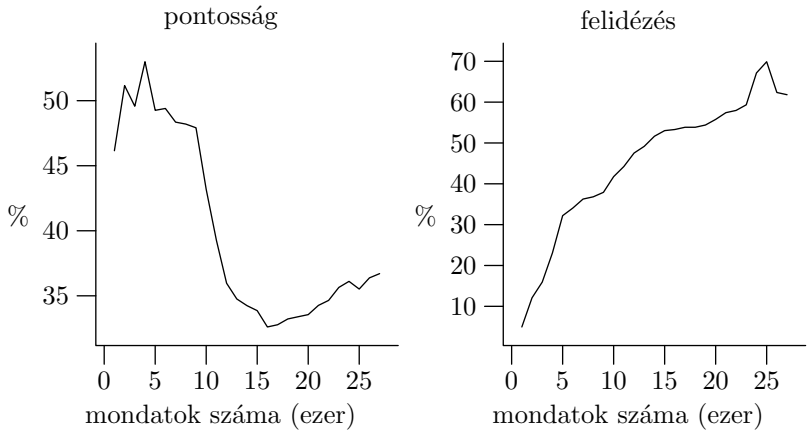
A pszicholingvisztikai párhuzam szemléltetése érdekében méréseink eredményét grafikusan is ábrázoljuk (4. ábra). A likelihood statisztika $1/\lambda$ reciproka jó mérőszáma annak, hogy a modell egy adott ige–vonzatkeret párt „mennyire" gondol helyesnek. Ez a görbe (bal grafikon) hasonló U-alakot mutat, mint a gyerekek tanulási görbéje (jobb grafikon). A tanulási görbe vízszintes tengelyén az idő szerepel (az átlagos mondathosszal jelölve): a kor előrehaladtával a gyerek több bemeneti adathoz jut, vagyis tökéletesíteni tudja mentális nyelvtanát, és a pontosan használt nyelvtani szerkezetek aránya nő. A likelihood próba eredményének vízszintes tengelyén a korpusz mérete szerepel, ami hasonló funkciót tölt be a gépi tanulás folyamatában. A gyereknyelvi korpuszok elemzése során arra az



4. ábra. A likelihood statisztika görbéje a *kér – kér valamiből* ige–vonzatkeret párra a Szeged Korpuszon (balra) és három magyar gyerek beszédprodukciójában a *kér* ige helyes vonzatkerettel való használatának aránya (jobbra).

eredményre jutottunk, hogy a jól érzékelhető, szisztematikus vonzatkerethibák

egy-egy igére vagy igecsoportra jellemzőek. Az 5. ábrán a likelihood hányados próba pontosságát és felidézését láthatjuk. A pontosság a kezdeti csökkenés után növekedésnek indul. Arra következtetünk, hogy nagyobb korpusz használatával a görbe szára még feljebb kúszna, vagyis még több helyes vonzatkeretet tudna a tanuló algoritmus kivonni a szövegből.



5. ábra. A likelihood hányados próba pontossága és felidézése a Szeged Korpuszon

Hivatkozások

1. Brent, M. R.: From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics 19, 2(1993) 243–262
2. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Proceedings of TSD 2004. Brno, vol. 3206 (2004)
3. 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. Association for Computational Linguistics, Prague, Czech Republic (2007) 209–212
4. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004) (2004)
5. Korhonen, A, G. Gorrell, McCarthy, D.: Statistical filtering and subcategorization frame acquisition. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong (2000) 199–206
6. Sass, B.: Extracting Idiomatic Hungarian Verb Frames. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.): Advances in Natural Language Processing. 5th International Conference on NLP, FinTAL, Turku, Finnország (2006) 303–309

7. Schulte im Walde, S.: The induction of verb frames and verb classes from corpora. In: Lüdeling, A., Kytö, M. (eds.): Corpus Linguistics. An International Handbook. Berlin, Mouton de Gruyter (2008)
8. Serény, A., Simon, E., Babarczy, A.: Automatic acquisition of Hungarian subcategorization frames. In: Hungarian Fuzzy Association 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI 2008), Budapest (2008) 443–454
9. 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. ELRA (2006) 1670–1673
10. Kornai A., Rebrus P., Vajda P., Halácsy P., Rung A., Trón V.: Általános célú morfológiai elemző kimeneti formalizmusa. In: Alexin Z., Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE Informatikai Tanszékcsoport, Szeged (2004) 172–176
11. Zeman, D., Sarkar, A.: Automatic extraction of subcategorization frames for Czech. In: Proceedings of the International Conference on Computational Linguistics (COLING '00)(2000) 691–697

VII. Poszter- és laptopos bemutatók

PACS: beszédvezérelt POI-kereső szolgáltatás

Csáki Tibor, Vajda Péter, Vámosi János

Ygomi Europe Kft.,
4034 Debrecen, Vágóhíd utca 2.
{csaki, vajda, vamosi}@connexis.com

Kivonat: Az utóbbi években a kéziszámítógépek elterjedése és a térképes alkalmazások fejlődése egyre több ember számára teszi mindennapossá a navigációs rendszerek használatát. Az autóban, vezetés közben használt rendszerek veszélye, hogy a vezető figyelmét könnyen megoszthatják. Ennek kiküszöbölésére egy olyan felhasználói felületet terveztünk, amely beszédfelismerésen alapul, így mentesíti a vezetőt a navigációs eszköz kézzel való használatától, valamint lerövidíti az információ megszerzéséhez szükséges időt. A felhasználói felület mögött komplex dialógusrendszer működik, ami szükség esetén emberi operátor beavatkozását is lehetővé teszi.

1 A PACS rendszer

A navigációs rendszerek fejlődése, a hordozható, GPS-sel felszerelt kéziszámítógépek, telefonok elterjedésével egyre többen vesznek igénybe telematikai szolgáltatásokat[1]. Ezzel a fejlődéssel párhuzamosan a felhasználók igényei is megnövekedtek: az útvonaltervezésen és a POI (Point of Interest) keresésen kívül ma már a helyfüggő, folyamatosan elérhető, valós idejű adatokat (pl. közlekedési információkat) kínáló szolgáltatások is mindennaposak.

Az autóvezetőknek nyújtott szolgáltatások egyik fontos kritériuma, hogy az alkalmazott eszköz a vezetők figyelmét a lehető legkisebb mértékben vonja el a forgalomtól a biztonságos közlekedés érdekében. Az autógyártó és telematikai cégek jelenleg kétféle megoldást alkalmaznak, azonban a fenti szempontból ma még ezek egyike sem kielégítő. Az egyik módszer az autóba beépített beszédfelismerő használata, amely még ma is egy drága megoldás és jelenlegi szintjén nem képes megbirkózni azokkal a nem ritkán komplex kérésekkel, amiket egy navigációs rendszer felé tehetünk. A másik megoldás az ügyfélszolgálati rendszereké, ahol a vezető egy telefonközpontossal kommunikálva jut a számára érdekes információhoz. Azonban ez a megoldás a telefonos operátorok alkalmazása miatt nem költséghatékony és a hívások időtartama, ezzel pedig költsége is nagyobb lehet, mint amit egy autóvezető elfogadhatónak tart.

A Ygomi Europe Kft-nél megvalósított PACS (People Assisted Computer System) rendszer egy olyan helyfüggő navigációs szolgáltatás folyamatát írja le, amely a fent

[1] Projektünket a Nemzeti Kutatási és Technológiai Hivatal a Jedlik Ányos program keretében támogatta (szerződésszám: OM-00102/2007).

említett két módszer előnyeit próbálja meg ötvözni. Tipikus felhasználási módja, amikor egy autóvezető útközben egy számára érdekes helyet (POI-t) keres, és ennek adatait szeretné letölteni gépkocsija navigációs rendszerébe. A PACS rendszer jellemzője a szerveroldali beszédfelismerés, és az automatikus beszédszintetizátorral (TTS-sel) előállított válasz, így a vezető egy rövid, géppel folytatott párbeszéd végén letöltheti autója navigációs rendszerébe a kívánt adatokat. Ugyanakkor a vezetőnek – ha szükséges – lehetősége van telefonos operátorral is felvenni a kapcsolatot, pl. ha bonyolultabb kérdést akar feltenni, vagy ha a beszédfelismerőnek nem sikerült egyértelműen felismerni a kérést. Utóbbi esetben a rendszer egy ún. Silent Agentet (néma operátort) használ, aki visszahallgatja a felvett kérést, és ha egyértelműen meg tudja válaszolni, akkor a megfelelő választ szintén TTS segítségével el tudja juttatni a vezetőhöz. Ennek a megoldásnak az előnye, hogy az operátornak nem kell beszélgetést folytatnia a klienssel, így a kérés kiszolgálása hatékonyabb és olcsóbb, mivel a „beszélgetés" időtartama rövidebb. A PACS korábbi prototípusait több európai autógyártóval is teszteltük. Ezekből a tesztekből kitűnik, hogy a hagyományos telefonközpontos megoldáshoz képest feleannyi az átlagos párbeszéd ideje, ha Silent Agentet használunk, illetve negyedannyi, ha az operátornak nem kell közbeavatkozni, azaz ha a beszédfelismerő helyesen működik [1]. Az operátornak csak akkor kell valódi párbeszédet folytatni az autó vezetőjével, ha a felvételből nem képes megállapítani a kérés tartalmát, vagy ha a vezetőnek az elsőre felajánlott válasz nem felel meg.

2 Megvalósítás

Az általunk bemutatott szoftver egy, a Ygomi Europe Kft.-nél elkészített PACS rendszert megvalósító demóalkalmazás, amely mobil kliensen keresztül is elérhető.

Az alkalmazás egyik fontos része a dialógusmenedzser, amely a tipikusan használt párbeszédeket vezérli. A dialógus véges állapotú automatával írható le, és a rendszer konfigurálásakor megadható. A dialógusok tervezésénél figyelembe vettük, hogy a vezető figyelmét minél kevésbé vonjuk el a vezetéstől. Ennek érdekében a dialógus úgy épül fel, hogy a gyakrabban előforduló kérések esetében a felhasználó kevesebb lépésben tudjon eljutni az eredményhez, valamint a rendszer által elmondott promptok hosszát is igyekeztük minimalizálni.

A rendszer fontosabb komponensei közé tartoznak a beszédfeldolgozást megvalósító modulok, ilyen a Carnegie Mellon University-vel (CMU) közösen fejlesztett Sphinx [3] beszédfelismerő rendszer, amelyhez nyelvi modelleket fejlesztettünk ki. A modellek építéséhez felhasználtuk a saját hanganyaggyűjtésből származó tipikusan előforduló kéréseket. Beszédszintetizátorként a Nuance RealSpeak nevű termékét használjuk, szemantikus elemzőként pedig a szintén a CMU-n készült Phoenix rendszert [2]. Ezek a nyelvi modulok más alkalmazásokra is kicserélhetők.

A rendszer több nyelvre készült el, lehetőség van a magyaron kívül angol, francia, német, spanyol és olasz nyelven is kéréseket megfogalmazni.

Hivatkozások

1. Masson, J.: Innovative Strategies For Improving Telematics Call Centre Operations. The Fully Networked Car Workshop ITU (2009)
2. Ward, W. H.: The Phoenix System: Understanding Spontaneous Speech.In: Proceedings of IEEE ICASSP (1991)
3. The CMU Sphinx Group Open Source Speech Recognition Engines http://cmusphinx.sourceforge.net

Jelentés-egyértelműsítés – egyértelmű jelentésítés?

Héja Enikő[1], Kuti Judit[1,2], Sass Bálint[1]

[1]MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
1068, Bp., Benczúr u. 33.
[2]ELTE BTK, Nyelvtudományi Doktori Iskola, Germanisztika Alprogram
{eheja, kutij, sass.balint}@nytud.hu

Kivonat: Az alábbiakban bemutatott esettanulmányunkban azt vizsgáljuk, hogy a magyar nyelvre létező, különböző típusú adatbázisok közül melyek mennyire alkalmas igei Wsd-célokra; emberi annotátorok között milyen mértékű egyetértést lehet elérni. A vizsgált adatbázisok az introspektív, illetve disztribúciós alapon készülő jelentéstárak közötti spektrumot hivatottak képviselni. Az eredmények arra utalnak, hogy a magyarra még nem létezik olyan, jelenlegi formájában késznek tekinthető jelentéstár, amely alapján kapott IAA-érték megfelelő viszonyítási alapot képezhet gépi WSD számára. A jelenlegi adatbázisok további, WSD-orientált fejlesztést igényelnek.

1 A feladat

A nyelvtechnológia egyik központi feladata megfelelő jelentés-egyértelműsítő rendszerek kialakítása. A jelentés-egyértelműsítés (a továbbiakban az angol terminus rövidítését használva WSD) számos alkalmazás számára elengedhetetlen; a legfontosabbak ezek közül a gépi fordítás, az információkivonatolás, illetve az információkinyerés. A jelentés-egyértelműsítés feladatát általánosan két alapvető lépésre bontjuk: (1) valamilyen jelentéstár kiválasztása, illetve létrehozása (2) a jelentéstárban szereplő jelentések hozzárendelése a kívánt szóalakokhoz valamilyen algoritmus segítségével. A jelentés-egyértelműsítéssel foglalkozó kutatások általában az utóbbira helyezik a hangsúlyt: azt vizsgálják, hogy a már létező jelentéstárakat milyen algoritmusokkal lehetne a lehető legjobban jelentés-egyértelműsítésre használni (pl. Latent Semantic Analysis (LSA), Hyperspace Analogue to Language (HAL).[1]

A jelentéstár kiválasztása, minőségének ellenőrzése ehhez képest minimális figyelmet kap. A legkülönbözőbb jelentés-egyértelműsítő részfeladatokhoz (célszó jelentés-egyértelműsítése, automatikus kulcsszókinyerés, szemantikai szerepek címkézése stb.) nagy százalékban a WordNet különböző verzióit használják jelentés-egyértelműsítésre, míg más jelentéstárak használata (pl. FrameNet, VerbNet) háttérbe szorul. (A Senseval versenyeken használt jelentés-egyértelműsített korpuszok több mint fele valamilyen WordNet-típussal lett annotálva). Köztudott ugyan, hogy a felhasznált adatbázisok jellemzően nem kifejezetten a jelentés-egyértelműsítés céljából

1 Agirre és Edmonds *Word Sense Disambiguation* c. könyvének alcíme nem véletlenül "Algorithms and Applications"

készülnek, a jelentés-egyértelműsítéshez szükséges adatbázisok szerkesztési elveivel és ezek létrehozásának módszereivel jóval kevesebb cikk foglalkozik, mint az algoritmusok tárházával.

A jelentés-egyértelműsítés mint komplex feladat, valamint a jelentések egzakt meghatározása sem tekintető tehát megoldottnak.[2] Ennek híján viszont jelentésmegkülönböztető adatbázisok előállításánál a fejlesztők sokszor elsősorban saját intuíciójukra vannak utalva.

Ebből fakadóan az enumeratív lexikonok WSD-feladatokra való alkalmassága is megkérdőjelezhető. Véronis *Sense tagging: Does it make sense?* c. cikkében [5] a jelentés-egyértelműsítés viszonylagos megoldatlanságáért az intuíción alapuló jelentéstárak és a szerkesztési mód miatt fellépő inkonzisztenciákat teszi felelőssé. A következetlenségek kikerülésének — szerinte — egyetlen módja, ha a lexikon létrehozása során, a jelentések elkülönítésénél elsősorban megfigyelhető disztribúciós jelenségekre támaszkodunk.

Az intuitív jelentésfogalom problematikus voltát, illetve az enumeratív lexikonok jelentés-egyértelműsítés céljára való alkalmatlanságát Véronis két kísérlettel próbálta meg alátámasztani. Az első kísérletben megmutatta, hogy az annotátorok közötti egyetértés már abban a kérdésben is alacsony (igék esetén 0.37), hogy egy szóalak egy- vagy többjelentésű-e. A második kísérletben 60 szó 3724 előfordulásához kellett hozzárendelni a kísérleti személyeknek a *Petit Larousse* értelmező szótárban felsorolt jelentések közül a kontextusnak megfelelőt. Az annotátorok között ebben a kísérletben is alacsony egyetértés volt (igék esetén 0.41), ami a feladat nehézségét támasztja alá.

2 A kísérlet

Esettanulmányunkban a fent említett második kísérletet végeztük el magyarra, kiegészítve azzal, hogy három különböző jelentéstár használatával nyert eredményeket hasonlítottuk össze. Kísérletünkben igei jelentésekkel foglalkoztunk. 15 ige 30 előfordulását annotáltattuk be 5-5 kísérleti személlyel a Magyar Értelmező Kéziszótár (a továbbiakban ÉKSz), a Magyar WordNet [3] (a továbbiakban HuWN) jelentéseivel, illetve az "Igei szerkezetek gyakorisági szótára" [4] (a továbbiakban ISZGYSZ) adatbázis igei szerkezeteivel. A kísérleti személyek az adatbázisokban megadott kategóriacimkék választásán kívül "nincs" és "nem tudom" választ is adhattak.

Utóbbi adattár automatikusan gyűjtött gyakori, különböző specifikusságú ige + főnévi csoport szerkezeteket tartalmaz a vonzatkeretektől a komplex igéken át a szólásokig (ige + esetragok / névutó + leggyakoribb lemmák). Az adatbázist előállító algoritmus szigorúan disztribúciós alapon gyűjt, és a bővítményi szavak eloszlásából képes megállapítani, hogy adott bővítmény kötött vagy szabad. Az adatbázis definíciókat nem tartalmaz, viszont minden szerkezethez ad korpuszból gyűjtött példákat,

2 Agirre és Edmonds így ír a célzott WSD-ről a *Word Sense Disambiguation* c. könyvük bevezetőjében: "... explicit WSD has not yet been convincingly demonstrated to have a significant positive effect on any application."

melyek a szerkezet jelentését hivatottak megvilágítani. A magyar igei WordNet egy, az angol nyelvű Princeton WordNet 2.0 verziójára épülő, de annak struktúrájához nem mereven ragaszkodó lexikális adatbázis, amelynek alapegysége a fogalom / jelentés, nem pedig a tradicionális szótárak alapegysége, a szó / lexéma. A magyar igei WordNet, amellett hogy a PWN-ben tárolt szemantikai relációkat kódolja mintegy 3000 fogalom között, néhány új reláció bevezetésének segítségével igyekszik lehetővé tenni az igék aspektuális jellemzőinek kódolását is. Fontos továbbá, hogy a magyar igei WordNet készítésekor már vonzatkeretekre vonatkozó, automatikusan kinyert információkat is figyelembe vettünk. Ezért a WordNet-beli jelentésmegkülönböz-tetések egyrészt nem pusztán introspekción, másrészt nem az angol nyelvű PWN jelentésmegkülönböztetésein alapulnak.[3] A HuWN ily módon módszertanilag az introspektív jelentésmegkülönböztetéseken alapuló ÉKSz., és a pusztán disztribúciós alapokon nyugvó ISZGYSZ között helyezkedik el.

Esettanulmányunkban azt vizsgáltuk, hogy a magyarban rendelkezésre álló jelentéstárak alapján milyen fokú egyetértést lehet emberi annotátorok között elérni, illetve hogy van-e az annotátorok közötti egyetértésben különbség az adatbázis fajtájától függően. Választ vártunk arra a kérdésre is, hogy milyen tulajdonságokkal kell rendelkeznie egy olyan jelentéstárnak, amelyet *kifejezetten* jelentés-egyértelműsítés céljából készítenek.

Az annotátorok közötti egyetértést (inter-annotator agreement — IAA) Fleiss-féle multi π érték szerint számoltuk, Artstein és Poesio (2008: 563-564) alapján [1]. A Cohen-féle κ–val szemben e mértéknek előnye, hogy képes elvonatkoztatni az egyes annotátorok esetlegességeitől. A Fleiss-féle multi π az összes annotátor adataiból becsült átlageloszlásból számolja a várható egyetértés mértékét, és azt mutatja meg, hogy a tapasztalt egyetértés hol helyezkedik el a várható egyetértés (0) és a teljes egyetértés (1) által meghatározott skálán. A mérték negatív értéket is felvehet, ha az egyetértés kisebb a véletlenszerűen elvártnál. Minél közelebb van tehát a kapott érték az 1-hez, annál nagyobb a valószínűsége, hogy az annotátorok közti egyetértés nem véletlen. A Fleiss-féle multi π érzéketlen olyan plusz kategóriákra, amelyeket soha egyetlen annotátor sem választott, azaz az eredményben nem jelenik meg, hogy hány kategóriából választhattak eredetileg az annotátorok.

3 Kiértékelés

Az alábbi táblázat az egyes adatbázisok szerinti IAA-értékeket mutatja az egyes igékre lebontva. A táblázatban szereplő eredmények számításakor a teljes érékű válasznak tekinthető "nincs" válaszokat önálló értékként kezeltük. Ugyanígy kezeltük a "nem tudom" válaszokat is, melyek előfordulási aránya mindössze 2-6% volt.

3 Az adatbázis építésének főbb módszertani lépéseihez ld. [3].

1. táblázat: A Fleiss-féle multi π mérték átlagolt értéke a három adatbázisra vonatkozóan.

	ÉKSz.	HuWN	ISZGYSZ	ÉKSz 2 (fő-jelentések)	választható jelentésszám **ÉKSz.** / *HuWN* / ISZGYSZ // ÉKsz2
emel	**0.450**	*0.753*	0.170	0.848	**13** / *10* / 16 // 5
feltesz	**0.493**	*0.693*	0.265	0.745	**14** / *7* / 8 // 7
fizet	**0.157**	*0.61*	0.259	0.278	**12** / *1* / 23 // 5
használ	**0.210**	*0.954*	0.336	0.611	**8** / *2* / 22 //
köt	**0.449**	*0.637*	0.237	0.535	**29** / *21* / 19 // 12
lép	**0.346**	*0.595*	0.443	0.601	**12** / *11* / 31 // 7
megold	**0.137**	*0.197*	0.255	0.449	**6** / *2* / 12 // 4
mutat	**0.187**	*0.153*	0.284	0.365	**13** / *4* / 27 //5
okoz	**0**	*0.59*	0.286	0	**2** / *3* / 26 // 2
rendelkezik	**0.195**	*0.469*	0.471	0.474	**6** / *3* / 15 // 4
segít	**0.112**	*0.371*	0.434	0.173	**7** / *4* / 19 // 5
szolgál	**0.279**	*0.516*	0.548	0.509	**15** / *8* / 16 // 7
tárgyal	**0.840**	*0.543*	0.407	0.840	**3** / *2* / 16
választ	**0.452**	*0.935*	0.444	0.713	**6** / *2* / 24 // 4
vállal	**0.207**	*0.311*	0.275	0.623	**6** / *3* / 26 // 3
átlagolt Fleiss-féle multi π	0.300	0.483	0.340	0.517	

A fenti adatok alapján a következő iránymutató következtetések vonhatók le: annotátorok közti egyetértés nagyságrendje összevethető Véronis kísérletének eredményeivel, minden adatbázis esetében. A szokásos (0.7-0.8) küszöbérték alapján egyik adatbázis szerinti IAA-érték sem lett olyan magas, amely alapján gépi WSD számára megbízható referenciakorpusz készíthető volna. A jelentéstárként használt adatbázis típusa nagyban befolyásolja, hogy milyen IAA-mértéket kapunk − jelenlegi állapotában az igei WordNet alapján kaptuk a legjobb értékeket, azt az ÉKSz-szerinti értékelést kivéve, amelyikben csak az ÉKSz fő jelentéscsoportjait vettük figyelembe, az ezek alatt meghatározott aljelentéseket (az adatbázis legfinomabb jelentésmegkülönböztetéseit) nem (l. a fenti táblázat jobbszélső oszlopát). A jelentések megkülönböztetésének finomsága, úgy tűnik, befolyásolja az IAA-értéket (ld a két ÉKSz-en alapuló IAA érték összehasonlítását (0.300 vs. 0,517)), ám pusztán az ige poliszémiájának mértéke nem tűnik relevánsnak az IAA-mérték szempontjából.

Véronis hipotézise a jelenlegi magyar nyelvre elérhető jelentéstárak alapján nem igazolható: a tisztán disztribúciós alapon készült ISZGYSZ jelenlegi formájában még nem válthatja fel a (legalább részben) introspektív alapon készült jelentéstárakat. Ennek okát az első kvalitatív elemzések alapján abban látjuk, hogy egyes annotátorok tisztán formai jegyek alapján rendeltek szóelőfordulásokhoz szerkezeteket, mások pedig az esetleges lemmákat, vagy akár az esetragokat / névutókat is szemantikailag

reprezentatív tartalommal töltötték meg. Az alábbi két tesztmondatban előforduló *emel* szó különböző annotációi jól illusztrálják ezt:

(1) Ezek az eredmények pedig az érdekképviseletek presztízsét emelik.
(2) A kipattanó labdát Makaay négy méterről a teljesen üres kapu fölé emelte.

Az (1) mondat esetében mind az öt annotátor különböző választ adott: hárman szemantikai értelmezés után az *emel magas-rA -t*, az *emel magas-bA -t*, *emel ár-A-t* kereteket választották, egy annotátor az *emel -t* keret mellett döntött, egy pedig "nincs" választ adott. A (2) mondat esetében egy annotátor az *emel -ba -t* szerkezetet választotta, hárman az *emel fölé -t* keretet, egy pedig "nincs" választ adott.

Esettanulmányunk alapján összegzésként elmondhatjuk, hogy a jelenleg rendelkezésünkre álló igei adatbázisok Wsd-céljára való alkalmazása további, WSD-orientált fejlesztést igényel. A továbbiakban az esettanulmány eredményeként kapott adatok további − kvalitatív és kvantitatív − elemzésével kívánjuk meghatározni, hogy milyen kritériumoknak kell egy WSD-célokra tervezett adatbázisnak megfelelnie.

Hivatkozások

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4) (2008) 555-596
2. Agirre, E., Edmonds, Ph.: Word sense disambiguation. Algorithms and Applications. (Text, Speech and Language Technology), Springer-Verlag New York, Inc., Secaucus, NJ, (2007)
3. Kuti J., Varasdi K., Gyarmati Á., Vajda P.: Hungarian WordNet and representation of verbal event structure. In: Acta Cybernetica, 18(2), (2007) 315-328
4. Sass B., Pajzs J.: FDVC - Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions for Hungarian. In: Abstracts of the eLexicography in the 21st century Conference, Louvain-la-Neuve, Belgium, (2009) 183-186
5. Véronis, J.: Sense tagging: does it make sense? In Wilson, A., Rayson, P. és McEnery, T. (Ed.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Frankfurt: Peter Lang (2003)

Jelentések gyakoriságának vizsgálata a Magyar WordNet-ben

Kiss Márton[1], Vincze Veronika[1], Alexin Zoltán[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{mkiss, vinczev}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
alexin@inf.u-szeged.hu

Kivonat: A WordNet strukturális felépítését és a Google keresőprogram szolgáltatásait felhasználva olyan kísérletet hajtottunk végre, amely vizsgálja a WordNetben előforduló szavak jelentéseinek gyakoriságát. A vizsgált szó jelentéseit a hiponímia – hipernímia relációkban lévő synsetek felhasználásával különbözteti meg (kiegészíti ezekkel a szavakkal a keresőkifejezést) és tárolja a Google által visszaadott becsült előfordulási számot. A megkülönböztetés eredményeképpen megállapítható, hogy egy adott jelentés relatív gyakorisága az összes jelentés előfordulására nézve. A kísérlet eredményeit összehasonlítottuk a SZTE Informatikai Tanszékcsoport által épített WSD korpuszban található jelentésgyakoriságokkal. E munkálatok fontos szerepet töltenek be egy magyar nyelvű jelentés-egyértelműsítő szoftver készítésében.

1 Bevezetés

Adott szó jelentései előfordulási arányainak meghatározásához a jelentés hiponímia és hipernímia relációkkal hivatkozott synseteket használtuk fel. Az alapötlet az volt, hogy egy jelentést meghatároz, ha a WordNet fastruktúrájában közeli (közvetlenül alatta vagy felette) elhelyezkedő synsetekkel fordul elő együtt egy weboldalon, egy dokumentumban.

2 A szópárok, kifejezések lekérdezése

A kutatás kezdetekor megvizsgáltuk azokat a módszereket, melyekkel nagy mennyiségű (napi több ezer, esetleg több tízezer) keresési eredményt lehet lekérdezni a Googletől. Négy megoldást vizsgáltunk, mely a Google kereső által visszaadott becsült találati számokat (ERC_{kif}) kéri le: HTML protokoll feletti lekérés, Google SOAP API, Google AJAX API, Google AJAX API használata HTTP protokoll felett. Erre az összehasonlításra azért volt szükség, mert a Google nem ad pontos találati számot, csak egy becslést közöl és ráadásul ez a becslés a különböző technikai megoldások-

ban sem egyezik meg. Ezeket a lehetőségeket összehasonlítottuk és kiválasztottuk a megfelelőt.

3 A jelentések gyakoriságának meghatározása

Meghatároztuk az A_{wi} tulajdonsághalmazokat, mellyel elkülöníthetünk egy adott jelentést. A w szóhoz tartozó A_{wi} tulajdonsághalmaz i: a WordNetben is használt jelentésindex; I: adott szó összes jelentésének halmaza; $i \in I$. Az A_{wi} *az* i jelentéséhez tartozó hiperním (w_{i_hip}) és hiponím (w_{i_hyp}) szavak, szókapcsolatok halmaza. Azzal a megkötéssel, hogy azon szavak vagy kifejezések, melyek a vizsgált w szó, más n jelentéseinél ($n \in I$ és $n \neq i$) is előfordultak, nem vettük figyelembe, tehát a közös ős- vagy gyerekhivatkozásokat kihagytuk.

A tulajdonsághalmazok meghatározása után lekérdeztük w szó összes $i \in I$ jelentését a jelentésekhez meghatározott A_{wi} tulajdonsághalmazban található összes szóval. A lekérdezésben használt kifejezés felépítése, tehát:

$$kif_n = w_i + n, \text{ ahol } n \in A_{wi} \quad (2)$$

Majd adott i jelentéshez tartozó becsült előfordulási számokat összegezzük:

$$w_{i\,ERC} = \sum_{n \in A_{wi}} \mathrm{ERC}_{\mathrm{kifn}} \quad (2)$$

Ezen értékek figyelembevételével már könnyen meghatározható volt adott w_i jelentés relatív gyakorisága.

4 A WSD korpusz és a jelentésgyakoriságok összehasonlítása

A WSD korpusz 39 synset szemantikus annotációját tartalmazza és minden synsethez 300-350 előfordulás található. A 39 synset jelentéseihez tartozó, a Google segítségével kapott relatív gyakoriságokat összehasonlítottuk a WSD korpuszban található gyakoriságokkal. A WSD korpusz a WordNet jelentéseinek felhasználásával készült, ugyanakkor a jelentések nem fedték egymást egy az egyben, így az összehasonlítás nehézkes munka volt, mert minden szóalak esetében meg kellett feleltetni a WordNet és a WSD korpusz jelentéseit egymásnak.

Az így összepárosított jelentések gyakoriságát vizsgáltuk a WSD korpuszban és a Google-ben. Az eredmények időnként egybevágtak a két módszert tekintve (pl. *század, jár*), más esetekben azonban a jelentésgyakoriságok éles eltérést mutattak (pl. *kormány*, *program*). Utóbbi jelenség valószínűleg a WSD korpusz tematikai egyöntetűségének köszönhető.

1. táblázat: A WSD korpusz és a kutatási eredményeink összehasonlítása.

szó:jelentés	korpusz %	Google %	szó:jelentés	kor-pusz %	Google %
jár: 3 volt	6,5	10.3	kormány: 1 irányítóeszköz	0,3	41,3
jár: 4 tánc	0,7	8.6	kormány: 3 biciklikormány	0,0	2,0
jár: 5 valahogyan	17,7	5.8	kormány: 2 szerv	98,4	56,7
jár: 7 működik	0,3	2.9	kormány: 3 egyéb	1,4	0
jár: 8 előfizető	0	10.4	program: 1 szabadidő	7,0	64,3
jár: 9 valakinek	12,3	30.0	program: 2 célok	74,8	0,2
jár: 10_együtt	18,1	8.6	program: 3_műsor	1,6	26,0
jár: 11 tartozik	1,3	18.9	program: 4 számítógép	16,5	9,5
jár: 12 egyéb	25,8	0	század_n_1 évszázad	99.7	97,0
jár: 13 valahol	1,0	1.3	század_n_2 katonai	0.3	3,0
jár: 14 valamiért	1,0	0			
jár: 15 közeledik	3,8	0			
jár: 16 valakivel	0,3	4.2			

5 Tervek, a kutatás folytatása

Kutatásaink célja egy magyar jelentés-egyértelműsítő rendszer előkészítése. Ehhez azonban szokásos módszerekkel tanuló korpuszt készíteni, amelyben az egyes jelentések megfelelő számban fordulnak elő, nem lehetséges. Mindenképpen olyan technikai megoldásokra van szükség, melyek az elérhető legnagyobb korpuszon (interneten) megtalálható dokumentumok alapján becslik meg az előfordulások gyakoriságát.

Hivatkozások

1. Szarvas György, Hatvani Csaba, Szauter Dóra, Almási Attila, Vincze Veronika, Csirik János: Magyar jelentés-egyértelműsített korpusz, Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Magyarország (2007), 158-165
2. Agirre et al.: Personalizing Page Rank for Word Sense Disambiguation, The First KYOTO Workshop, Amsterdam, Netherlands (2009)
3. Gabrilovich, Evgeniy, Markovitch, Shaul: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Israel, Haifa
4. Strube, Michael, Ponzetto, Simone Paolo: Genetic WikiRelate! Computing Semantic Relatedness Using Wikipedia, Heidelberg, Germany
5. Miháltz M.: Towards A Hybrid Approach To Word-sense Disambiguation In Machine Translation. In Proceeding Modern Approaches in Translation Technlologies Workshop at RANLP-2005, Borovets, Bulgaria (2005)

Szemantikai gráf alapú mondatelemző modul kidolgozása IS-NLI értelmezőhöz

Kovács László

[1] Miskolci Egyetem, Általános Informatikai Tanszék,
3515 Miskolc-Egyetemváros
kovacs@iit.uni-miskolc.hu

Kivonat: Az NLI lekérdező modulok egyik alapfeladata a természetes nyelven beérkező parancsok átkonvertálása a feldolgozó modul saját parancsnyelvére. Napjainkban az NLP transzformációs feladat megoldási módszerek között dominálnak a generatív vagy a statisztikai algoritmuson alapuló eljárások. A cikk egy fogalmi hálót mint közvetítő elemet tartalmazó NLP modul modelljét ismerteti. A kidolgozott rendszer a Dependency Grammar modellen alapul.

1 Bevezetés

Az információs rendszereket tekintve a természetes nyelvi feldolgozás (NLP, natural language processing) egyik legfontosabb alkalmazási területét az emberközeli lekérdező felületek jelentik. A lekérdező modulok egyik alapfeladata a természetes nyelven beérkező parancsok átkonvertálása a feldolgozó modul saját parancsnyelvére. A természetes nyelvi interfésszel rendelkező információs rendszerek gyökerei az 1970-re nyúlnak vissza. Az úttörő LUNAR projekt a holdkőzetek adatbázisában való lekérdezésekhez dolgozott ki NLI (természetes nyelvű interfész) felületet. A RENDEZVOUS (Codd, 1977) rendszer volt az első általános célú adatbázis NLI modul. Az NLP transzformációs feladat megoldási módszerek között dominálnak a generatív vagy a statisztikai algoritmuson alapuló eljárások. A generatív esetben, mely alatt most azt értjük, hogy a kódba beépítjük a két nyelv általunk fontosnak tartott szabályrendszerét és a meghatározzuk a két szabályrendszer közötti közvetlen konverziót. E módszer előnye az egzakt működés, a feltárt szabályrendszernek való pontos megfeleltetés. Hátránya viszont, hogy ismertnek és algoritmizálhatónak kell lennie az alkalmazott nyelvek nyelvtanának. Ekkor a transzformáció jósága a nyelvtan leképezés jóságától függ. A statisztikai módszereknél, melyek egyik leggyakrabban használt formája a Markov-modelleken alapuló módszerek, a tanítómintából kinyert valószínűségi szabályok alkotják a konverzió magját A legtöbb statisztika alapú módszernél vagy a szabad szöveges forrásokra építenek, vagy nyelvi annotációt alkalmaznak a tanulás hatékonyság javítására. A tapasztalatok azt mutatják, hogy az alapszöveg, a nyers szintaktika önmagában nem elegendő a nyelvtan hatékony feltárására. A nyelvtani annotáció más részről jelentős többletterhet jelent, és igen nagy tanító mintahalmazt igényel. A cikkben bemutatott módszer alapvonása, hogy a nyelvi konverziót egy köztes szemantikát leíró formalizmuson keresztül hajtjuk végre. A

szemantikai hálóval történő köztes tartalom annotáció egyik fő előnye a nagyobb rugalmasság a nyelvfüggetlenségben és így a különböző nyelvek közötti konverzióban. Az irodalomban viszonylag szerény az ilyen szemantikai mediátoron keresztüli parancsértelmezőre vonatkozó vizsgálatok száma, mivel ezen megoldás csak nemrég került a vizsgálatok központjába [10]. A dolgozat bemutatja a javasolt transzformációs modult, a nyelvi és a szemantikai reprezentációs alak közötti konverzió lépéseit. A modell egyik fontos eleme a megfelelő szemantika ábrázolásai mechanizmus kiválasztása. Elemzésünk azt mutatta, hogy a tradicionális szemantikai reprezentációk nem adnak kellő hatékonyságot a transzformációban, ezért egy HECG-nek elnevezett szemantikai háló került kidolgozásra a modulhoz. A HECG ontológia leírás és az interfész nyelvek közötti transzformáció több lépcsőben megy végbe. A konverzióhoz szükséges nyelvtani elemek deklaratívan adhatók meg mint működési paraméterek. A konverzió első fázisában a szemantikai gráfból egy szógráf képezhető. A szógráf jellegében igen sok közös vonással rendelkezik a word dependency graph nyelvi reprezentációval, mely mintegy átmenetet képez a szintaxis és a szemantika között. A leképzés második fázisában a szógráfból szavak szekvenciája generálódik. A kidolgozott séma mintarendszer keretekben működik és a hatékony implementáció kidolgozása esetén fontos alkalmazási területeket kiszolgálására válhat alkalmassá.

2 A szófüggőség alapú nyelvtan modellek

A nyelvtanok egyik elterjedt osztályozása szempontja, hogy mit tekintünk mondat egységnek: a szavakat, szóláncokat vagy a szavak közötti függőségi rendszert. A függőség alapú rendszerek fő jellemzője a fejfüggő asszimmetria és az a törekvés, hogy a fej és tagelemek közötti kapcsolatot szemantikai alapokon nyugvó függőségi relációkkal írjuk le. A függőségi nyelvtan (Dependency Grammar) modelljét a francia Lucien Tesnière [1] dolgozta ki. A modell alapegysége a stemma, amely a szavak között fennálló szintaktikai függőségi viszony grafikus reprezentációjának tekinthető. A modell értelmezésében az ige tekinthető a legmagasabb helyen álló szónak, amely felügyeli, vezérli az alatta elhelyezkedő kiegészítőket, csatolmányokat. A csatolmányok maguk is lehetnek összetettek, rendelkezhetnek saját csatolmányokkal.

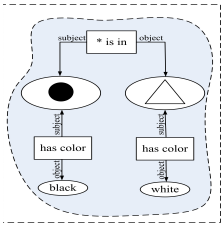
Tesnière elmélete jelentős hatással volt a nyelvészek széles táborára, azokra, akik a szemantika fontosságát a szintaktika elé helyezték. A [2] mű részletes áttekintést ad erről a területről.

Klein és Simmons [3] ezen függőségi nyelvtant alkalmazta gépi fordítást végző rendszerükre. A Valency elmélet [4] és Meaning-Text elmélet [5] néhány példái a mai is folyó függőségi nyelvtant támogató kutatásoknak. Schank is ezen irányból kiindulva alkotta meg a Conceptual Dependency Graph [7] modelljét, melynek sajátossága, hogy a háló elemei a szavak helyett a fogalmakat reprezentálják. A függőségi nyelvtan szerepét széles körben elemző mű a [8]. A függőségi nyelvtanok egy további lényeges bázisa a Case Grammar modell [9] is, melyben a függőségi élek címkézettek. Az Extensible Dependency Grammar [10] és a Word Grammar [6] olyan újszerű nyelvi modelleket képviselnek, melynek célja egy egységes modellbe egyesíteni a szemantikai és szintaktikai elemeket.

A függőségi nyelvtanok egyik előnyös tulajdonsága a magyar nyelv vonatkozásában, hogy a rendszer tudja kezelni a szabadsorrendű szerkezeteket is. Emellett lehetőség van nem folytonos szóláncot alkotó struktúrák kezelésére is. A DG alapú reprezentációban az élek így nemcsak szintaktikai szereppel bírnak, alkalmasak a szemantikai szerep jelölésére is.

3 A HECG szemantikai háló

A szemantikát, a jelentést megadó hatékony leírások közé tartozik a szemantikai háló, amely egy ontológiai modellt valósít meg. A szemantikai háló (Sloman 2003) egy olyan irányított gráf, melynek csomópontjai a fogalmakat reprezentálják és a köztük lévő élek a különböző relációkat jelölik. Az ontológia területe a problémák, a vizsgált világ fogalmi szinten történő leírásával foglalkozik. Az ontológiai rendszerek egyik lényeges vonása, amely megkülönbözteti őket a hagyományos szemantikai modellektől, hogy szabályalapú logikai kezelő nyelvvel is rendelkeznek. A mögötte álló következtető motor segítségével ellenőrizni lehet a modell konzisztenciáját, illetve új tények levezetését is biztosítani tudja a rendszer. Az ontológialeíró nyelvek között a két leginkább elterjedt nyelv az RDF és az OWL. Az RDF nyelvben az ábrázolás alapelemei körébe az erőforrások, a literálok és az állítások tartoznak. Az erőforrásoknak két fő típusa van: egyed és tulajdonság. Az állítás egy (p,s,o) hármassal adható meg, ahol a p egy tulajdonság, s egy erőforrás és o egy erőforrás vagy literál. Jelentését tekintve a p egy predikátumot, egy állítmányt takar. Az s szimbólum a szubjektum, az alany, míg az o az objektum, az érték. A RDF modellben az állítások vonatkozhatnak nemcsak elemi egyedekre, hanem más állításokra is. Az OWL nyelv az RDF nyelv kiterjesztésének tekinthető. A hozzátett új funkcióelemek köre magába foglalja az adattípus kezelést, a tulajdonság minősítését, a számosság ellenőrzést és egyéb új megszorítási elemeket.



1. ábra. ECG modell minta.

A kidolgozott HECG modell egy olyan fogalmi modellt jelöl, melyben a szerkezet építő elemei az egymásba foglalható elemi állításatomok. Egy elemi állítás magja az ige vagy predikátum. A predikátumhoz csatolható elemeket argumentumoknak nevezzük. Mind az élek, mind az elemek címkézettek, ahol a címke több elemi információt hordoz, mint a kapcsolat szemantikai tartalma, a kapcsolat megvalósulási megszorításai. A kapott fogalmi hálóból egy fókusz-állítás megadásával egy kapcsolati fa feszíthető ki, amely az elemek függőségi rendszerét is kifejezi. Az 1. ábra egy mintahálót mutat be.

4 A mintarendszer architektúrája, működése

A kidolgozott rendszer kétirányú konverziót valósít meg a HECG modell és egy szimbolikus nyelv között. A konverzió menete az alábbi alaplépésekre bontható fel:

- a hálóhoz a kijelölt predikátum alapján egy kifeszítő, függőségi fa generálása
- a fához egy szó-fa generálása, ahol a fogalmak mögé a hozzá csatolható szavak kerülnek be egy megadott tezauruszból
- a szavak módosítása az élekhez rendelt nyelvtani ragok alapján
- a szavakból a mondat generálása a sorrendiségi megszorításokat figyelembe véve.

A fordított irányú konverziónál elsőként a mondat elemeit határozzuk meg szavakra és morfémákra bontással. Az elemzés főbb lépései:

- Morfémaelemző segítségével a szavak szerkezetének feltárása
- A szavak morfémaelemzésével a ragok meghatározása
- A szótövek alapján a szó fogalmi kategóriáinak kijelölése
- A ragok alapján a kapcsolható argumentum élek kijelölése
- A szógráfhoz rendelt sorrendiség előírás összevetése a beérkező mondat sorrendiségével
- A vizsgált fogalomháló és a mintamondat távolság mértékének meghatározása
- A legközelebbi háló kiválasztása, mint a mondat jelentését reprezentáló háló.

A megadott algoritmus segítségével a mintarendszerben a magyar nyelv adott témakörhöz tartozó mondatait egy predikátum kalkulusbeli formára alakította, mely a későbbi lépésekben SQL vagy más nyelvre konvertálható tovább.

Hivatkozások

1. Tesnière, L.: Éléments de syntaxe structurale. Paris: Klincksieck (1959)
2. Sowa, J. F.: Semantic networks. In: Shapiro, S. C. (ed.): Encyclopedia of Artificial Intelligence. 2nd ed., Wiley. (1992)
3. Klein, S., Simmons, R. F. Syntactic dependence and the computer generation of coherent discourse. Mechanical Translation 7 (1963)
4. Hudson, D R.: Language Networks: The new Word Grammar. Oxford University Press (2007)
5. Mel'cuk, I. A.: Towards a linguistic ''Meaning ⇔ Text' model. In: Kiefer, F. (ed.): Trends in Soviet Theoretical Linguistics. Dordrecht, Reidel (1973) 35–57
6. Steele, J. (ed.): Meaning-Text Theory. Ottawa, University of Ottawa Press (1990)
7. McEnery, A., Xiao, R., Tono, Y.: Corpus-Based Language Studies: An Advanced Resource Book. In: Ser. Routledge Applied Linguistics. Routledge (2005)
8. Hudson, R.: Recent developments in dependency theory. In Jacobs, J., v. Stechow, A., Sternefeld, W., Vennemann, T. (eds.): Syntax. Ein internationales Handbuch zeitgenöössischer Forschung. Berlin, Walter de Gruyter (1993) 329–338
9. Fillmore, C. J.: The case for case. In: Bach, E., Harms, R. T. (eds.): Universals in Linguistic Theory. New York, Holt, Rinehart and Winston (1968) 1–88
10 Debusmann, R.: Extensible Dependency Grammar: A modular grammar formalism based on multigraph description. PhD thesis (2006)

Szekvenciajelölés gráfalapú, részben felügyelt tanulási módszerrel

Molnár Gábor József[1], Farkas Richárd[2]

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

Kivonat: A felügyelt tanulás fő problémája, hogy az egyedek kézi jelölése költséges és időigényes. Ez különösen igaz a szekvenciajelölés esetében, ahol egy tanítóhalmaz elkészítése több ezer token átvizsgálását igényli. Természetesen adódik az az igény, hogy olyan módszereket dolgozzunk ki, amelyek kevesebb tanítópélda ellenére is megfelelő modellt képesek építeni. Továbbá a klasszikus, szekvenciajelölésre használt algoritmusok kis méretű tanítóhalmazokon legtöbbször rosszul teljesítenek. Ezzel szemben a részben felügyelt tanulás éppen az előző igénynek próbál eleget tenni. Kísérleteinkben arra igyekeztünk rámutatni, hogy kis számú tanítópéldán alkalmazva a gráfalapú, részben felügyelt tanulási módszereket, azok jobb eredményt érnek el, mint a manapság gyakran alkalmazott szekvenciajelölők.

1 Bevezetés

Számos valós életbeli osztályozási probléma létezik, amelyekhez nem áll rendelkezésre megfelelő egyedszámú tanítóhalmaz. Az egyedek manuális jelölése gyakran költséges és időigényes. Ez különösen igaz a természetes nyelvi feldolgozás problémáinál, pl. a szekvenciajelölésnél, ahol gyakran több százezer tokenes tanítóadatbázisra van szükség. A probléma megoldására, a részben felügyelt tanulás módszere kínálhat megoldást.

Részben felügyelt esetben jelölt és jelöletlen példáink is vannak. Célunk a jelöletlen példák közötti mintázatok felismerésének segítségével, és a jelölt adatokból származó információ felhasználásával jelöléseket hozzárendelni a jelöletlen példákhoz. Azt várjuk, hogy ilyen módon kevesebb jelölt példa mellett is tanulható megfelelő pontosságú modell. A részben felügyelt tanulási technikákról egy kitűnő áttekintést ad [4].

A részben felügyelt tanulás egyik legfiatalabb részterületei a gráf alapú módszerek [1]. Ebben az esetben az egyedek alkotják a gráf pontjait, a gráf élei pedig a köztük lévő hasonlóságot reprezentálják. Ezeknél a módszereknél a kiértékelő adatbázist is felhasználjuk annak jelölései nélkül, hiszen a célunk nem az ismeretlen példákat jól klasszifikáló modell építése (induktív megközelítés), hanem a kiértékelő adatbázis felcímkézése (transzduktív megközelítés).

2 Szekvenciajelölés gráfok felhasználásával

Szekvenciajelölésen egy olyan osztályozási problémát értünk, ahol egyedek (tokenek) sorozatához (szekvenciához) rendelünk jelöléssorozatot. Tipikus szekvenciajelölési probléma a tulajdonnév-felismerés, ahol a mondatok szavait jelöljük be aszerint, hogy azok mely tulajdonnévosztályba tartoznak. Ebben a cikkben egy adott tulajdonnévosztályt jelöltünk szekvenciákban (bináris szekvenciajelölés), azaz mondatokban. A problémát a gráfalapú részben felügyelt tanulási paradigmába illesztve a gráf pontjainak a szekvenciák tokenjei felelnek meg. A tokenek között két éltípust különböztetünk meg: egyrészt, hogy megtartsuk a tokenek sorrendiségét és szekvenciához tartozását, az egyes tokeneket összekötöttük az őket megelőző és a rákövetkező tokennel; másrészt az előző pontban említett hasonlóság reprezentálására szolgálnak. Ez a módszer szoros összefüggésben áll a skip-chain CRF-fel [3], ami azt a tényt használja ki, hogy ha egy token többször fordul elő a dokumentumban, akkor az előfordulások nagy valószínűséggel ugyanabból az osztályból származnak, ezért kombinálja az azonos előfordulások jellemzőit, és olyan címkézésre törekszik, amely az ismétlődő tokeneket azonosnak tekinti. Ezzel szemben a gráfalapú részben felügyelt tanulási módszerek nemcsak az azonos előfordulások, hanem az aktuális tokenhez leghasonlóbb tokenek jellemzőit is képesek felhasználni azáltal, hogy a gráfban ezek a tokenek szomszédsági kapcsolatban állnak.

A gráf pontjainak felcímkézését egyszerű propagáló algoritmusokkal végeztük [1]. Propagálás során az a célunk, hogy a tanítópéldák jelöléseit eljuttassuk a szomszédos gráfpontokon keresztül a jelöletlen pontokhoz, a példákat összekötő élek súlyait figyelembe véve.

3 Módszer

KNN-gráfot használtunk, amelyben egy adott pontból csak a K leghasonlóbb szomszédba megy él. A KNN-gráf felépítésének időigénye (O(n*logn)) kisebb, mintha teljes gráfot építenénk fel ($O(n^2)$), és tárigénye is kevesebb ($O(n^2)$ helyett O(K*n)). Mindezek ellenére a KNN-gráf használata újszerű megközelítés, hiszen a publikált rendszerek jelentős része teljes gráfokat használ. Érdemes megjegyezni, hogy – ennek következményeként – a magukat kimondottan nagy adatbázisokon működőnek valló algoritmusok is csak néhány ezer pontra működnek elfogadható ideig [2].

A gráf pontjai közt értelmezett hasonlósági metrikát a Hamming-távolságból származtattuk: két token jellemzővektorát véve nem az eltérések, hanem az egyezések számát tekintettük. A gráf építése során a jellemzők súlyozásra kerültek az alapján, hogy az adott jellemző csak az osztályozandó tulajdonnevek (CC), a tulajdonnevek és nem tulajdonnevek (NC) vagy csak nem tulajdonnevek között fordul elő (NN). Minden jellemzőre összeszámoltuk, hogy hányszor fordul elő az egyes csoportokban. A gráf legközelebbi szomszédjának keresésekor két pont jellemzőinek metszetét véve a hasonlóságok (w) megadására kétféle módszert használtunk:

1. Csak azokat a jellemzőket vettük számításba, amelyek a tulajdonnevek között szerepeltek:

$$w = CC. \quad (\mathbf{0})$$

2. A hasonlóságot az alábbi csoportok gyakoriságát felhasználó képlet segítségével adtuk meg:

$$w = CC*(1-NC)*(1-NN). \quad (\mathbf{1})$$

Az általunk használt algoritmus a label propagation volt [1]. Ez minden iterációban frissíti a pontok címkéjét a következő képlet szerint:

$$y_i^{t+1} = \frac{\sum_{j \in K_i} \left(w_{ij} * y_j^t \right)}{\sum_{j \in K_i} w_{ij}}. \quad (\mathbf{2})$$

(2)-ben y_i^{t+1} jelöli az i. pont címkéjét a (t+1). iterációban; K_i az i. pontból kimenő élek végpontjainak halmazát; w_{ij} pedig az i. pontból a j. pontba tartó él súlyát. Az iterációk után minden ponthoz y_i szerint rendelünk címkét.

Propagálás során a gráf élsúlyait a szomszédos pontok címkéjének a.priorijával is normáltuk (CMN) [1]. Ezzel a módszerrel igyekeztünk kiküszöbölni azt a problémát, hogy a szekvenciákban előforduló pozitív példák (tulajdonnevek) száma lényegesen kisebb, mint a negatív példáké. Az alábbi képlet szerint normalizáltunk:

$$\hat{w}_{ij} = w_{ij} + \frac{\lambda}{p\left(y_j\right)}. \quad (\mathbf{3})$$

(3)-ban $\hat{w}_{ij}$ az élsúly normalizáció utáni értékét; $p(y_j)$ az y_j–nek megfelelő címke a.priori értékét jelenti; λ pedig egy normalizációs tényező, ahol $\lambda \in [0;1]$.

4 Kísérletek, tapasztalatok

Kísérleteinket a Reuters hírkorpuszon végeztük. A korpusz tanítóhalmazának 3000 pontját választottuk ki véletlenszerűen. A tanítóadatbázisban négyféle tulajdonnévosztály került felcímkézésre (személyek, szervezetek, helyek, egyéb). Egy teszteset alatt csak egy adott osztályra fókuszáltunk, a többi osztályba tartozó tokeneket ekkor nem kezeltük tulajdonnévként. A tesztekhez a korpusz kiértékelő adatbázisának 3000 véletlenszerűen választott pontját használtuk fel. A gráfban a K értékét 10-nek választottuk meg. Az algoritmusok kiértékelése során használt referenciaalgoritmusnak a CRF valószínűségi tanulót használtuk [3]. Kétféle paraméterrel kísérleteztünk:

1. A mondatokon belül szomszédos tokenek közötti éleket (tokenélek) súlyoztuk egy konstans értékkel.
2. CMN esetén a λ értékét változtattuk

A referencia legjobb eredményei a személynevekre adódtak. Ebben az esetben 62.1%-os F-measure értéket kaptunk. A legrosszabbul pedig a szervezeteket címkézte fel a CRF, ahol az F-measure 2.7%-os lett. Az eltérés valószínűleg a tanítóhalmazban található pozitív példák száma miatt tapasztalható.

A részben felügyelt tanulás eredményei 100 iteráció után a következőképpen alakultak erre a két osztályra. Személynevek esetén a címkepropagálás szignifikánsan alulmaradt a CRF-fel szemben. A legjobb eredményt (F-measure = 19.5%) akkor értük el, amikor a tokenélek súlyait 0.5-re, a CMN normalizációs tényezőjét pedig 0.0-re állítottuk. Utóbbi azt jelenti, hogy a CMN normalizáció egyáltalán nem segített a személyneveknél. Bár a szervezetek esetében a gráfalapú módszerek legjobb eredménye csupán 4.1%-os F-measure-t eredményezett, a CRF-hez képest mégis javulást értünk el. Ebben az esetben a CMN segített az eredmény javításában, a λ értéke 0.05 volt; a tokenélek súlya pedig az előző esethez hasonlóan 0.5.

5 Konklúzió

Összességében azt mondhatjuk, hogy bár a gráf alapú módszereink kis adatbázisok esetén bizonyos esetekben jobban működnek, mint a szekvenciajelölők, a legtöbb esetben ez nem mondható el azok jól megfogalmazható matematikai háttere ellenére. Ezért további kísérleteket folytatunk a CMN-nel történő normalizálásra és a tokenélek súlyának nem konstans értékű megválasztására. A jövőbeli terveink között szerepel továbbá a K értékének és a címkézett pontok száma hatásának vizsgálata.

Köszönetnyilvánítás

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

Hivatkozások

1. Chapelle, O., Schölkopf, B., Zien, A. (eds.).: Semi Supervised Learning. 11. fejezet, The MIT Press (2006)
2. Farkas R.: Részben felügyelt tanulási módszerek a tulajdonnév felismerésben. In: V. Magyar Számítógépes Nyelvészeti Konferencia (2007) 166-176
3. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. In: Getoor, L., Taskar, B. (eds.): Introduction to Statistical Relational Learning, The MIT Press (2007)
4. Zhu, X.: Semi-Supervised Learning Literature Survey. Technical Report Computer Sciences 1530, University of Wisconsin-Madison (2005)

Szintaktikai elemzés szerepe a biológiai eseménykinyerés kulcsszavainak detektálásában

Móra György[1], Molnár Zsolt[2], Farkas Richárd[3]

[1] SZTE, Informatikai Tanszékcsoport,
H-6720 Szeged, Árpád tér 2.
gymora@inf.u-szeged.hu
[2] Acheuron Hungary, Kemo- és Bioinformatikai Csoport,
H-6720 Szeged Tiszavirág u. 11.
zsoltm@acheuron.hu
[3] SZTE, MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
H-6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: Angol nyelvű élettudományi cikkekben szereplő biológiai események kulcsszavainak meghatározásához felhasznált hat nyelvi elemzőt hasonlítottunk össze. Biológiai esemény a szövegben leírt biológiai tény, folyamat. Az esemény kulcskifejezése az eseményt indukáló szövegrész, argumentumai a szövegben található biológiai entitások, mint például fehérjék, gének. A szövegből nyert statisztikai és nyelvi jellemzők felhasználásával döntési fa és szupport vektor gépi osztályozókat tanítottunk. A modellek teljesítménye közvetve információval szolgál az adott nyelvi elemző felhasználhatóságáról a kulcsszókinyerési feladaton.

1 Bevezetés

A tudományos publikációkban rejlő hasznos információk megszerzése sokszor komoly problémát jelent az információáradattal küzdő kutató számára. A biológia és az élettudományok területén ezért egyre nagyobb igény mutatkozik olyan információkinyerő rendszerekre, amelyek a publikációkból (szabadalmak, újságcikkek, konferenciakiadványok) tényeket, adatokat nyernek ki kereshető, strukturált formában. Az elmúlt években az érdeklődés fókusza az interaktáló fehérjepárok azonosításáról az összetettebb, részletesebb adatok kinyerésére tevődött át.

Az ún. *biológiai események* nem csak kétszereplősek lehetnek, egy vagy akár több fehérje is szerepelhet egy eseményben. Emellett az események más eseményekre is hivatkozhatnak, komplexebb tudásbázist létrehozva. A biológiai események jóval pontosabb adatokat tartalmaznak a biokémiai, sejtbiológiai folyamatokról, történésekről, mint a fehérje-interakciók, így értékesebb, piacképesebb adatbázisokat lehet építeni belőlük.

A gépi tanuláson alapuló eseménykinyerő rendszerek fejlesztése a GENIA Event Corpus [2] megjelenéséhez köthető, amely az első komplex biológiai eseményeket tartalmazó manuálisan annotált korpusz. A BioNLP2009 Shared Task on Event

Extraction elnevezésű eseménykinyerési verseny [1] volt az első, amely ezt a problémát tűzte ki feladatául.

Egy teljes biológiai esemény egy, az eseményt indukáló kulcskifejezésből, a résztvevő entitásokból és az őket összekötő esemény típusából áll. Az entitások fehérjék, gének és egyéb molekulák nevei. A versenyen ezen entitások szövegbeli előfordulásait ismertnek tekintették. A résztvevő rendszerek túlnyomó többsége két részfeladatra bontotta a problémát. Első lépésben az indukáló kulcsszavakat azonosították, majd szereplőket rendeltek ezekhez az entitáshalmazból. A verseny tapasztalatai, valamint a legjobban teljesítő rendszerek eredményei alapján megállapítható, hogy a függőségi és más szintaktikai elemzők kimenetéből nyert jellemzők jelentősen javítják a gépi tanulási rendszerek teljesítményét.

Cikkünkben a kulcskifejezés azonosításának részproblémájára koncentrálunk és négy különböző szintaktikai elemző kimenetének felhasználásával nyert jellemzőkészletet hasonlítunk össze. A gépi tanuló modellek teljesítményét értékelve a verseny által biztosított adathalmazokon, megállapítható, mely elemzők vagy melyek kombinációja adja a legjobb eredményt, illetve tárgyaljuk az egyes elemzők (melyeknek elméleti alapjai is különböznek) előnyeit, hátrányait, alkalmazhatóságuknak feltételeit. A jellemzőkészlet a szintaktikai és függőségi elemzők eredményein kívül a szavak más egyéb tulajdonságait is tartalmazza, ám ezek minden elemző esetében megegyeznek. A jellemzőkészlet mintájául a BioNLP2009 Shared Task on Event Extraction verseny első helyezettjének kulcskifejezés jelölő rendszere szolgált [3].

A különböző elemzőknek a feladaton elért eredményeit összehasonlítva láthatóak azok előnyei, illetve hátrányai a kulcsszódetektálásban. Az eltérő nyelvészeti megközelítések különböző összefüggések kinyerésére alkalmasak, ezért is fontos a feladatnak megfelelő kiválasztása.

2 Nyelvi elemzők

A vizsgált nyelvi elemzők két csoportba sorolhatóak. A függőségi elemzők (*dependency parser*ek) a mondat szavai közötti kapcsolatokat függőségi fa formájában ábrázolják. A fa minden pontjához egy szót rendelnek – amelyeknek pontosan egy őse van –, kivéve a virtuális gyökérelemet. A pontok és őseik közötti élek, valamint ezek címkéi definiálják egy mondat szerkezetét. Szabad szórendű nyelvek elemzésére különösen alkalmas, lévén a fa szerkezete a szavak sorrendjétől nem, csak a közöttük lévő nyelvi kapcsolattól függ.

A másik csoportba a frázisstrukturált nyelvtant használó elemzők tartoznak, amelyek a mondatokat hierarchikus formában, konstituensfaként írják le. A csomópontok igei, főnévi, stb. nyelvi csoportokat jelentenek, a fa gyökerében a mondatot reprezentáló pont van. Két egymás melletti csoport alkothat egy magasabb szintű csoportot, így a szavak sorrendjétől is függ, mely szavak képezhetnek egy csoportot. A fa pontjai nem a mondat szavainak felelnek meg, mint a függőségi fánál, hanem a mondatot alkotó hierarchikus szerkezeteket jelölik. A függőségi formátumtól eltérően itt a pontok címkéi tartalmazzák a felhasználandó információt, az élek az egyes csoportok elemeit, azok felbontását adják meg hierarchikus formában.

A *PCFG (Probabilistic Context-Free Grammar)* elemzők környezetfüggetlen nyelvtan segítségével elemzik a mondatokat. Az egyes csoportok valószínűségeit kombinálva határozzák meg a szöveg legvalószínűbb konstituens elemzését.

A *HPSG (Head-driven phrase structure grammar)* elemzők összetett, struktúrált "szótárak" és szabályok alapján építik fel a frázisok hierarchiáját. Minden frázisnak van egy feje, amely kitüntetett szerepű a kifejezés felépítésében. A szavak és frázisok tulajdonságait egymásba ágyazódó hierarchikus kulcs-érték párok adják meg. Ez a frázisstruktúra felbontható a beágyazások mentén, és faszerkezetben ábrázolható.

A cikkben felhasznált nyelvi elemzők:

3. ***Bikel:*** *Mike Collins* függőségi elemzőjének *Dan Bikel* által implementált változata
4. ***CCG:*** A *C&C Tools* függőségi elemzője biológiai doménre
5. ***Enju:*** Valószínűségi *HPSG* modellt használó szintaktikai elemző. Akár több lehetséges elemzési kimenetet is generál a valószínűségeik sorrendjében. A felhasznált változatot a *GENIA* korpuszon tanították.
6. ***Gdep:*** A *KsDep* függőségi elemző *GENIA* korpuszon újratanított változata
7. ***McClosky-Charniak:*** *Charniak és Johnson statisztikai* elemzőjének *David McClosky* által továbbfejlesztett biológiai doménre adaptált öntanulást alkalmazó változata
8. ***Stanford:*** *PCFG* elemző, frázisstruktúrált és függőségi formájú kimenettel.

3 Kulcsszavak detektálása

A biológiai események az élettudományi cikkekben szereplő valamilyen biológiai tényt vagy folyamatot írnak le. Az eseményeket jelző szövegrészlet az esemény kulcskifejezése. Az egyszerű statisztikai modellek helyett a nyelvtani elemzőkkel előállított, a szavak mondatban betöltött szerepét leíró jellemzők használata válik elterjedtté [1]. A kulcsszavak meghatározása osztályozási feladatként, gépi tanulási módszerek segítségével történt. A tanítóadatbázis a versenyen kiadott *train* halmaz volt, míg a kiértékelést a *development* halmazon végeztünk. Az *infogain* alapján leszűrt kétezer legjobb jellemzőn tanított *C4.5* döntési fa (*Weka J48*) és az összes jellemző felhasználásával tanított szupport vektor modellek (*libsvm*) eredményeit mértük meg.

A jellemzőkészlet mintájául a BioNLP2009 SharedTask on Event Extraction versenyen legjobban szereplő rendszer kulcsszódetektáló rendszere szolgált. A jellemzők három nagyobb csoportra oszthatók:

- **Token jellemzők:** A mondatok szavakra bontását a GeniaTagger tokenizálójával végeztük. A jellemzőkészlet tartalmazta a szavak gyökerét, amit a *Porter stemmer* állított elő, a szavak karakterenként vett bi- és trigramjait.

- **Numerikus jellemzők:** Ezek a jellemzők a szó adott tokenszámú környezetében és a mondatban található biológiai entitások számát, a mondatban található egyedi szavak számát adják meg.
- **Nyelvi jellemzők:** A nyelvi jellemzőket a függőségi fa az adott szóból kiinduló 1-3 mélységű útvonalai és az útvonalak végén található szavak mondatbeli funkciói alkották. A frázisstruktúrált elemzők kimenetét a függőségihez hasonló formában használtuk fel. Az *Enju* kivételével az összes ilyen elemző kimenete rendelkezésre állt *Stanford függőségi formátum*ra alakítva.
- **Szomszédos szavak:** Minden szóhoz a nyelvi fában a szülő szó, a gyerek szavak, illetve a szavak közvetlen környezetében található tokenek összes tokenjellemzőjét hozzárendeltük.

4 Eredmények

Jelen munkában a különböző nyelvi elemzők használhatóságát az általuk előállított jellemzők felhasználásával tanított kulcsszódetektáló modellek eredményeivel jellemezzük (1. táblázat). A C 4.5 modell kiértékelését keresztvalidációval is elvégeztük. Az elemzők nagy része biológiai doménre készült, de vannak közöttük általános szövegen tanítottak is. A kis eltéréseket és a magas pontosságot az okozza, hogy a szavak csak kis aránya kulcsszó, így a "nem kulcsszó" osztály előfordulása magas.

A keresztvalidáció során az elemzők nem mutattak jelentős eltérést, de a *dvelopment set*-en a *Stanford parser* teljesített legjobban a döntési fa modellel, a tanító adatbázis relatív méretének csökkenésével javult a teljesítménye.

1. táblázat: A különböző nyelvi elemzők teljesítménye a kulcsszó-meghatározási feladaton.

	Bikel	**CCG**	**Enju**	**GDep**	**M-C**	**Stanford**
C 4.5	96,696	96,660	96,655	96,783	96,681	**96,925**
C 4.5 k.v.	97,618	97,638	97,583	97,635	97,645	97,617
libSvm	96,730	**96,804**	96,450	96,552	96,635	96,408

Köszönetnyilvánítás

A kutatást – részben – a BAROSS_DA07-DA_TECH_07-2008-0028 projekt támogatta.

Hivatkozások

1. Kim, J-D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extractionin Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop (2009)
2. Kim, J., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature, BMC Bioinformatics (vol. 9) (2008)
3. Bjorne, J., Heimonen, J., Ginter, F. Airola, A., Pahikkala, T., Salakoski, T.: Extracting Complex Biological Events with Rich Graph-Based Feature Sets, BioNLP2009 Workshop Companion Volume for Shared Task Association for Computational Linguistics (2009)

Kutatók honlapjainak automatikus osztályozása pozitív és jelöletlen tanulás módszerével

Nagy István[1], Farkas Richárd[2]

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

1 Bevezetés

Az utóbbi években a kutatók kapcsolatainak feltérképezése és feldolgozása igen intenzíven kutatott területté vált [1]. Egyes kutatók weboldalán számos hasznos életrajzi információ található, úgymint a témavezetők vagy diákok neve, érdeklődési kör, nemzetiség, affiliációk, tudományos fokozatuk stb. [2]. Ezen adatok normalizált változatainak segítségével könnyen feltárható a kutatók közötti kollegiális kapcsolat, (az egy időben és helyen együtt dolgozók) ami nagyban különbözhet az együtt publikálóktól. Mindazonáltal lehetőség nyílik az olyan jellegű kérdések megválaszolására, mint hogy *az amerikai vagy az európai kutatók változtatják gyakrabban a munkahelyüket*.

Az ilyen jellegű feladatok megoldására használt webbányász rendszerek az internet redundanciáját használják ki [3], vagyis azon az elméleten nyugszanak, miszerint minél hasznosabb egy információ, annál többször fordul elő a weben. Ezért használhatóak olyan pontosságra optimalizált algoritmusok, amelyek automatikusan képesek összegyűjteni az adatokat, ugyanakkor nem céljuk az adott információnak az összes elérhető forrásból való kinyerése. Az egyes kutatókról elérhető életrajzi információ sok esetben csak a saját honlapjukon férhető hozzá, ezért, ellentétben a jelenleg alkalmazott megoldásokkal, elengedhetetlen ezen adatok minden esetben való felkutatása [2].

Ebben a cikkben olyan megoldásokat ismertetünk, amelyek automatikusan képesek azonosítani az egyes kutatókhoz tartozó oldalakat. A probléma nehézségét az adja, hogy egy egyszerű webes keresés eredményeként gyakran előfordulhat, hogy a találati lista számos irreleváns oldalt tartalmaz. Ennek egyik lehetséges oka lehet: egy, a keresett kutatóval azonos nevű színész, politikus, esetleg sportoló honlapja kerül a találati listába. Ugyanakkor nehézséget jelenthet az adott kutató által írt könyveket, publikációkat ajánló oldalak kiszűrése is. Ezért az egyes kutatók internetes oldalainak azonosítása érdekében a kereséshez használt online keresők eredményeit automatikusan, „kutató honlap” és „irreleváns honlap” csoportokba kell sorolni. A probléma megoldásához az utóbbi években igen intenzíven kutatott, *pozitív és jelöletlen mintából tanulás* standard módszereit és néhány általunk megkonstruált algoritmust mutatunk be.

2 Kutatók honlapjainak automatikus azonosítása

Amióta az internet különböző információk óriási adatbázisává vált, a honlapok automatikus osztályozása vagy kategorizálása igen intenzíven kutatott terület lett. A probléma megoldására adott legígéretesebb megközelítések a pozitív és jelöletlen tanulás valamely változatát alkalmazták, melyeknek legfőbb előnye a klasszikus osztályozókkal szemben, hogy a tanulás során nincs szükségük negatív példákra.

Az egyes kutatók honlapjainak az azonosítása során (azok kiválasztása a webes keresés találatai közül) 89 kutató letöltött honlapján, annotátorok által előzetesen bejelölt affiliációkat tartalmazó korpuszt használtuk. Amennyiben egy oldal tartalmazott jelölt affiliációt, akkor azt pozitív példának tekintettük, egyébként negatívnak. Az így kialakított dokumentumhalmaz 177 pozitív és 229 negatív példát tartalmazott. Az osztályozáshoz a modell által kialakított nagydimenziós térben is hatékony döntési fákat alkalmaztuk. Ezen megközelítés legnagyobb előnye, hogy az ember számára könnyen értelmezhető outputot generál, ráadásul éppen diszkrét jellemzők feldolgozására fejlesztették ki.

Az adott feladatot hatféleképpen oldottuk meg, melynek eredményeit az első táblázat hivatott illusztrálni. A korpuszt, a szövegbányászati modellek első, és egyben egyik legszélesebb körben használt dokumentum reprezentációs eszközével, a vektortérmodellel illusztráltuk. A különböző megközelítések alapvetően az egyes dokumentumokat leíró vektorokban különböztek. Ennek alapvető oka, hogy megpróbáltuk különböző tartalom alapján elvégezni a honlapok osztályozását. Az első táblázat első sorában egy dokumentumot a hozzátartozó URL és az abból kialakított n-gramok illusztrálják. A második, harmadik és negyedik sorban egy online keresés során elérhető snipet információk segítségével reprezentáltuk a teret. Az utolsó két sorban az adott honlap teljes szöveges tartalma és a hozzá tartozó webcím jelentette a reprezentáció alapját.

1. táblázat: Kutatók honlapjainak osztályozása.

Megközelítés	Pontosság	Fedés	F-mérték
URL	0,785	0,786	0,786
Snipet + URL	0,763	0,764	0,763
Snipet	0,828	0,828	0,826
Snipet + szűrők	0,845	0,845	**0,845**
Honlap + URL	0,79	0,791	0,79
Honlap + URL + szűrők	0,853	0,852	**0,852**

Az első táblázatban jól látható, hogy a keresés során elérhető snipet adatok és a honlapok teljes tartalmát különböző szűrőkkel és az URL-lel kiegészítve sikerült a legjobb eredményt elérni. Ennek megfelelően a későbbiekben ezen megközelítések eredményeit ismertetjük.

A pozitív és jelöletlen példákból való tanuláshoz a fentiekben leírt korpuszt alkalmaztuk. Minden negatív dokumentumot, valamint minden második pozitívot „*jelöletlen*” címkével láttunk el. Az így kialakult dokumentumhalmazt még kiegészítettünk további 30 kutatóhoz tartozó csaknem 200 újonnan letöltött dokumentummal, amik

szintén „*jelöletlen*” címkét kaptak. Ugyanakkor a kiértékelés természetesen az eredeti korpuszon történt.

2. táblázat: Pozitív és jelöletlen tanulás eredményei.

Algoritmus	Pozitív F (honlap)	F (honlap)	Pozitív F (snipet)	F (snipet)
PEBL	0,25	0,68	0,61	0,26
PEBLII	0,62	0,57	0,62	0,57
Tf-idf PEBL	0,65	0,62	0,62	0,57
Rocchio	0,61	0,26	0,61	0,26
Rocchio-Cluster	0,61	0,26	0,61	0,26
Rocchio PEBL	0,60	0,55	0,63	0,57
Spy	0,43	0,69	0,42	0,71
Modosított PEBL	**0,78**	**0,806**	**0,72**	**0,745**
Szavaztatás	**0,76**	**0,769**	**0,82**	**0,837**

A második táblázatban a tanulás pozitív és jelöletlen példákból különböző népszerű [4, 5, 6] és ezek általunk módosított algoritmusainak eredményei láthatók. A második és harmadik oszlopban a honlapok teljes szöveges tartalmából és a hozzájuk tartozó internetcímből képzett n-gramokból álltak az egyes dokumentumokat leíró vektorok, míg a harmadik és negyedik oszlopok csak a keresés során elérhető snipet adatokat tartalmazták.

A pozitív és jelöletlen tanulás egyik első, úttörő algoritmusa a PEBL (más néven 1-DNF vagy M-C) [4]. A megközelítés lényege, hogy a pozitív halmazban leggyakrabban előforduló szavak kigyűjtése után, azokat a dokumentumokat jelöljük negatívnak a jelöletlen halmazból, amelyekben egyetlenegyszer sem fordult elő ezen szavakból. Hátránya, hogy gyakran egyetlen dokumentumot sem jelöl negatívnak (a snipet esetben is így történt). Éppen ezért a PEBLII algoritmusnál [5] könnyítettek a feltételeken. Ebben az eseten akkor kerül be egy szó a pozitív szólistába, ha annak frekvenciája nagyobb a jelöletlen halmazbelinél, ugyanakkor meghalad egy bizonyos értéket. Az általunk kidolgozott tf-idf PEBL esetében, hogy az adott problémára minél inkább jellemző szavak kerüljenek a pozitív szólistába, ezért a mindkét halmaz tf-idf súlyozása után, szintén azok szavak kerülnek kiválogatásra, amelyek frekvenciája magasabb a pozitív halmazon. Mindhárom algoritmus igen hatékonynak bizonyul, amennyiben sikerül a helyes paraméterezést beállítani. A Rocchio algoritmus lényege, hogy az egyes tf-idf súlyok és a halmazok alapján minden csoporthoz egy-egy középpontot határoz meg, és az egyes elemeket ezekhez rendeli. A Rocchio-Cluster az előző megközelítés egy finomítása, miszerint a jelöletlen halmazt összefüggő csoportokra bontjuk, majd minden egyes halmazhoz meghatározzuk a középpontokat. A Spy megközelítés [6] lényege, hogy a pozitív példák egy részét a jelöletlenek közé másoljuk, ezáltal megkönnyítve a jelöletlen halmazban a pozitív dokumentumok „leleplezését”. Az általunk megvalósított módosított PEBL algoritmus lényege, hogy a pozitív szólistába egészen addig kerülnek bele a jellemző szavak, amíg a kezdeti negatív halmaz mérete meg nem egyezik a pozitívéval. A Rocchio PEBL algoritmus negatív középpontját a módosított PEBL megközelítés által kijelölt halmazon számoljuk ki, ezáltal az távolabb kerül a pozitív középponttól. Végül a szavaztatás megköze-

lítés esetében akkor kerül egy elem a kezdeti negatív halmazba, ha a Spy, Rocchio vagy a módosított PEBL algoritmusok közül legalább kettő negatívnak jelölte.

A második táblázatban jól látható, hogy az általunk megvalósított és módosított algoritmusok érték el a legjobb eredményeket. Továbbá a jelen feladat során kiemelten fontos pozitív fedésben is a legjobbak közt teljesítettek.

Köszönetnyilvánítás

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

Hivatkozások

1. Said, Y. H., Wegman, E. J., Sharabati, W. K., Rigsby, J. T.: Social networks of author-coauthor relationships. Computational Statistics & Data Analysis, 52(4) (2008) 2177–2184
2. Nagy, I., Farkas, R., Jelasity, M.: Researcher affiliation extraction from homepages. In: Proceedings of the NLPIR4DL Workshop at ACL (2009)
3. Califf, M. E., Mooney, R. J.: Relational learning of pattern-match rules for information extraction. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence (1999) 328–334
4. Yu, H., Han, J., Chang, K. C.: PEBL: positive example based learning for Web page classification using SVM. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (2002) 239-248
5. Zuo, W., Yu, H., Peng, T.: A New PU Learning Algorithm for Text Classification A New PU Learning Algorithm for Text Classification. In: MICAI 2005: Advances in Artificial Intelligence (2005) 824-832
6. Li. X., Li, L. B., Ng, S.-K.: Learning to Classify Documents with Only a Small Positive Training Set. In: Machine Learning: ECML 2007 (2007) 201-213

A spontán beszéd prozódiai frázisszerkezetének modellezése és felhasználása a beszédfelismerésben[1]

Pápay Kinga

DE BTK Általános és Alkalmazott Nyelvészeti Tanszék
4032 Debrecen, Egyetem tér 1.
kinga.papay@gmail.com

A spontán beszéd egységeinek jelölése, felismerése, illetőleg elkülönítése az automatikus beszédfelismerés egyik alapvető problémája nemzetközi szinten is [2], [6], [7], [9], [10]. Amellett, hogy a prozódiai és egyéb kulcsok a szöveg típusától függően különböznek, további problematikus pont, hogy a prozódiai megvalósítás és a szintaktikai szerkezet közötti kapcsolat feltérképezhetősége nemzetközi vita tárgya [3], [4], [5], [8]. Magyar vonatkozásban tovább nehezíti a fejlesztéseket a prozódiailag felcímkézett, országos nagyságú, spontán beszéd adatbázisok hiánya. Ugyanakkor megfelelő akusztikai előfeldolgozással (a szegmentális tartományban végzett lényegkiemelés jellemző vektorainak használata mellett a szupraszegmentális, prozódiai jellemzőkön alapuló lényegkiemeléssel), valamint a spontán beszéd szupraszegmentális jellemzőinek kutatásával a beszédfelismerő rendszerek hatékonysága növelhető: minél többet tudunk bevonni az emberi beszédfelismerési folyamat szintjei – akusztikai, fonetikai-fonológiai, szintaktikai, szemantikai, illetve pragmatikai szint – közül a gépi beszédfelismerésbe, annál biztosabb lesz a működése.

A kutatás célja a magyar spontán beszéd prozódiai frázisokra (IP-kre) bontása, a prozódiai határok megállapítása és ennek bekapcsolása a beszédfelismerő rendszerbe. A kutatás a spontán beszéd vizsgálatán keresztül járul hozzá a pontosabb ismeretekhez a prozódiáról, különös tekintettel a beágyazásokra – a beágyazott részek lokalizálásával, a tonális folytonosság szabályainak megállapításával és rendszerbe illesztésével kísérletet teszünk a felismerő hatékonyságának növelésére. Az elméleti nyelvészet prozódiával kapcsolatos aktuális eredményeit használjuk fel. A prozódia ráillik a szintaktikai csoportosításra az alapvető tagolásban, de további, szemantikai és pragmatikai funkciói is vannak, amelyek ki vannak fejezve a prozódia egy származtatott szintjén. A prozódia az elsődleges elemét, a dallamot használja fel a csoportosításhoz; a dallamvariációk rekurzív használatára utal, hogy minél mélyebb a beágyazás, annál alacsonyabb frekvencián kezdődik a dallam. A prozódia reprezentálja a szintaktikai szegmentumok diszkontinuitását és a tonális kontúrok kapcsolódnak egymáshoz – a szintaktikai diszkontinuitás prozódiai reprezentációja az ún. könyvjelző-hatás. Ez a tulajdonság tágítja a hozzárendelések lehetőségét a szintaxis és a prozódia között, és a prozódiai frázisok kapcsolódása a felismerő szempontjából is modellezhető. A vizsgálatok a beágyazások, alárendelések és mellérendelések, illetve az újrakezdések és hezitálások prozódiai jellegzetességeire terjednek ki, különös tekintettel a tonális folytonosság, a nem folytonos tonális összeillés és a tonális rekonstrukció elvére [4], [5]. E prozódiai jellemzők felhasználásának eredménye lehet a keresési tér csökkené-

1 A kutatás Az ember-gép kommunikáció technológiájának elméleti alapjai című, TÁMOP-4.2.2-08/1/2008-0009 jelű projekt keretein belül zajlik.

se (lehetőséget adhat a felismerés során futó Viterbi-algoritmus szakaszolására), zajos körülmények között robusztusabbá teheti a felismerő működését (ezáltal gyorsul és pontosabb lesz a felismerés), illetve felismerheti a megakadásjelenségeket (szintén a pontosabb felismeréshez járul hozzá).

A szupraszegmentális hangszerkezet egyes elemei, a prozódiai jegyek lényegében a három akusztikai jellemző különböző időtartományokra érvényes – szó- vagy mondatszintű – kombinációi. A beszéd alapfrekvencia-, energia- és időviszonyainak vizsgálatát statisztikai módszerekkel végezzük magyar nyelvű, megfelelő spontán beszéd adatbázison. A spontán beszéd adatbázis gyűjtése és felhasználása specifikusabbá teszi a felismerőt, hiszen a spontán beszédben még gyakoribbak azok a jelenségek, amelyek az automatikusan futó algoritmust megzavarhatják: szótévesztések, javítások, újrakezdések, változtatások a közlés közben, hevesebb érzelmek stb. A méréseket, illetve az annotálást (szegmentálás, címkézés és feliratozás) a Praat hangelemző szoftver [1] segítségével végezzük; ennek során az adott tagmondat hullámformájához rendeljük annak alapfrekvencia- és intenzitásgörbéjét. A vizsgálatok után következik a szabályalkotás, illetve a statisztikai modellezés, valamint ezek bekapcsolása a HTK beszédfelismerő rendszerbe [11] – az új modult a rendszerbe illesztve annak vizsgálata következik, hogy milyen mértékben javítható a beszédfelismerés hatékonysága. Statisztikai modellezés esetén a betanítás során az adatbázis hangfájljaiból az előfeldolgozással nyert szupraszegmentális jellemző vektorok, valamint az adatbázis szegmentálási és címkézési adatai használhatók fel a prozódiai modellek felépítéséhez. A prozódiai szegmentálás ismeretében a hipotézis gráfok újrasúlyozhatók, így a végeredmény kiértékelését már a prozódia alapján nyert információ is befolyásolja [2], [6], [9], [10].

Hivatkozások

1. Boersma, P., Weenink, D.: Praat: doing phonetics by computer 5.1.14. Institute of Phonetic Sciences, University of Amsterdam (2009) http://www.praat.org
2. Borostyán G., Szaszák Gy., Vicsi K.: Folyamatos beszéd szószintű szegmentálása szupraszegmentális jegyek alapján. In: Alexin Z., Csendes D. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE Informatikai Tanszékcsoport, Szeged (2004) 319 – 326
3. Chomsky, N., Halle, M.: The Sound Pattern of English. Harper and Row, New York (1968)
4. Hunyadi, L.: Grouping, the cognitive basis of recursion in language. In: Kertész, A. (ed.): Argumentum, 2. Kossuth Egyetemi Kiadó, Debrecen (2006) 67 – 114
5. Hunyadi L.: Cognitive grouping and recursion in prosody. In: van der Hulst, Harry (ed.): Recursion and Human Language. Mouton de Guyter, Berlin New York (2009)
6. Németh Zs., Szaszák Gy., Vicsi K.: Prozódiai információ használata az automatikus beszédfelismerésben; mondatmodalitás felismerése. In: Alexin Z., Csendes D. (eds.): V. Magyar Számítógépes Nyelvészeti Konferencia. SZTE Informatikai Tanszékcsoport, Szeged (2007) 69 – 80
7. Rabiner, L.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Hills, NJ (1993)
8. Selkirk, E. O.: Phonology and Syntax: The Relation between Sound and Structure. MIT Press, Cambridge (1984)
9. Szaszák, Gy.: A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben. PhD értekezés, Budapest (2009)

10. Szaszák, Gy., Vicsi, K.: Folyamatos beszéd szószintű szegmentálása szupraszegmentális jegyek alapján II. In: Alexin Z., Csendes D. (eds.): III. Magyar Számítógépes Nyelvészeti Konferencia. SZTE Informatikai Tanszékcsoport, Szeged (2005) 360-370
11. Young, S. et al.: The HTK Book (for version 3.4). Cambridge University, Cambridge (2009)

„Amikor nagyapa agyonlövette apát” Fordítások minőségvizsgálata statisztikai alapon

Puskás László

PTE BTK, Pszichológia Doktori Iskola
laszlopuskas@gmail.com

A poszter egy az önéletrajzi emlékezet körébe tartozó olasz művet és annak magyar fordítását megvizsgálva igyekszik olyan statisztikai összefüggések bemutatására, melyek alapján a hibás fordítások, illetve a fordítások bizonyos típusú hibái kiszűrhetőek. A poszter Puskás László Fordítások statisztikai alapú minőségvizsgálata tartalomelemzéssel című előadásának kulisszatitkaiba enged bepillantást, illetve a módszer technikai részleteit igyekszik bemutatni, különös figyelmet fordítva azokra a részletekre, amelyek az előadás keretei között tartalmi és terjedelmi korlátok miatt nem kerülhettek bemutatásra. A posztert és az előadást egyben vitaindítónak is szánom.

A poszter a vizsgálat eredményeinek technikai hátterének részleteibe kíván bepillantást nyújtani, miközben a következő feltevések igazolására törekszik:

I. meghatározott típusú szövegek esetén, az olasz szövegrészekben szereplő szavak száma szinte mindig nagyobb a magyar szövegrészekben szereplő szavak számánál;

II. az olasz és a magyar szövegrészben szereplő szavak számának eltérése arányaiban általában jól behatárolható, de a mondat szintjén nem, csak a szövegrész szintjén alkalmazható;

III. az együtt járások elsősorban a történetek elbeszéléséhez, vagyis a narratív szemléletmódhoz köthetőek;

IV. a szavak számának eltérése a szövegben szereplő szófajok arányainak eltérésével is együtt jár;

V. az eljárás általános alkalmazása lehetővé teszi, hogy olasz és más idegen nyelvű szövegek hibás fordítását nagy valószínűséggel felismerjük, azaz a módszerrel nyelvfüggetlenül hasonlítsunk össze idegen nyelvű szövegeket magyar nyelvű fordításaikkal.

A poszteren be kívánom mutatni, hogy a szóban forgó eljárás a szavak számának milyen eltérési arányai alapján képes kimutatni a fordítás valószínűsíthető hibáit, és milyen érzékenységgel. A vizsgált szöveget szövegrészekre osztjuk, és a szövegrészeket összehasonlítjuk azok fordításával. Az eljárás érzékenységét befolyásolja a vizsgált szövegrészek hosszának kiválasztása, így foglalkozom a kiválasztott beszédszakaszok méretének kérdésével is.

A vizsgálati eljárás elsősorban az összehasonlított szövegrészek szóstatisztikái közötti eltérést elemzi, de a vizsgálat tárgyát képezi az összehasonlított szövegrészekben

szereplő karakterek száma is. A különböző statisztikai lekérdezéseket Word programmal valósítottam meg.

Az eljárás azon a korábban már vizsgált feltevésen alapul, hogy a különböző nyelvek különböző gondolkodásformákra, és a külvilág különböző észlelésére adnak lehetőséget. A nyelv és a gondolkodásmód kölcsönhatásával először Wilhelm von Humboldt foglalkozott a XIX. század elején. A XX. század második felében a kulturális antropológia kezdett foglalkozni a nyelvek, a gondolkozásmód és a kultúra összefüggéseivel. Edward Sapir és Bejamin Whorf különböző amerikai indián nyelveket hasonlított össze európai nyelvekkel. Azt találták, hogy az amerikai indián nyelvek és ezzel együtt a kultúrák tér-, idő- és okságszemlélete is eltér az európai nyelvekétől, illetve kultúrákétól. A Sapir-Whorf hipotézis szerint a nyelv struktúrája és szemléletmódja meghatározza a valóságlátást és a külvilágból jövő ingerek érzékelését. Ahogy az előadásomban is, a poszteren is ennek a gondolatnak egy sajátos megközelítésével kívánok foglalkozni: hogyan adható át egy gondolat két különböző szerkezetű nyelv között anélkül, hogy az átadott gondolat megváltozna, és statisztikai módszerekkel hogyan szűrhetők ki a fordítási hibák. A magyar nyelv a legtöbb európai nyelvtől különbözik. Az eltérő szerkezetű nyelvek fordítása során egy eltérő szerkezetű szöveg jön létre. Mivel az eltérések általában szisztematikusak, statisztikai alapon vizsgálhatóak. Feltételezésem szerint a nem megfelelő módon, szerkezetben átadott fordítás a megfelelő szerkezetű fordítástól eltérő statisztikai paraméterekkel rendelkezik, amely számszerűsíthető, ezzel kimutatva a hibás fordítást. Az előadásomnak ezt a gondolatmenetét a poszteren inkább az elemzés technikai lebonyolításának nézőpontjából közelítem meg. Bár a Sapir-Whorf hipotézisnek azt a részét sokan vitatják, hogy a nyelv határozná meg a kultúrát, mondván, hogy az eltérő nyelvet beszélő emberek eleve különböző kultúrában nőnek fel – tehát magát az okokozati összefüggést vitatják –, azzal nem vitatkoznak, hogy a különböző nyelveket beszélő emberek kulturális sajátosságai eltérhetnek egymástól.

Végül azzal foglalkozom, milyen lehetőségei vannak az eljárás jövőbeni alkalmazásának. Egyrészt mennyire lehet automatizálni technikailag egy szöveg részekre bontását, illetve mennyiben szükséges, másrészt foglalkozni kívánok azzal, hogyan alkalmazható más nyelvek esetében az eljárás, valamint milyen további lehetőségek vannak a megfogalmazott feltételezések igazolására, megerősítésére, és milyen új lehetőségeket nyithat ez bizonyos fordítási hibák kiküszöbölésére, illetve hogyan egyszerűsítheti lefordított szövegek ellenőrzését, hibajavítását.

A néma szünetek időtartamának hatása az érzelmi állapot észlelésére

Szabó Eszter[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Kognitív Tudományi Tanszék
eszabo@cogsci.bme.hu

Kivonat: Korábbi kutatások alapján a szomorú érzelmi állapotot a néma szünetek megnyúlása, míg a vidám érzelmi állapotot ezek lerövidülése jellemzi. Jelen vizsgálat azt kutatja, hogy a spontán monológokban a szünetek hossza hogyan befolyásolja a beszéd érzelmi töltetének észlelését. Semleges tartalmú monológokat a Praat program segítségével úgy módosítottunk, hogy az eredetileg is meglévő szüneteket mesterségesen megnyújtottuk, illetve lerövidítettük, korábbi kutatások adatai alapján. Hipotézisünk az volt, hogy ha minden más feltétel azonos, a hosszabb szüneteket tartalmazó semleges témájú monológokat szomorúbbnak, míg a rövidebb szüneteket tartalmazó monológokat vidámabbnak fogják észlelni a megítélők, mint az eredeti beszédeket. A kutatás hozzájárulhat azoknak a szoftvereknek a fejlesztéséhez, amelyek a beszéd érzelmi töltetét elemzik.

1 Bevezetés

Az érzelmek és a beszéd kapcsolata az elmúlt években nem csak a pszicholingvisztikában, hanem az informatikában, a szintetizált beszédet előállító, illetve az automatikus beszédfelismeréssel foglalkozó kutatók körében is fontos szerephez jutott. Mivel a legújabb kutatások célja, hogy egyre inkább az emberéhez hasonló beszédet tudjunk előállítani, illetve hogy a természetes nyelv mondatait is fel tudja ismerni programunk, egyre több kutatás foglalkozik a beszéd prozódiájával. A beszéd gyakran érzelmekkel telített, és a természetesnek ható szintetizált beszéd elkészítéséhez, illetve a természetes környezetben elhangzott érzelemteli beszéd felismerésének eléréséhez fontos, hogy tudjuk, a beszéd mely tulajdonságai hogyan változnak egy-egy érzelem esetében.

A beszédet legtöbbször fonetikai szempontból elemzik, és olyan fizikai változókat vesznek figyelembe, mint a hangmagasság, a hangerő, az időtartamok. Több kutatás alapján úgy tűnik, hogy az érzelemfelismerés szempontjából a hangmagasság a legfontosabb paraméter. Ugyanakkor pl. Kienast és munkatársai [1] kiemelik, hogy a beszédtempó, az időtartam, az artikuláció pontossága mind hasznos paraméterek a beszélő érzelmi állapotának megállapításához.

2 Hosszú idejű változók vizsgálata

2.1 Korábbi kutatások

A legtöbb az érzelem és a beszéd kapcsolatot vizsgáló kutatás rövid, néhány szavas, vagy egymondatos beszédmintákat elemez, amelyek általában színészektől származnak. Ugyanakkor a természetes beszédben gyakoriak a monológok, a több mondatos, hosszabb beszédek. A hosszabb szakaszok elemzésének szükségességét, ennek hiányát a mesterséges beszédészleléses kutatások is kiemelik (ld. Ververidis és Kotropoulos, [2]).

Elmesélt élettörténetekben a szomorúság és reménytelenség érzésének kifejezésekor kutatók azt figyelték meg, hogy mélyebb, halkabb, erőtlenebb, monoton vagy intonáció nélkülivé vált a beszéd. Gyakoribbá váltak a szünetek – különösen a hosszú (egy másodpercnél hosszabb) szünetek, akár egy szintaktikai egységen belül is. Vidám érzelmek esetén ezekkel ellentétes módon változik a beszéd: magasabb, hangosabb, dallamosabb és gyorsabb lesz (Deppermann és Lucius-Hoene [3]; Scherer [4]).

Egy korábbi kutatásunkban (Szabó [5]) a szomorú és a vidám érzelmi állapot hatását vizsgáltuk kísérleti fonetikai módszerekkel. Itt az önéletrajzi emlékezés módszerét és zenét használva sikerült egy olyan kísérleti elrendezést létrehozni, amelyben a kísérleti személyek szomorú, illetve vidám érzelmi állapotokat éltek át, és ez hatással volt a beszédükre: a beszédtempóra, a beszédbeli szünetek hosszára, a szünetek arányára a teljes időtartamhoz viszonyítva, valamint a hangerőre. Félperces szakaszokat elemezve a hangmagasságban nem mutatkoztak különbségek, és az idői jegyek tűntek fontosnak.

2.2 Jelen kutatás

Jelen vizsgálatunkban arra kerestük a választ, hogy a szünetek hosszának milyen hatása van az érzelemazonosításra. Tehát ha minden más feltétel azonos, akkor csupán a szünetek hosszának megváltozása hogyan befolyásolja a hallgatókat abban, hogy milyen érzelmi állapotot tulajdonítanak a beszélőnek.

A spontán beszédben is elhangozható monológok modellálására semleges témájú beszédrészleteket használtunk fel, amelyeket nem hivatásos színészek mondtak el egy-egy beszélgetés során. A beszédekről felvételek készültek, amelyeket aztán a Praat program segítségével módosítottunk. A monológokban egyébként is benne lévő néma szüneteket az irodalomban megtalálható adatok alapján mesterségesen megnyújtottuk, illetve lerövidítettük, és percepciós tesztelésnek vetettük alá. A megítélőknek arra kellett választ adniuk, hogy a hallott beszédrészleteket milyen érzelmi állapotban mondhatta a beszélő.

Hipotézisünk az volt, hogy a hosszabb néma szüneteket tartalmazó monológokat szomorúbbnak, míg a rövidebb szüneteket tartalmazókat vidámabbnak fogják megítélni, mint az eredeti formájában hagyott beszédrészleteket. A kutatás nagyban hozzájárulhat a beszéd érzelmi töltetét felismerni kívánó szoftverek fejlesztéséhez.

Hivatkozások

1. Deppermann, A. & Lucius-Hoene, G.: Trauma erzählen – kommunikative, sprachliche und stimmliche Verfahren der Darstellung traumatischer Erlebnisse. Psychotherapie und Spzialwissenschaft. Zeitschrift für Qualitative Forschung und klinische Praxis. 1. (2005) 35–73
2. Kienast, M., Paeschke, A., & Sendlmeier, W.: Articulatory reduction in emotional speech. Proceedings of Eurospeech 1999, Budapest, Hungary (1999) 117–120
3. Scherer, K. R.: Vocal communication of emotion: A rewiev of research paradigms. Speech Communication 40. (2003) 227–256
4. Szabó E.: A szomorú és a vidám érzelmi állapot megjelenése a beszédben. Magyar Pszichológiai Szemle, *63, 4*, (2008) 651–668
5. Ververidis, D. és Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication, 48 (2006) 1162–1181

Automatikus intonációs osztályozó felhasználása hallássérültek beszédterápiájában

Szaszák György, Nagy Katalin, Sztahó Dávid, Vicsi Klára

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

Kivonat A BME-TMIT egy prozódiai rejtett Markov-modell alapú modalitásfelismerőt fejlesztett ki, amely szupraszegmentális akusztikai előfeldolgozás után tagmondatok és mondatok határait és a mondat modalitását ismeri fel. Cikkünkben bemutatjuk a modalitásfelismerő automatikus intonációs osztályozásra való felhasználását hallássérültek vagy idegen nyelvet tanulók beszédterápiájában. A rendszer teljesítményét ép hallású bemondóktól származó anyagon optimalizáljuk, majd vizsgáljuk a hallássérült bemondók által bemondott mondatok automatikus osztályozásában. A jobb összehasonlíthatóság érdekében az eredményeket szubjektív lehallgatási tesztek eredményeivel is összevetjük.

1. Bevezetés

A beszéd szupraszegmentális szintje - a prozódia - igen fontos az emberi beszédpercepcióban, és hatékonyan felhasználható a gépi beszédtechnológiában is [1]. A jó minőségű beszédszintézis például elképzelhetetlen a prozódia megfelelő modellezése nélkül [2]. A prozódia beszédfelismerésbeli felhasználása kevésbé elterjedt, mindazonáltal számos kutatás igazolja, hogy a beszédfolyam automatikus tagolásában, a beszédfelismerés eredményességének növelésében, a szintaktikai és szemantikai szintű információ kinyerésében fontos szerepe van (Vö.: [4], [5], [6]).

Az emberi beszédben gyakorlatilag a prozódia az egyetlen akusztikai jellemző, amely a modalitásra utal, néhányan ezt a lehetőséget is vizsgálták már [7], [8]. Az utóbb hivatkozott műben a szerzők olyan rejtett Markov-modell (HMM) alapú rendszert muattak be, amely az F0 és az energia menete alapján végez modalitásfelismerést. Jelen cikkükben a szerzők ezt modalitásfelismerőt vizsgálják beszédterápiás rendszerbe ágyazottan.

A számítógépes beszédterápiás rendszerek interaktív felületet biztosítanak a nyelvtanulónak, amelyet a hallássérültek hatékonyan használhatnak helyes beszéd - a helyes artikuláció vagy a helyes hangsúlyozás és intonáció - elsajátításához. A vizuális visszacsatolás révén ugyanis értékelhetik saját kiejtésüket, "produktumukat", ily módon kiváltva a hiányzó auditív visszacsatolást [9]. A módszert a prozódia elsajátítására használva bizonyított, hogy a vizuális visszacsatolás hatékonyabb, mint a puszta auditív [10], különösen, ha a tanuló referenciamintát is lát - például a kívánatos F0-kontúrét.

A legtöbb napjainkban elérhető beszédterápiás rendszer a helyes artikuláció tanítására koncentrál, emellett a prozódia gyakran elhanyagolt szerepbe szorul. A létező alkalmazások egy csoportja távolságszámítás alapján automatikusan értékeli a tanuló kiejtését (vö. SPECO, [11]), míg más rendszerekben HMM fonéma modelleket használnak a kiértékeléshez [12].

Célunk a prozódia oktatása és automatikus kiértékelésének megvalósítása magyar nyelven. Az így előálló rendszert hallássérült gyerekek használhatják a helyes hangsúlyozás és a modalitásnak megfelelő intonáció elsajátítására. Az automatikus kiértékelés elvégzésére a már említett modalitásfelismerőt adaptáljuk [8], ennek során egy speciálisan erre a célra kialakított ún. intonációs beszédadatbázist is felhasználunk.

2. A modalitásfelismerő

Jelen cikk alapja a korábban már részletesen bemutatott [8] HMM alapon intonáció osztályozását végző modalitásfelismerő. Ez az osztályozó magyar nyelvre 7 különböző modalitás elkülönítésére alkalmas, pontosabbana szükséges csönd és nem mondatzáró modelleket leszámítva a véglegesen elkülönítendő modalitások száma 5, mégpedig: kijelentő, kiegészítendő kérdő, igen-nem kérdő, felkiáltó vagy felszólító, választó.

3. Az intonációs adatbázis

A modalitásfelismerő betanításához külön intonációs adatbázis készült a budapesti Dr. Török Béla - hallássérültekre specializált - Általános Iskolában. Az adatbázis anyagát a tervezett feladatoknak megfelelően állítottuk össze: abban minden modalitású mondat előfordul, mind hosszabb, mind rövidebb, akár egyetlen szóból álló mondat formájában. A felvételeket 60 ép hallású és 19 hallássérült gyermekkel készítettük el. Az előbbi csoport a betanításhoz, míg utóbbi a végső rendszer teszteléshez szükséges.

Az adatbázisban az egyes modalitásoknak megfelelően címkéztük az intonációs kontúrokat. A címkézés kritériuma a megvalósult intonáció, amelyet szakértő ítélt meg. A nem pontos vagy nem helyesen intonált bemondásokat nem használtuk fel. Az osztályozás során használt osztályokat és megfelelő címkéiket az 1. táblázat tartalmazza. Ne feledjük, hogy az 1. táblázatban szereplő 6 osztályon kívül a csönd is modellezendő.

4. Az intonációs sémák betanítása

Az intonációs sémák HMM-jeit az 1. táblázatban szereplő osztályokra az intonációs adatbázis ép hallású beszélőkkel készített részének 2/3-án tanítottuk be. A fennmaradó 1/3 validálási célokat szolgál. A tanított HMM-ek 7 állapotú, balról jobbra felépítésű, a kibocsátási valószínűséget 1 vagy 2 Gauss komponenssel leíró modellek. A használt prozódiai-akusztikai jellemzők az F0 és az energia.

1. táblázat. *A címkézéshez használt intonációs osztályok.*

Intonáció	Címke	Példa	Kontúr
Ereszkedő	DE	Anna áll.	
Eső	FA	Miért áll ott?	
Emelkedő-eső	AF	Anna áll ott?	
Eső-ereszkedő	FD	Gyere ide!	
Lebegő	FL	Ez Anna, és ...	
Emelkedő	RI	Nem?	

Előbbit oktávugrások ellen szűrjük, és logaritmikus tartományban lineárisan interpoláljuk a zöngétlen helyeken. Mindkét jellemző értékét 25 pontos átlagoló szűrővel szűrjük 10 ms keretidő mellett, majd első és második deriváltjaikat is kiszámítjuk.

5. Validálás

Az intonációs osztályozóként használandó modalitásfelismerő előzetes tesztelése az ép hallású bemondások betanításból kihagyott 1/3-án történt. Az egyes mondatokból olyan csoportokat képeztünk, amelyek a beszédterápiás eszközben egy-egy konrét feladatnak felelnek meg. Az eredmények tévesztési mátrix formájában a 2. táblázatban láthatóak (%-os értékekel megadva). Az eső ereszkedő osztályt (FD) az optimalizálás során az esőbe (FA) olvasztottuk be.

2. táblázat. *Tévesztési mátrix az ép hallású gyermekek által produkált intonáció gépi osztályozásában.*

Referencia	Osztályozás [%]				
	DE	FA	AF	FL	RI
DE	**97.67**	2.33	0.00	0.00	0.00
FA	1.61	**82.26**	8.06	6.45	1.61
AF	0.00	0.00	**93.10**	3.45	3.45
FL	2.56	2.56	2.56	**92.31**	0.00
RI	0.00	0.00	0.00	0.00	**100.0**

6. Az intonációs osztályozás tesztelése

Az intonáció osztályozására használt modalitásfelismerő végső tesztelése a hallássérült, és emiatt beszédhibával is rendelkező gyerekektől származó felvételeken

történt. Az osztályozás szerepe ebben az esetben a kiejtés intonáció szempontjából történő értékelése, a kiejtést akkor tekintjük helyesnek, ha a modalitásfelismerő a kívánt intonációt ismeri fel. Ezek a tesztek egyben megfelelnek a modalitásfelismerő beszédterápiás rendszerben történő használatának. A teszteredmények az 3. táblázatban láthatók. Felhívjuk a figyelmet arra, hogy az eredmények nem a modalitásfelismerőt minősítik (arra ugyanis a 2. táblázat vonatkozik), hanem azt mutatják, hogyan alakult a gyermekek által helyesen vagy helytelenül kiejtett intonációinak aránya az egyes intonációtípusokéra a gépi osztályozás esetében.

3. táblázat. *Beszédhibás gyermekek által produkált intonáció osztályozása modalitásfelismerővel.*

Kívánt	Osztályozás [in]				
kiejtés	DE	FA	AF	FL	RI
DE	**33.0**	35.0	0.0	32.0	0.0
FA	9.5	**62.3**	0.0	28.1	0.0
AF	15.5	15.5	**53.5**	15.5	0.0
FL	16.9	32.3	0.0	**50.7**	0.0
RI	0.0	10.0	0.0	30.0	**60.0**

A tesztek alaposabb kiértékelésének érdekében emberi hallgatók is értékelték a beszédhibás gyermekek által használt intonációt szubjektív lehallgatási tesztek keretében. A 21 hallgató ugyanazokra az intonációosztályokra osztályozott, mint a gépi rendszer azzal a kivétellel, hogy a szubjektív hallgatók teljes bizonytalanság (UC) esetén kihagyhatták az adott elem értékelését. Az eredmények a 4. táblázatban láthatók.

4. táblázat. *Beszédhibás gyermekek által produkált intonáció osztályozása szubjektív lehallgatási tesztek során.*

Kívánt	Osztályozás [%]					
kiejtés	DE	FA	AF	FL	RI	UC
DE	**89.0**	1.0	1.5	5.5	0.5	2.5
FA	17.0	**75.0**	1.5	0.5	0.0	6.0
AF	11.4	2.5	**79.6**	0.5	1.0	5.0
FL	44.0	3.5	10.5	**33.5**	0.0	8.5
RI	17.0	1.0	0.5	3.0	**70.0**	8.5

A szubjektív lehallgatási tesztek és az automatikus osztályozás eredményeit összevetve az osztályozási teljesítmények jól párhuzamba állíthatók, kivéve az ereszkedő (DE) és a lebegő (FL) intonációtípusokat. Ennek oka az, hogy a szubjektív lehallgatók valószínűleg ódzkodtak a kissé szofisztikált lebegő kategória

használatától, és akkor is ereszkedő intonációra döntöttek, ha az intonáció valójában bizonytalan, lebegő volt (mintegy alkalmazkodtak a beszédhibás beszélő beszédmódjához). Ugyanerre vezethetők vissza a szubjektív lehallgatás során tapasztalt nagyobb elfogadási hajlandóság, illetve arra is, hogy a szubjektív lehallgatók nyelvtani információra is támaszkodhattak a lehallgatás során, jóllehet természetesen azt az utasítást kapták, hogy a grammatikai vonatkozásoktól tekintsenek el.

Az eredményeket részletesen összehasonlítva azt tapasztaltuk, hogy a szubjektív lehallgatók legalább 50%-a által a kívánttal megegyezőnek elfogadott intonációt a gépi osztályozás csupán az esetek 9%-ában nem fogadta el. A gépi osztályozás tehát szigorúbb, de véleményünk szerint elfogadható osztályozást valósít meg, ami kívánatos is a helyes kiejtés elsajátításában, hiszen a helyes, és nem a még elfogadható kiejtésformák megerősítése az elsődleges cél.

Hivatkozások

[1] Kompe, R.: Prosody in Speech Understanding Systems. LNAI 1307, Springer (1997)
[2] Fujisaki, H., Ohno, S.: The Use of a Generative Model of F0 Contours for Multilingual Speech Synthesis. 4^{th} Int. Conf. on Signal Proc., Vol. 1 (1998) 714–717
[3] Hunyadi, L.: Hungarian Sentence Prosody and Universal Grammar. Peter Lang (2002)
[4] Szaszák, Gy., Vicsi, K.: Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition. In: A. Esposito et al. (eds.): Verbal and Nonverbal Communication Behaviours. Springer. (2007) 138-150
[5] Hirose, K. et al.: Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of Fundamental Frequency Contours. ISCA Tutorial and Research WS on Prosody. Red Bank, USA (2001) 61-66
[6] Veilleux, N. M., Ostendorf, M.: Prosody/parse scoring and its application in ATIS. In: Proc. of ARPA Human Language Technology Workshop '93 (1993) 335-40
[7] Král, P., Klečková, J. Cerisara C.: Sentence Modality Recognition in French based on Prosody. In: Proc. of World Academy of Science, Engineering and Technology, Vol. 8 (2005) 185-188.
[8] Vicsi, K., Szaszák, Gy.: Using Prosody for the Improvement of ASR: Sentence Modality Recognition. Interspeech 2008, ISCA Archive. http:www.isca-speech.org/archive/ (2008)
[9] Vicsi, K.: Computer-Assisted Pronunciation Teaching and Training Methods Based on the Dynamic Spectro-Temporal Characteristics of Speech. In: Divenyi, P. L. et al. (eds.): Dynamics of Speech Production and Perception. IOS Press (2006) 283-304
[10] James, E.: The acquisition of prosodic features of speech using a speech visualizer. IRAL, 14(3) (1976) 227-243
[11] Vicsi, K., Csatári, F., Bakcsi, Z., Tantos, A.: Distance score evaluation of the visualized speech spectra at audio-visual articulation training. In: Proc. Eurospeech (1999) 1911-1914
[12] Narusa, J.: Computer-aided spoken language training with enhanced visual and auditory feedback. In: Proc. Eurospeech (1999) 183-186

Morfoszintaktikailag annotált néprajzi korpusz[1]

Szauter Dóra[1], Vincze Veronika[1], Almási Attila[1], Alexin Zoltán[2], Kiss Márton[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{szauter, vinczev, mkiss}@inf.u-szeged.hu, vizipal@gmail.com
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
alexin@inf.u-szeged.hu

Kivonat: Az első, néprajzi tematikájú, nyelvileg elemzett magyar nyelvű korpusz szövegállománya a Néprajzi Múzeum Ethnológiai Adattárából származik. A szövegek két téma köré csoportosulnak: népi hiedelemvilág és táltosszövegek. A korpusz tartalmazza a szövegszavak lehetséges és az adott kontextusban helytálló morfoszintaktikai MSD-kódjait. A korpusz bővíthető más jellegű néprajzi szövegekkel, illetve a későbbiekben lehetséges lesz az állomány szintaktikai annotációjának elvégzése is.

1 Bevezetés

Cikkünkben bemutatjuk az első, néprajzi tematikájú, nyelvileg elemzett magyar nyelvű korpuszt. Újdonságot jelent egyrészről a korpusz tematikája, hiszen néprajzi témájú szöveges adatbázisok eddig nem vagy alig bizonyultak elérhetőnek elektronikus formátumban (a néprajzi adatbázis-építés nehézségeiről l. [2]), másrészről – tudomásunk szerint – magyar nyelvű néprajzi szövegeket még nem vetettek még alá számítógépes nyelvészeti elemzésnek.

A néprajzi korpusz feldolgozása követi a Szeged Treebankben [1] megszokott jelölésrendszert. A korpusz ebben az esetben is TEI XML formában készül, amelyben a szöveget szakaszokra, bekezdésekre, mondatok és szavakra bontják fel. Minden egyes szó mellett szerepelnek majd a lehetséges morfoszintaktikai elemzései, illetve a kontextusnak megfelelően kiválasztott morfoszintaktikai kód. A munka elvégzéséhez a kutatók azokat a szoftvereket fogják használni, amelyeket korábban a Szeged Treebank elkészítéséhez is igénybe vettek. Szükség esetén kisebb javításokat és korrekciókat végeznek a programokon.

A néprajzzal foglalkozó kutatók számára ez a fajta munka újdonságot jelent, mivel korábban a feldolgozásokat többnyire kézzel végezték. Sok esetben az összegyűjtött szövegek számítógépes formára hozása – begépelése, rögzítése sem történt még meg. Vélhetően ez a kisebb, mintegy 110 ezer szövegszó méretű korpusz elegendő vonzerőt gyakorol majd a néprajzos szakma képviselőire, hogy további anyagokat gyűjt-

[1] Az itt ismertetett kutatást az NKTH Jedlik Ányos program 2008, MASZEKER (Modell Alapú Szemantikus Kereső Rendszer) kódnevű kutatás-fejlesztési projektje támogatta.

senek össze, adjanak át feldolgozásra, s a tőlünk visszakapott anyag pedig újabb eredményeket hozhat a kutatásban.

A következőkben részletesen bemutatjuk a korpuszt, ismertetjük a nyelvi annotáció folyamatát, végül statisztikai adatokat közlünk az adatbázisról.

2 A korpusz tematikája

A néprajzi korpusz két témából tartalmaz szövegeket: népi hiedelemvilág (2704 szöveg) és táltosszövegek (432 szöveg). A szövegek lejegyzése a XX. század elején történt, a történelmi Magyarország csaknem minden tájegységéről származnak adatok. Az eredeti kéziratok a Néprajzi Múzeum Ethnológiai Adattárában találhatók, és gyűjteményes formában, könyv alakban is hozzáférhetőek [4].

A hiedelemszövegek a hétköznapi élet szinte valamennyi területéről tartalmaznak közléseket: az emberi élet fő állomásai (születés, keresztelő, férjszerzés, betegség, halál, túlvilág), időjárás, jeles napok, háziállatok. A rövid, egymondatos hiedelmeket hol magyarázat kíséri, hol rövid elbeszélések illusztrálják. A gyűjteményben egyszerű leírásokon kívül versformába szedett ráolvasások is találhatók. Bizonyos hiedelmek több változatban is előfordulnak. A szövegekből gyakran népszokáselemekre is következtethetünk:

Ha a menyasszony cipőjét ellopják a lakodalom éjjelén s lekaparva a talpáról a földet felteszik a füstre – ez a házas társak nyugodt életét megrontja.

A hiedelemközlésekhez fűzött megjegyzések az adott közösség életéről is hordoznak információt:

Ha a fiatal asszony közvetlen esküvő után 3-szor egymásután belenéz a kutba: meghal minden gyereke. Ez a szokás általános lett nálunk!

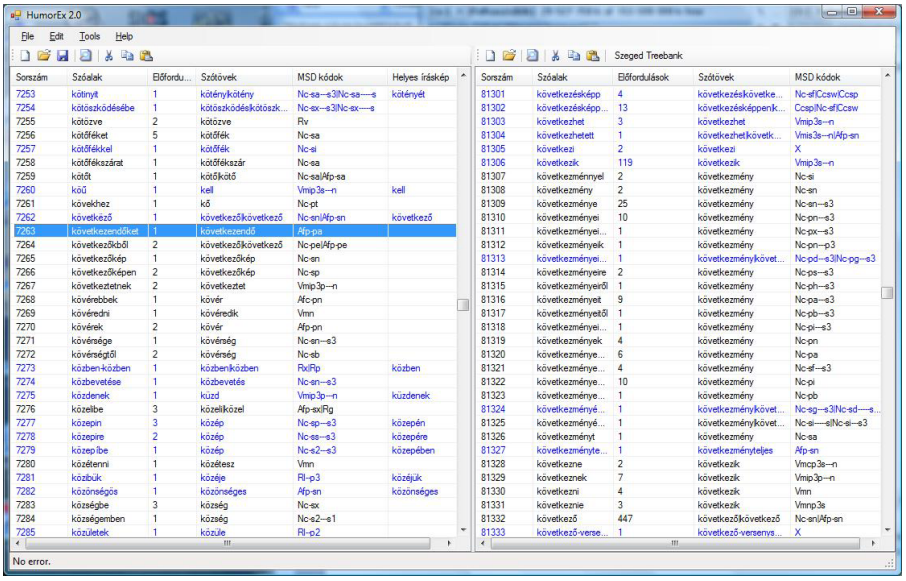
A táltosszövegekben a Kárpát-medence több tájegységéről található információ garabonciásokról, tudósemberekről, tudósasszonyokról, táltosokról, illetve azok ismertetőjegyeiről és képességeiről, leginkább róluk szóló rövid elbeszélések formájában, a tájegységnek megfelelő nyelvváltozatban.

3 Morfoszintaktikai annotáció

A korpusz szövegállományának digitalizálását követően Dalányi Sándor, a Stockholmi Egyetem kutatója kezdett foglalkozni az anyaggal. Egy közös kutatásfejlesztési projekt keretében jutottunk hozzá a szövegekhez, melyeken számítógép segítségével végzünk további nyelvi elemzéseket.

A feldolgozás első lépése a korpuszban található szavak összegyűjtése és morfoszintaktikai elemzése volt. A kapott 25 034 szóból álló listát a kutatók két részre bontották aszerint, hogy az adott szó megtalálható-e a Szeged Treebankben. Az ismert és korábban már elemzett szavakat ebben a munkafázisban félretettük, kódolá-

sukat egy az egyben átemeltük a Szeged Treebankből, és csak a korábban elő nem forduló, ismeretlen szavakkal foglalkoztunk. 14347 ilyen szó fordult elő a néprajzi szövegekben. Az annotálási munkálatokhoz az 1. ábrán látható programot használtuk. Először a szavakhoz számítógépes elemzéssel morfoszintaktikai kódokat rendeltünk, amelyeket azután át kellett nézni és jóvá kellett hagyni. Továbbra is tartottuk magunkat ahhoz, hogy a nyelvi elemzésben az Értelmező Kéziszótár kiadásaira támaszkodunk, annak a kategóriarendszerét vesszük át.



1. ábra. A szövegszavak morfológiai annotálásához készített szoftver.

A program két panelből áll, amelyekbe egyrészt az eddig feldolgozatlan szavak listáját (bal oldal), illetve a Szeged Treebank szótárát lehet betölteni (jobb oldal). Amennyiben az új szó csak kis mértékben, pl. esetragban tért el egy korábban már elemzett szótól, akkor a korábbi szóhoz rendelt morfológiai kódokat korrekcióval át lehet emelni. A programnak van is egy ilyen másolási funkciója. Új elemként jelent meg a program baloldali paneljében egy oszlop, amelyben a szavak ma szokásos írásmódját lehet megadni. Ha ez a modern alak előfordult a Szeged Treebank szótárában, akkor a program át tudta emelni a kódot a meglévő adatbázisból. A múlt századi vagy annál is régebbi népies vagy tájnyelvi szövegekben található szavaknál gyakori jelenség, hogy a helyesírásuk megváltozott.

A tájnyelvi szavak (*goroboncás*, *slájer*) mellett sajátos problémát jelentettek a következő esetek:

- népies helyesírású szavak (*ígízís*, *abbú*): ezek mellett feltüntettük a sztenderd magyar helyesírású alakot (*igézés*, *abból*), és ezek MSD-kódja a legtöbbször már átemelhető volt a Szeged Korpuszból. Amennyiben a szóalak nem szerepelt benne, akkor természetesen megadtuk a megfelelő kódo(ka)t.

- ha a népies helyesírású szó egybevág egy másik, létező szóalakkal (*mellül*, *aggyá*): ezek különös figyelmet igényeltek az egyértelműsítésnél, hiszen már volt egy – sztenderd helyesírás szerinti – MSD-kódjuk, azonban a szövegekben többnyire a népies változat fordult elő, így külön meg kellett adni annak sztenderd alakját (*mellől*, *adjál*) és MSD-kódját/kódjait.

A korpusz jelenleg morfoszintaktikai annotációt tartalmaz az MSD-kódrendszert [3] követve: minden szövegszó mellett szerepel annak összes lehetséges morfoszintaktikai kódja, és ezek közül az adott kontextusba illő is jelölve lesz (ez a munkafázis jelenleg zajlik).

4 Statisztika

A korpusz 109760 szövegszót tartalmaz összesen (a hiedelemszövegekben 65715, a táltosszövegekben 44045 szövegszó szerepel). Mivel a szövegszavak egyértelműsítése még folyamatban van, további statisztikákat például a morfoszintaktikailag egy-, illetve többértelmű szavak arányáról a későbbiekben közlünk.

5 További tervek

A morfoszintaktikai elemzésen kívül szintaktikailag is elemezni kívánjuk a teljes szövegállományt (dependenciaelemzés). A korpusz később esetleg más jellegű szövegekkel (például népmesék) is bővíthető.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceddings 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
2. Pávai I.: A néprajzi adatbázis-építés akadályai. Néprajzi Hírek 1-4 (1996) 86-89
3. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf
4. Verebélyi K. (szerk.): Néphit szövegek. Magyar Néprajzi Társaság, Budapest (1998)

Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban

Vincze Veronika

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
vinczev@inf.u-szeged.hu

Kivonat: A félig kompozicionális főnév + ige szerkezetek számítógépes nyelvészeti kezelésének megkönnyítésére hozzuk létre a Szeged Korpusz egy olyan változatát, amelyben e kifejezések és altípusaik annotálva vannak. Az elkészült korpusz tanító adatbázisként szolgálhat a szerkezetek automatikus azonosításához, így hozzájárulhat többek között a gépi fordítás és az információkinyerés eredményességéhez.

1 Bevezetés

A számítógépes nyelvészeti alkalmazások számára az egyik legnagyobb kihívást a kollokációk megfelelő kezelése jelenti. Kollokációk gyakran előfordulnak a nyelvhasználatban, és viselkedésük sokszor eltér a kompozicionális kifejezésekétől, ezért különleges bánásmódot igényelnek.

2 Félig kompozicionális szerkezetek

A kollokációk egyik altípusának tekinthetők a félig kompozicionális főnév + ige szerkezetek (*tanácsot ad, döntést hoz, virágba borul…*) [1], melyekben a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában. E szerkezetek számítógépes nyelvészeti kezelése nem problémamentes. Mivel jelentésük nem teljesen kompozicionális, a szerkezet részeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Továbbá, a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*), 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. Mindezen jellemzők miatt a félig kompozicionális főnév + ige szerkezetek felismerése és megfelelő kezelése kulcsfontosságú a számítógépes nyelvészeti alkalmazásokban, például a gépi fordításban és az információkinyerésben.

Egy félig kompozicionális szerkezeteket tartalmazó adatbázis létezése igencsak megkönnyítené az ilyen szerkezetek automatikus felismerését (így azok megfelelő kezelését is). Más nyelvekre léteznek már ilyen korpuszok: például hozzáférhető egy többszavas igéket tartalmazó adatbázis az észtre [2, 3] és a prepozíciós vonzattal rendelkező igék adatbázisa a németre [4]. Ezek nyomán hozzuk létre az első olyan magyar nyelvű korpuszt, melyben a félig kompozicionális főnév + ige szerkezetek be vannak jelölve. Az annotáció alapját a Szeged Treebank 2.0 képezi [5], mivel ez az adatbázis már tartalmaz morfoszintaktikai annotációt és szintaktikai elemzést is. Az annotáció során a szerkezet <FX></FX> tagek közé kerül, és jelölni lehet a szerkezet altípusát is. Jelenleg az üzleti hírek és az újsághírek annotációja készült el teljesen, a jogi szövegeke annotációja folyamatban van, azonban terveink szerint a teljes korpusz anyagára kiterjesztjük az annotációt.

A félig kompozicionális szerkezetek a prototipikus főnév + ige mintán kívül előfordulhatnak más szintaktikai mintázatban is, például igenévi alakban vagy főnévi (képzett) változatban. A korpuszban az alábbiak szerint vannak megjelölve a különféle altípusok (példákkal illusztrálva):

Főnév + ige kombinációja <verb>: *bejelentést tesz*

Igenevek <part>
Folyamatos melléknévi igenév: *életbe lépő (intézkedés)*
Befejezett melléknévi igenév: *csődbe ment (cég)*
Beálló melléknévi igenév: *fontolóra veendő (ajánlat)*
Főnévi igenév: *forgalomba hozni*
Határozói igenév: *ajánlatot téve*
Igei igenév: *(jogszabály) adta lehetőség*

Főnévi változat <nom>: *bérbe vétel*

Előfordulhat, hogy a főnévi és az igei komponens nem egymás mellett fordul elő a mondatban. Ezeket az eseteket is jelöljük, és a <split> altípusba soroljuk őket:

Különálló szerkezet <split>: *előadást fog tartani*

Mivel a Szeged Treebank már eleve tartalmaz szintaktikai annotációt, a félig kompozicionális szerkezetek jelölése során figyelembe vesszük a frázishatárokat is: a szerkezet főnévi komponensének legkülső határát jelöljük meg mint a szerkezet részét, nem csak pusztán a főnévi fejet. Ennélfogva a főnévi komponens esetleges jelzői is bekerülnek a szerkezetbe:

<FX>**nyilvános** ajánlatot tesz</FX>

A melléknévi igeneves alakban előforduló szerkezetek esetében pedig könnyen előfordulhat, hogy a szerkezetben más NP is szerepel:

<FX>**Nyíregyházán** tartott ülésén</FX>

A tárgyesetű főnévi komponenst tartalmazó szerkezetek nominalizációja kétféleképpen is történhet: összetett szóval, illetve birtokos szerkezettel:

<FX>szerződéskötés</FX>
<FX>adásvételi szerződések megkötése</FX>

A korpuszban mindkét típust jelöljük.

3 Statisztika

Az adatbázis jelenlegi formájában 407 félig kompozicionális szerkezetet tartalmaz 1745 előfordulásban az alábbi eloszlásban:

1. táblázat: A félig kompozicionális szerkezetek száma típus szerint.

	verb	**part**	**nom**	**split**	**összesen**
üzleti hírek	565	270	90	40	965
újsághírek	205	92	31	24	352
összesen	770 58.5%	362 27.5%	121 9.2%	64 4.8%	1317 100%

4 A korpusz hasznosíthatósága

A korpusz eredményesen használható mint tanító adatbázis a szerkezetek gépi úton történő azonosításához, melynek nyomán a különféle számítógépes nyelvészeti alkalmazások – például gépi fordítás és információkinyerés – pontossága is javulhat.

A gépi fordítás során a programnak először is fel kell ismernie, hogy az adott főnév és ige összetartozik (egy kollokáció két részét alkotják), továbbá – mivel egy adott szerkezet és idegen nyelvű megfelelője esetében a főnévi komponens megegyezik (azaz általában szó szerint fordítható), míg az ige eltérő [6] – a fordítóprogram az adott főnévhez társított megfelelő igét egy célnyelvi tanulókorpusz alapján készített gyakorisági mutató segítségével tudja kiválasztani.

Információkinyerésnél, különösen relációk kinyerésekor rendkívül fontos a mondatok megfelelő szintaktikai elemzése. A félig kompozicionális szerkezetek főnévi komponensének és a szerkezet egyéb vonzatainak szintaktikai státusa azonban vitatott [7]. Információkinyerés szempontjából a komplex predikátum feltételezése a legígéretesebb, azaz a szerkezetet egy egységként kezeljük, és ennek vannak vonzatai. Így például *A cég bérbe vette a raktárt* mondatból kinyerhető viszonyok a következők: **bérbe vétel** esemény, szereplői: **a cég**, **a raktár**. Ezzel szemben, ha az elemző nem ismeri fel a félig kompozicionális szerkezetet, így a főnévi komponens különleges szintaktikai státusát sem, a következő (helytelen) eredményt adja: **vétel** esemény, szereplői: **a cég**, **bér**, **a raktár**. Az elemző program betanítására szintén jól használható a létrehozott korpusz.

Köszönetnyilvánítás

A szerző köszönetet mond Szarvas Györgynek az annotációs eszköz kifejlesztésében nyújtott önzetlen segítségéért.

A kutatást – részben – a TUDORKA és MASZEKER programok keretében az NKTH támogatta.

Hivatkozások

1. 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)
2. Kaalep, H.-J., Muischnek, K.: Multi-Word Verbs in a Flective Language: The Case of Estonian. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Context. Trento, Italy (2006) 57-64
3. Kaalep, H.-J., Muischnek, K.: Multi-Word Verbs of Estonian: a Database and a Corpus. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco (2008) 23-26
4. Krenn, B.: Description of Evaluation Resource – German PP-verb data. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco (2008) 7-10
5. 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
6. Vincze V.: Angol–magyar főnév + ige szerkezetek és igei párjaik. In: Váradi T. (szerk.): II. Alkalmazott Nyelvészeti Doktorandusz Konferencia. Budapest: MTA Nyelvtudományi Intézet (2009) 113-123
7. Alonso Ramos, M.: Towards the Synthesis of Support Verb Constructions. In: Wanner, L. (ed.): Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk. Benjamins, Amsterdam / Philadelphia (2007) 97-138

Magyar nyelvi elemző modulok az UIMA keretrendszerhez

Zsibrita János[1], Nagy István[1], Farkas Richárd[2]

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

1 Az UIMA keretrendszer

Az UIMA (Unstructured Information Management Application) keretrendszer [1] célja olyan szoftverrendszerek fejlesztésének támogatása, amelyek nagy mennyiségű strukturálatlan adat elemzését célozzák meg. Az Apache UIMA[1] az UIMA specifikáció nyílt forráskódú implementációja, amely kifejezetten szöveges dokumentumok feldolgozását támogatja.

Az UIMA keretrendszer platformfüggetlen, törekszik az elemzés során minél inkább szabványos megoldások használatára. Fő célja, hogy az egyes elemző modulok könnyen beilleszthetőek legyenek elemzési láncokba (letöltöm és már használom is) és hogy a felhasználó számára megkönnyítse a leginkább megfelelő komponens kiválasztását (azonos feladatot ellátó komponensek gyorsan cserélhetőek).

A keretrendszer lehetőséget ad egy komplex probléma kisebb részproblémákra történő szétbontására, mint például: mondatra bontás, tokenizálás, tulajdonnévfelismerés. Minden feldolgozási egység egy meghatározott interfészt implementál (Java vagy C++ nyelven), a keretrendszer felügyeli az elemzési lánc összeállítását és futtatást, gondoskodik az egységek közötti adatáramlásról, performanciamérésről stb. A programozónak csak az adott modul megírására kell fókuszálnia, minden egyebet a keretrendszer hajt végre.

2 Magyar nyelvi elemző modulok

A Szegedi Tudományegyetem Informatikai Tanszékcsoportjánál elkészítettünk egy magyar nyelvi elemző láncot JAVA programozási nyelven. A munka elsősorban meglévő JAVA nyelvű modulok magyar nyelvre adaptálásából és létező magyar nyelvi modulok „JAVA-sításából" állt. A JAVA nyelvű modulok egyrészről könnyedén beilleszthetőek az utóbbi években népszerűvé vált UIMA keretrendszer alá, másrészről könnyen építhetőek be webes alkalmazásokba (például Google Web Toolkit).

[1] http://incubator.apache.org/uima

Az elemzési folyamat első lépése a szöveg mondatokra bontása, ehhez a Northwestern University nyelvi csomagjának (MorphAdorner) [2] *SentenceSplitter*ét használtuk, kiegészítve a beépített szótárat azon speciális magyar rövidítésekkel, amelyek után bár a szövegben . áll, mégsem mondatvégek. Ilyen például a *zrt.*, a *szül.* vagy a hónapnevek rövidítései. Második lépésben a mondatokon belüli tokenek azonosítása történik, szintén a MorphAdorner-ben található *Tokenizer* segítségével.

Az így kapott tokenek morfológiai elemzése a magyar nyelvre készült, szintén szabad forrású, Hunspell [3] rendszer JAVA-sított verziójával történik. A lehetséges morfológiai kódok halmazából a szövegben betöltött szerep (szófaji kódok és szótövek) kiválasztásához a Stanford Maximum Entrópia POS taggert [4] tanítottuk a Szeged Korpuszon.

Ezeken felül UIMA modulként is használható a magyar nyelvű újsághíreken tanult tulajdonnév-felismerő algoritmusunk. Ez saját paraméterezhető jellemzőkészletet és a MALLET Conditional Random Fields implementációt[2] használja.

Az így megalkotott elemzési lánc segítségével lehetővé vált magyar nyelvű szövegek standard elemzési eszközökkel történő feldolgozása, illetve egyszerűbbé válik egy feladatot megoldó különböző algoritmusok cseréje és tesztelése.

Köszönetnyilvánítás

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

Hivatkozások

1. Gotz, T., Suhre, O.: Design and implementation of the UIMA Common Analysis System, IBM Systems Journal (2004)
2. Kumar, A.: MONK Project: Architecture Overview. Technical Report of the Northwestern University (2009)
3. Németh, L., Halácsy, P., Kornai, A., Trón, V: Nyílt forráskódú morfológiai elemző. Magyar Számítógépes Nyelvészeti Konferencia (2004)
4. Toutanova, K., Klein, D., Manning, C., Singer, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL (2003) 252-259

[2] http://mallet.cs.umass.edu/

VIII. Stratégiai Kutatási Terv

Stratégiai Kutatási Terv

Nyelv- és Beszédtechnológiai Platform
e-mail: info@hlt-platform.hu

1. Vezetői összefoglaló

A Nyelv- és Beszédtechnológiai Platform a szektor vezető kutatóműhelyeit és ipari partnereit tömörítő stratégiai szövetség. A Platform Stratégiai Kutatási Tervének célja az, hogy megfogalmazza a hazai nyelv- és beszédtechnológia fejlődésének irányait, e technológiák nyelvfüggő elemeinek „kötelező" hazai feladatait, rámutasson a nemzetközi kitörési lehetőségekre, és meghatározza az ezek realizálásához szükséges lépéseket. Jelen dokumentum szándékunk szerint a gazdasági, kormányzati döntéshozók, az ágazati szereplők számára jövőbetekintő stratégiaként, kutatási sarokpontokat és módszereket meghatározó iránymutatásként szolgál, amely az alábbi megállapításokat tartja kulcsfontosságúnak:

- A szektor mai gazdasági, társadalmi környezete a helyzetelemzésben felvázolt kedvezőtlen jelenségek és akadályok ellenére a hajtóerők, a motivációk tekintetében nagyon ígéretes. A magyar nyelv- és beszédtechnológia rendelkezik olyan jelentős erősségekkel, mint a szaktudás, élenjáró technológia, aktív nemzetközi kutatói kapcsolatok, amelyekre a sikeres előrelépés alapozható.
- A jövő tudásalapú gazdaságának és társadalmának nélkülözhetetlen alkotóelemei azok a technológiák, melyek hatékonyan támogatják a természetes emberi kommunikációt. Ezek kifejlesztését szolgálják a legfontosabb stratégiai célok: a kutatási infrastruktúra kialakítása, a természetes nyelven megfogalmazott információ megértésének számítógépes támogatása, az automatikus gépi megértés megvalósítása, az interdiszciplináris kutatások előtérbe helyezése.
- Nemzetközi kitörési pontokat ad a robusztus beszédfelismerési technikák fejlesztése, a nagyszótáras, folyamatos többnyelvű gépi beszédfelismerés hatásfokának javítása, az idegen nyelvű szövegek megértését támogató gépi fordításra, illetve a szöveges tartalmak elemzését végző szemantikus technológiákra irányuló fejlesztés, az emberi beszédértés, a kogníció nemzetközi szinten előrehaladott kutatásaiba történő bekapcsolódás, az eredmények alkalmazásra kész technológiába való beépítése.
- A technológiai fejlesztésekkel együtt járnak a kutatás-fejlesztés hatékonyságának és gyakorlati alkalmazásának javítását szolgáló tevékenységek: szakmai kommunikációs központ kialakítása, a szabványosítás, a kutatói utánpótlás koordinált képzése, a kutatásfinanszírozási keretek hosszú távú meghatározása.

A Stratégiai Kutatási Terv törzsanyaga elsősorban a szakpolitikának, döntéshozóknak szóló összegző, iránymutató dokumentum, míg a szakmai(bb) érdeklődésű olvasó a bizonyos kérdéseket részletesen tárgyaló Jelenkép és Jövőkép mellékletekből kaphat további információt.

2. Bevezetés

2.1. Nyelv- és beszédtechnológia a tudásalapú társadalomban és gazdaságban

Az emberihez közel álló technológiák teljesítményét az ember adott területen mutatkozó képességéhez szokás viszonyítani. Feltehetők tehát például az alábbi kérdések: tud-e egy robot egy tűbe cérnaszálat befűzni, tud-e egy nyelvtechnológiai eszköz gyorsírni, tud-e egy beszédfelismerőből és -előállítóból álló számítógépes dialógusrendszer egy koktélparti hangzavarában működni. A válasz a kb. 150 éve művelt robottechnikában a „majdnem", a kb. 50 éve művelt nyelv- és beszédtechnológiában pedig az, hogy sajnos még nem. De a robotok azért igen hasznosak például az oxigénsátorban ápoltak ellátásában, a nyelv- és beszédtechnológia fejlesztései pedig például az írott szöveg, illetve rögzített hanganyag akár hatalmas halmazában az általunk meghatározott információ megtalálásában. Ezek a gondolatok arra kívánnak rámutatni, hogy az embert utánzó technológiák egyre csak közelítik — de valószínűleg a maguk teljességében soha nem érik el — az emberi teljesítőképességet, mindazonáltal egyes tulajdonságaik révén (például sterilitás a robotikában, fáradhatatlanság és gyorsaság a nyelv- és beszédtechnológiában) már akkor is hasznosak (a szó gazdasági értelmében is), amikor az emberihez hasonló tökéletességtől még elég messze állnak. És ahogy a robotika nem maradt abba 50 év után, a nyelv- és beszédtechnológiát is folytonosan fejleszteni kell, hogy egyre több területen legyen gazdaságilag is hasznos helyettesítője az emberi munkaerőnek, illetve elvégzője az ember által fel nem vállalt mennyiségű munkával járó feladatoknak.

A számítógépek és egyéb infokommunikációs eszközök mindennapi életünkben játszott szerepe, s ezzel együtt a ránk zúduló információ mennyisége folyamatosan növekszik. Alapvető fontosságúak tehát azok a módszerek, melyekkel könnyebben, gyorsabban és kényelmesebben tudjuk elérni a számunkra fontos információt, és csak azt. A nyelv- és beszédtechnológia ebben tud segíteni: az informatikusok, mérnökök, pszichológusok és nyelvészek együttműködéséből kialakult kutatási terület célja, hogy olyan új technológiákat és alkalmazásokat állítson elő, melyek az emberi kommunikációt természetesen és hatékonyan szolgálják ki (l. 5.2.). A természetes nyelven történő információáramlás és az emberi tudás számítógépes támogatása egyre nagyobb szerepet játszik nemcsak az európai gazdaságban, hanem az esélyegyenlőség és az életminőség javításában is. Ezt felismerve az Európai Unió régóta kiemelt figyelmet fordít a nyelv- és beszédtechnológiai fejlesztésekre. A kérdés prioritását egyértelműen jelzi, hogy e törekvések az európai információs társadalom előmozdítására irányuló i2010[1]

[1] http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm

kezdeményezés részévé váltak. Az i2010 által megjelölt három kiemelt fontosságú területen (információs tér, kutatási ráfordítás és innováció, társadalmi integráció) a nyelv- és beszédtechnológiának kulcsszerep jut:

- egy változatos és minőségi tartalmat és szolgáltatásokat, biztonságos és gyors kommunikációs lehetőségeket elérhető áron nyújtó információs társadalomban a tartalom és szolgáltatások széles körének kialakításában;
- az Európa felzárkózását biztosító infokommunikációs-technológiai kutatások terén a kutatást és az innovációt érintő európai szintű befektetések hatékony felhasználásában, az innováció előrevitelében;
- valamennyi polgár életminőségének javításához szükséges közszolgáltatások mindenki számára hozzáférhetővé tételében.

Nyelv- és beszédtechnológia által támogatott tartalom és szolgáltatások nélkül az információs társadalom életképtelen, e technológiák nélkül Európa kulturális öröksége a digitális kor számára elveszhet. Ehhez a kontextushoz, a megfogalmazott prioritásokhoz és irányelvekhez Magyarországnak is igazodnia kell. Az infokommunikációs technológiák új minőségi szintre emelése csak akkor lehetséges, ha áttörést érünk el a jelenleg mutatkozó **nyelvi korlátok** leküzdésében. A magyar nyelv- és beszédtechnológia ebből a szempontból igen speciális helyzetben van. Ugyan a nemzetközi kutatás-fejlesztés jelentős eredményeit tekintve világos, hogy sok más nyelvhez hasonlóan bizonyos mértékig követi a vezető angolközpontú technológiákat, de a magyar nyelv radikálisan egyedi jellege új módszerek kidolgozását követeli meg, melyek nemzetközi szinten is érdeklődésre tarthatnak számot, nemcsak a magyarhoz hasonló tipológiájú nyelvek esetében. Az eddigi itthoni eredmények azt mutatják (l. a Jelen- és Jövőképet), hogy e tekintetben életképes és fejlődő nyelvi középhatalom vagyunk, és a stratégiai terv középpontjába a fent megfogalmazott célok elérését biztosító technológiákat kell állítani.

2.2. Helyzetelemzés

A magyar nyelv- és beszédtechnológiai kutatás-fejlesztés eddigi eredményei nemzetközileg elismertek, számos területen világszínvonalat képviselnek. Részletes bemutatásuk a Jelenképben található, jelen fejezet a kutatás-fejlesztési tevékenység gazdasági, társadalmi környezetét jellemző sajátosságokra, az előtte álló akadályokra tér ki röviden.

A nyelv- és beszédtechnológiai fejlesztések mai gazdasági, társadalmi környezete a hajtóerők, a motivációk tekintetében nagyon kedvező. Az előző részben vázolt európai léptékű célok, a globalizáció, a telekommunikácós, hálózati technológiák rohamos előretörése, a felhasználóközpontúság követelménye az ágazat soha nem látott, ugrásszerű fejlődéséhez vezethet már a következő 5 éven belül. Ehhez azonban számos akadályt kell leküzdeni.

Mint több más hazai iparág, a magyar nyelv- és beszédtechnológia fejlesztései is sok esetben a magyar nyelvet beszélők számából következően olyan szűk piaccal találkoznak, amely számos esetben önmagában nem képes finanszírozni a létrehozásához szükséges kutatás-fejlesztési tevékenységet. A jelenleg rendelkezésre álló

erőforrásai és kapacitásai nem teszik lehetővé magas költségű innovációs tevékenység külső támogatástól független folytatását (ennek illusztrálását l. Jövőkép *A közeljövő kutatásai* fejezet). Az állami és vállalati kutatás-fejlesztési ráfordítások mértéke nemzetgazdasági szinten is nagyon alacsony, ez alól természetesen ez a szektor sem kivétel, és ez nemcsak a magyarnál jelentősen erősebb gazdasággal rendelkező országokkal való összehasonlításban van így, hanem a régió hozzánk hasonló méretű országaival szemben is (pl. Csehország, Szlovénia).

Az elmaradás és forráshiány más vonatkozásban is észrevehető, a kutatói utánpótlás, szakemberképzés területén az alulfinanszírozottság már rövid távon is kritikus versenyhátrányhoz vezet. Az ipari és a kutatás-fejlesztési szféra közötti mobilitás alacsony és erősen egyirányú, a kommunikáció korlátozott. Egyrészről az ipari szférából a kutatás felé nehezen mozdulnak el a szakemberek. Ennek egyik oka, hogy különösen a nonprofit intézményekben dolgozó kutatók juttatásai jóval alacsonyabbak, mint a gazdasági szférában a hasonló szakértelemmel rendelkező munkaerőé. Emellett a nyelv- és beszédtechnológiához szükséges és használható magas szintű tudás piaci értéke jóval nagyobb annál, mint amit az állami intézmények nyújtani tudnak, így a kutatás-fejlesztés területéről már most jelentős az elvándorlás a nem innovatív, alkalmazó munkakörökbe, illetve külföldre. Másrészt az ipari igények ritkán jutnak el a kutatás-fejlesztési szervezetekhez, azok kutatási eredményei pedig elvétve hasznosulnak az iparban. Hozzájárulhat ehhez az, hogy nincs szervezett, irányított és naprakész, a fejlesztéseket bemutató és közvetítő kommunikáció(s csatorna), valamint az egyes szervezetek sem koordinálják egymás között tevékenységeiket a szűkös erőforrások minél hatékonyabb felhasználásának érdekében — ezért a fejlesztések fragmentáltak maradnak, sokszor párhuzamosan zajlanak, és az eredmények nem épülnek egymásra. A Platform egyik küldetése éppen egy ilyen kommunikációs csatorna megteremtése és működtetése.

Meg kell említeni, hogy nem elhanyagolható akadályt jelent a pályázatok elkészítéséhez és a támogatások elszámolásához szükséges bonyolult adminisztráció működtetése, melynek költségei nem számolhatók el, és nem állnak arányban a kapott támogatás mértékével. Akadályt jelent a szakterületi fejlesztéseket célzó pályázati kiírások, a rendelkezésre álló pályázati támogatás jelentős visszaesése, és a meglévő pályázatokban az ipari szereplők számára általában előírt belső erőforrások hiánya is.

Összegzésképpen megállapítható, hogy a hazai nyelv- és beszédtechnológiában meglévő kitörési lehetőségek csak akkor realizálódhatnak, ha a vonatkozó kutatás-fejlesztési politikában és gyakorlatban mihamarabb jelentős változás történik. (Ehhez kíván segítséget nyújtani a jelen tanulmány.)

2.3. Küldetésünk

A Nyelv- és Beszédtechnológiai Platformot élenjáró magyarországi kutató-fejlesztő közösségek hozták létre azzal a céllal, hogy összehangolt munkával erősítsék és elősegítsék az innovációt a nyelv- és beszédtechnológia területén, így hozzájáruljanak a magyar technológiai fejlődéshez, a nemzetgazdaság versenyképességének

növeléséhez. A Platform hivatalos keretet nyújtva összefogja a jelentősebb hazai nyelv- és beszédtechnológiai kutatás-fejlesztést végző tudásközpontokat, és ezáltal

- elősegíti az eddig viszonylagos elszigeteltségben működő központokban felhalmozódott magas szintű tudás megosztását illetve integrációját;
- feltérképezi a nyelv- és beszédtechnológiai kutatásoknak a nemzetgazdaság számára legfontosabb fejlesztési és kutatási irányait a magyar adottságok (erőforrások, érdekviszonyok) figyelembevételével;
- részletes stratégiai és arra épülő megvalósítási terveket dolgoz ki, amelyek megvalósítását kialakított koordinációs eszközeivel a későbbiekben is elősegíti;
- javaslatait szakpolitikai csatornákon keresztül eljuttatja a kormányzat megfelelő szerveihez és segít azoknak a kormányzati stratégiákkal és megvalósítási tervekkel való összehangolásában;
- közvetíti az informatikai szektor érdekelt résztvevői felé a Platform elemzéseit, stratégiáit, javaslatait, megvalósítási programját és annak eredményeit;
- megjeleníti és képviseli a magyar szempontokat és érdekeket, valamint a hozzájuk kapcsolódó konkrét javaslatokat a nemzetközi központok és piaci szereplők számára;
- elősegíti a Platform eredményeinek tudatosítását a magyar gazdaság potenciális felhasználói felé, különös tekintettel a kis- és középvállalkozásokra.

3. Stratégiai célok

Bevezető

A magyar nyelv- és beszédtechnológiai kutatás-fejlesztés általános stratégiai célja az, hogy a nyelv- és beszédtechnológia az infokommunikációs technológiákon belül húzóágazattá fejlődhessen. Ehhez a magyar nyelv- és beszédtechnológiai fejlesztések stratégiájának az alábbi kérdésekben kell iránymutatást adnia:

- Melyek azok a kutatás-fejlesztési területek, ahova a ráfordításokat irányítani kell, és amelyek a versenyképesség növelését eredményezik? Figyelembe kell venni a rendelkezésre álló erőforrásokat, és ezeket a kiemelt kutatás-fejlesztési területekre kell koncentrálni, melyeket oly módon célszerű kiválasztani, hogy azok termékekben, szolgáltatásokban hasznosuló eredmények létrehozását szolgálják.
- Melyek azok a jelenlegitől eltérő kutatásfinanszírozási keretek, amelyek biztosítják a kutatás-fejlesztési erőforrásokat a tartós eredményesség érdekében, ösztönzik az ipari szereplőket saját kutatás-fejlesztési ráfordításaik növelésében, és megalapozzák a kutatóhelyek betöltéséhez szükséges személyi állományt?
- Mit lehet tenni annak érdekében, hogy a sikeres kutatási-fejlesztési projektek eredményei ne maradjanak a fejlesztő műhelyek zárt közösségén belül, a gyakorlati hasznosítás lehetőségét kizárva? Ennek érdekében miként lehet szorosabbá és szervezettebbé tenni a kapcsolatot a fejlesztésben és a hasznosításban érdekelt felek között?

A kutatás-fejlesztési tevékenységek tágabb kontextusát ugyan nem lehet figyelmen kívül hagyni, így általánosságban a világtrendeket követő pozícióból globális vezető helyre való előretörésre nincs reális alap, mindazonáltal a magyar nyelv sajátosságaiból adódó specifikus kihívásokra adott válaszokból származó eredmények „exportálhatók". Ennek kihasználása az ágazat világpiaci pozícióit már rövid (2-5 éves) távlatban is jelentősen erősítheti, ami indokolja a nemzeti nyelvre irányuló kutatás-fejlesztés stratégiai fontosságát.

A helyzetelemzésben felvázolt kedvezőtlen jelenségek és akadályok ellenére a magyar nyelv- és beszédtechnológia rendelkezik olyan jelentős erősségekkel, mint a szaktudás, élenjáró technológia, aktív nemzetközi kutatói kapcsolatok, amelyekre a sikeres előrelépés alapozható, amennyiben a kutatás-fejlesztési erőfeszítések és erőforrások az ország számára kitörési pontokat adó területekre összpontosulnak. A következő fejezet ezeket a területeket foglalja össze, valamint ismerteti a Platform által stratégiainak ítélt célokat, melyek elérését a 4. fejezetben tárgyalt eszközökkel és módszerekkel kívánja elősegíteni.

3.1. Nemzeti kutatási infrastruktúra kialakítása és szolgáltatása a nyelv- és beszédtechnológia területén

Az utóbbi években a kutatás-fejlesztés elsőrendű prioritásai között megjelent az integrált, egységes, mindenki számára elérhető és könnyen kiterjeszthető kutatási infrastruktúrák létrehozása. Az Európai Unió ESFRI (European Strategy Forum on Research Infrastructure) kezdeményezése, nagyszabású, számos európai intézményt magában foglaló és a Platform működési területét is érintő, magyar részvétellel is futó projektek (CLARIN, FLARENET, DARIAH), illetve a vonatkozó hazai vállalkozás (NEKIFUT) elindítása egyértelműen jelzik a kérdés stratégiai fontosságát.

A nyelv- és beszédtechnológia területén sikerrel alkalmazható módszerek és eljárások jellegéből (l. 4.1. fejezet) következik, hogy *korszerű kutatási eredmények és alkalmazások nem jöhetnek létre a megfelelő erőforrások, írott és beszélt nyelvi adatbázisok, alapvető sztenderdizált feldolgozó eszközök nélkül*; ezek a nyelv- és beszédtechnológia elengedhetetlen szükségletei a fejlesztésben és az elért eredmények kiértékelésében is. Számos területen voltaképpen ezek tartalmazzák a nyelvi tudás legnagyobb részét, a modern technológiák sok esetben „csupán" ennek a tudásnak a kivonatolását, használhatóvá tételét végzik.

A nemzeti nyelv- és beszédtechnológia hatékonyságáért, a Platform stratégiai céljaiért a legtöbbet a nyelvi erőforrások fejlesztésével, azok szolgáltatásával és alkalmazásával lehet tenni. A nyelv- és beszédtechnológia területén a nemzeti kutatási infrastruktúra kialakításának az elsődleges feladata a különféle hozzáadott értéket tartalmazó erőforrások definiálása, folyamatos létrehozása, illetve a meglévők menedzselése. Fontos kiemelni, hogy ezek a nyelvi adatbázisok mindenki számára szabadon elérhetővé és felhasználhatóvá kell, hogy váljanak.

3.2. Kutatásszervezés

Technológiatranszfer, kommunikáció. Az ipar és a kutatók közötti párbeszéd javítása érdekében szükség van az információátadás módszereinek fejlesztésére, a kutatás-fejlesztési eredmények és erőforrások rendszerezésére és hozzáférhetővé tételére, valamint hatékony kommunikáció kialakítására.

Létre kell hozni a terület technológiatranszfer-központját, amely a kialakítandó nemzeti kutatási infrastruktúrát a Platform által kidolgozott alapelvek szerint, a modern autorizációs és autentikációs technológiákat kihasználva egységes keretben (akár az egységes nemzeti kutatási infrastruktúra részeként) szolgáltatja, és hozzáférhetővé teszi mind a kutatási, mind az ipari szereplők, illetve akár a nagyközönség számára is. (Ez természetesen a korszerű hálózati technológiák korában nem jelenti az erőforrások egy adott fizikai helyre történő koncentrálását, hanem virtuális központként is értelmezhető.) Feladata továbbá az ipari szereplők kutatás-fejlesztési igényeinek felmérése, valamint az országban rendelkezésre álló tudás és a hozzáférhető eredmények, módszerek feltérképezése és az információ közvetítése a lehetséges partnerek felé. Ennek eszköze többek között a Platform által létrehozott nyelv- és beszédtechnológiai kutatás-fejlesztéssel kapcsolatos internetes portál, amely mind a szűkebb szakmai, mind pedig a nem szakmabeli érdeklődőknek szolgáltat információt, és széles körben ismerteti az új kutatási eredményeket.

A kutatás-fejlesztési eredményeket az ország határain túlra is exportálni kell, törekedni kell a magyarra kifejlesztett eszközök, módszerek más nyelvekre történő alkalmazására. Elsődleges célok lehetnek azok a környező országok, ahol viszonylag fejletlen az ágazat, mint például Szlovákia, Ukrajna vagy a volt Jugoszlávia egyes területei. Úgy válhatunk igazán *regionális központtá*, ha megmutatjuk, hogy a környéket segíteni tudó potenciállal is rendelkezünk.

Szabványosítás. Nemzetközi versenyképességünk növelése érdekében kulcsfontosságú, hogy a már létező és a létrejövő új technológiák megfeleljenek a meglevő szabványoknak, illeszkedjenek az egyre erősödő sztenderdizáló törekvésekhez. Ennek érdekében a ma használatos szabványokat széles körben ismertté kell tenni, az új szabványok kialakításában aktívan részt kell venni. Ki kell dolgozni egy nemzetközi gyakorlatba illeszkedő, összehasonlítható eredményeket biztosító kiértékelési módszertant, az ehhez szükséges szabványosított adatbázisok kifejlesztésével és az egyes területekhez kapcsolódó alapfogalmak meghatározásával.

Előtérbe kell helyezni a széles körben való felhasználhatóság, testreszabhatóság, fenntarthatóság és további fejlesztés elősegítése érdekében a nyílt forráson alapuló fejlesztéseket, figyelembe véve természetesen az üzleti érdekeltségeket.

Oktatás, kutatói utánpótlás. A kutatói utánpótlás képzését koordinálni kell, az egyes területek legkiválóbb szakembereit be kell vonni az oktatásba. A piac által felvehető munkaerő méretéből adódóan a szakképzésben résztvevők száma nem lehet tömeges, ezért az oktatás hatékonyságát növelheti a képzési erőforrások koncentrációja és egységesítése: azonos ismeretek oktatásához közös tananyagmodulok kidolgozása, ezek kommunikációs hálózatokon keresztül történő

szabad hozzáférhetősége. A fiatal kutatók számára ösztöndíjakat kell létesíteni, az ipar és az oktatási intézmények közötti kapcsolat megerősítésének keretében lehetővé kell tenni képzésük egy részének kihelyezését ipari szereplőkhöz.

Kutatásfinanszírozás. A kutatás-fejlesztés talpon maradása és a gyakorlati alkalmazás hatékonyságának javítása érdekében elengedhetetlenül fontos olyan kutatásfinanszírozási keretek kialakítása, amelyek

- lehetővé teszik interdiszciplináris, nagy költségigényű, de stratégiai fontosságú eredményeket hozó kutatások magasan képzett, a nemzetközi kapcsolatokat aktívan kihasználó kutatói teamek közreműködésével történő megvalósítását,
- hosszú távon biztosítják a megkérdőjelezhetetlen szakmai teljesítménnyel rendelkező műhelyek fennmaradását.

Ehhez a fiatal, tehetséges, az oktatásból kikerülő kutatói utánpótlást alkalmazni és megtartani tudó kutatóhelyek megteremtésén túl olyan pályázati kiírásokra van szükség, melyek meghatározott stratégiai területeket vesznek célba, és ahol az átlátható értékelési folyamat eredményeként a terület szempontjából releváns kritériumrendszer alapján, erős *szakmai* kontroll alkalmazásával a színvonalas, valódi innovációt tartalmazó pályaművek kapnak támogatást.

Együttműködés. Törekedni kell a hatékony és gördülékeny információcsere, az interdiszciplináris kutatás-fejlesztési tevékenység megalapozása érdekében a határterületekkel való együttműködés rendszeressé tételére. Különösen fontos a rokon technológiai területeken létrejött vagy szerveződő platformokkal, klaszterekkel történő együttműködés. Ennek egyik lehetséges formája a különböző szakterületek kiemelkedő teljesítménnyel rendelkező képviselőivel való rendszeres szakmai találkozás, szakmai rendezvények szervezése.

A szakpolitikai csatornákon keresztüli rendszeres konzultáció a kormányzat képviselőivel segít a szektor javaslatainak, stratégiájának a kormányzati stratégiákkal és megvalósítási tervekkel való összehangolásában.

3.3. Nyelvi információ kezelése, tárolása és feldolgozása

Nyelvalapú tudásmenedzsment. A digitális formában elérhető tartalmak robbanásszerű növekedése miatt a rendelkezésre álló képi, hangzó és szöveges információ további feldolgozás nélkül gyakorlatilag kezelhetetlen. Szinte nincs is az életnek olyan területe (tudomány, politika, gazdaság, oktatás, kultúra, adminisztráció stb.), ahol megengedhetnénk magunknak, hogy az elektronikus formában elérhető információkat ne hasznosítsuk. A hatékony információkezelés része az is, hogy kérdéseinkre több nyelven is releváns válaszokat kapjunk, amely rendkívül nagy fontossággal bír Magyarország nyelvi integrációja szempontjából. A nagy mennyiségű hangzó vagy szöveges információ feldolgozása során az alábbi feladatokat kell megoldanunk.

Egyfelől fontos, hogy a felhasználók felmerülő kérdéseikre minél hamarabb választ találjanak (***információ-visszakeresés***, *information retrieval*). Ennek a feladatnak a megoldását tűzték ki maguk elé a keresőmotorok fejlesztésével foglalkozó cégek, például a Google és a Yahoo!. A böngészők következő generációjának célja a ***szemantikai keresés*** és a lekérdezett információ ***strukturált megjelenítése*** (l. Google Squared, Wolfram Alpha, Bing és a megjelenés előtt álló Yebol), mely feladatok a nyelvi információra csak közvetetten támaszkodó statisztikai, gépi tanulási módszerek mellett magas szintű nyelvfeldolgozást is megkövetelnek. Ez a magyar vonatkozásában azt jelenti, hogy nekünk is ki kell, illetve tovább kell fejlesztenünk azokat az eszközöket, amelyek a weben található információ ilyen magas szintű hozzáférését lehetővé teszik. Ide tartozik a morfológiai egyértelműsítés, a szintaktikai elemzés és a tulajdonnév-felismerés.

Másfelől, a természetes nyelvi információ feldolgozásával nemcsak a releváns dokumentumokat szűrhetjük ki, hanem a strukturálatlan természetes nyelvi szövegben található információt adatbázisba szervezhetjük, hogy ezek hatékonyan lekérdezhetőek legyenek már létező adatbázis-kezelő technológiákkal (***információkinyerés***, *information extraction*).

A Platform tagjai már számos információkinyeréshez kapcsolódó kutatást végeztek és jelentős eredményeket tudnak felmutatni a szükséges részfeladatok megoldásában, azonban még számos nyitott kérdésre kell választ találni. [2]

A nyelvi kulturális örökség digitális korba való átmentése. A Platform stratégiájának középpontjában olyan technológiák állnak, amelyek egy életképes és rohamosan fejlődő nyelvi középhatalom képét vetítik előre. E célok mellett azonban a magyar nyelv- és beszédtechnológia **értékőrző**, sőt bizonyos esetekben **értékmentő** szerepéről sem szabad elfeledkeznünk. A magyar nyelvtechnológia még számos nyelvfeldolgozó eszközzel adós, például uráli nyelvrokonaink nyelveire. E nyelvek egy- és többnyelvű szótárainak, korpuszainak és egyéb erőforrásainak fejlesztése is elsősorban a magyar nyelvtechnológiától várható. Olyan kihalófélben levő rokon nyelvek, mint a nganaszan, a nyenyec, a mari vagy a komi nyelvi rendszerének dokumentálása, írott és hangzó megnyilatkozásainak digitalizálása és automatikus feldolgozása már az elmúlt években megkezdődött, és a jövőben is feladatunknak érezzük az értékmentő munka folytatását. Hangsúlyoznunk kell, hogy ezeket a nyelveket általában már csak néhány beszélő használja, vagyis a nyelvi jelenségek dokumentálása lehetőségének utolsó órájában vagyunk. Ezzel a célkitűzéssel a Platform teljes mértékben illeszkedik a világtrendek vonalába, amit az is mutat, hogy az amerikai Linguistic Data Consortium nemrég kifejezetten a kisebb nyelveket vette célba a „Ritkábban tanított nyelvek” (Less Commonly Taught Languages) program keretében.

[2] Új kihívás, hogy az elektronikus nyelvi tudás mind nagyobb mennyiségben hangzó anyagok formájában áll elő. Az ezekből történő információkinyerés és -visszakeresés első és egyelőre legkritikusabb lépése a beszéd-szöveg átalakítás. Tehát a nyelvalapú információmenedzsment multimédiás kiterjesztése érdekében kiemelt stratégiai cél a nagyszótáras, folyamatos többnyelvű gépi beszédfelismerés hatásfokának javítása (l. 4.3.).

A magyarországi és **határon túli magyar** nyelvváltozatokat feltérképező kutatásokban is jelentős támogatást tudnak nyújtani a nyelv- és beszédtechnológia művelői a beszélt és írott nyelvváltozatok digitális rögzítése és automatikus feldolgozása terén.

Az **automatikus szövegfeldolgozás** technológiái jelentős segítséget nyújtanak abban, hogy az ország írott kultúrkincsét a digitális korszakba átmentsük. A régi magyar szövegemlékek egyszerű beszkennelése még nem teszi hozzáférhetővé a bennük lévő szöveget, hanem szükséges a szöveg kinyerése, automatikus morfológiai és szintaktikai elemzése. Ez biztosítja a szövegek olyan részletes keresését és elemzését, amilyenre a nyelvtörténészeknek, kutatóknak valójában szükségük van, és amelynek elkészülte a magyar nyelvtörténet kutatásának hatalmas lendületet adhat. Hasonló értékmentő, az adott nyelv történeti korpuszának megépítését célzó projektek a világ minden táján folynak.

A **beszédfelismerési technológia** a nagy nemzeti hang/film/multimédia archívumok szövegtartalom szerinti kereshetőségét biztosíthatja. Az alaptechnológia már ma is elérhető magyar nyelven, azonban a speciális tartalmakhoz történő adaptáció (pl. régi filmhíradók nyelvi és hanganyagához történő lexikai, stilisztikai és akusztikai adaptáció) jelentősen növelheti a használhatóságot.

3.4. A természetes nyelven történő kommunikáció számítógépes támogatása

Természetes ember-gép kommunikáció. A **szűkebb értelemben** vett ember-gép kommunikáció legfőbb feladata az emberi igények közlése a gépekkel és a kapott válaszok hasznosságának növelése. A fejlődési tendenciák azt mutatják, hogy az embernek egyre kevésbé kell alkalmazkodnia a gépekhez, a gépek többféle módú kapcsolódást is elfogadnak, ezek a kapcsolódási felületek rugalmasan, a felhasználó képességeit és a környezetet is figyelembe véve alakulnak. Vagyis a kommunikációt természetesen és hatékonyan kiszolgáló új technológiák révén egyre könnyebben értjük meg egymást a számítógépekkel és egyéb elektronikai eszközökkel.

Az ember-gép kapcsolódási módok közül még mindig az érintésalapú kommunikáció a legelterjedtebb, ugyanakkor a legtermészetesebb emberi kommunikáció a beszéd: ez a hajtóereje az egyre nagyobb volumenű beszédalapú ember-gép kapcsolati kutatás-fejlesztéseknek. Szűk keresztmetszetet jelent a gépi beszédfelismerés és beszédértés emberi szinttől elmaradó hatásfoka, itt hosszú távú kutatások szükségesek. A gépileg előállított beszéd érthetősége, természetessége és stílusának a témához, beszélőhöz való illeszkedése is kulcsfontosságú a sikeres alkalmazásokhoz. Meg kell említenünk, hogy a beszédkapcsolat esetén automatikusan emberihez hasonló reakciókat várunk a géptől, így a dialógus- és mesterségesintellingencia-kutatás is előtérbe kerül.

A beszédfelismeréssel rokon, elsősorban a bemenő jel feldolgozását végző modul megváltoztatását igénylő feladatok az írás-, jelbeszéd- és gesztusfelismerés. Ezek jelentőségét látjuk a mostaninál természetesebb multimodális interfészekben, melyek például a szemgolyó mozgásának követésével arra is odafigyelnek, hogy mire néz éppen az ember. Célunk, hogy rövid időn belül élőszó és/vagy

gesztusok segítségével is lehetővé váljon az internet böngészése és általában az emberi inputot igénylő számítógépes programok irányítása.

Tágabb értelemben véve az ember-gép együttélésen azt értjük, hogy az ember többletképességeket kaphat a gépektől. A gépek segítenek bizonyos funkciókat, például az értékelés, a diagnosztika vagy a döntés-előkészítés területén. A gépek részben vagy egészben át is vehetnek bizonyos funkciókat, például az információfeldolgozás, statisztika, megjelenítés és tájékoztatás területein. Általában a hihetetlen tömegű információ közti tájékozódást, akár a szakember, akár a laikus számára, rendkívüli módon megkönnyíti a nyelvtechnológia. Továbbá a nyelv- és beszédtechnológiai fejlesztés eredményeit hasznosító, az oktatás hatékonyságát növelő szoftverek kiválóan alkalmazhatók a logopédiában, az idegennyelv-oktatásban és a magyar mint idegen nyelv tanításában egyaránt. (A gyakorlati felhasználási területekről részletesebben l. az 5. fejezetet, illetve a Jövőkép *Kiemelt alkalmazások* fejezetét.)

Fogyatékkal élők és hátrányos helyzetűek információs társadalmi integrációjának elősegítése. A nyelv- és beszédtechnológia fejlesztéseit alkalmazó infokommunikációs eszközök komoly elősegítői nemcsak a gazdaság fejlődésének, hanem az **esélyegyenlőség** és az **életminőség** javításának is. A fogyatékkal élők társadalmi integrációjának elősegítésében kulcsfontosságú az ember-gép kommunikáció megkönnyítése. A tudásalapú társadalomban az integráció elengedhetetlen lépése, hogy olyan tartalmakhoz is hozzájussanak a fogyatékkal élők, amelyeket számukra primér módon nem hozzáférhető médiumokon keresztül közvetítenek. A beszédszintézisre és -felismerésre alapuló technológiák, amelyek más médiumokra „fordítanak” és tesznek elérhetővé információt, mind a siketek és nagyothallók, mind a vakok és gyengénlátók számára ezt az integrációs lépést könnyítik meg.

Különösen fontos a tanulásban akadályozott vagy nyelvi zavarral küzdő **gyermekek** felzárkóztatása az oktatásban, hiszen a nyelvi készségek alsó tagozatban történő fejlesztése teremti meg az alapját annak, hogy későbbi tanulmányaik során az értelmi képességüknek megfelelő nyelvi teljesítményt tudjanak nyújtani.

A szociálisan hátrányos helyzetű tanulók esélyegyenlőségére való törekvésben is kulcsszerepe lehet az iskolai környezetben alkalmazott nyelvtechnológiának, azaz a tanulók információs társadalomba való integrálásának. Az informatika rohamos fejlődésének következtében a hardverek árcsökkenése Magyarországon is egyre közelebb hozza azt az időt, amikor a számítógéppel közvetíthető tudás elérhető lesz mindenki számára. Elengedhetetlen az olyan szoftverek kidolgozása, melyek célja nemcsak a logopédiai vagy részképesség-fejlesztés, hanem az általános szókincs és kifejezőkészség javítása is. Különösen fontos lehet ez utóbbi a magyart második nyelvükként beszélő tanulók számára. A magyar nyelvre készített alkalmazások fejlesztésén kívül a magyar nyelvtechnológia feladata az országban kisebbségként élő közösségek nyelvén elérhető alkalmazások fejlesztése is.

Többnyelvűség az Európai Unióban, a nyelvi korlátok leküzdése. Az Európai Unió fontos elve a nyelvek sokféleségének tisztelete és a nyelvi alapon

történő megkülönböztetés tilalma. Az EU 23 hivatalos nyelve **egyenrangú**. Az „ahány nyelven tudsz, annyi ember vagy" mottó jegyében kialakított EU többnyelvűségi politika három célkitűzése, hogy:

1. támogassa a nyelvi sokféleséget, ösztönözze a nyelvtanulást, Unió-szerte elősegítse hivatalos nyelveinek mind szélesebb körű ismeretét és használatát;
2. a több nyelven folyó munka költségeinek leszorításával elősegítse az egészséges többnyelvű gazdaságot az egységes európai piacon;
3. lehetővé tegye, hogy anyagi helyzetétől, egészségi állapotától és lakóhelyétől függetlenül valamennyi európai polgár élvezhesse az információs társadalom előnyeit, saját nyelvén jusson hozzá az uniós információkhoz.

A fentieknek megfelelően tehát cél, hogy bármely nyelven nyilvánosságra hozott hangzó vagy írott közlemény az EU bármely polgára számára egyenlő eséllyel hozzáférhető legyen. Ami ennél is fontosabb, hogy a befogadó az információt **meg tudja érteni**, vagy legalábbis a releváns tartalmat egyszerűen ki tudja nyerni belőle. A nyelvtechnológiai kutatások egyik stratégiai célja éppen ez: a (nagy mennyiségű) természetes nyelven megfogalmazott információ megértésének számítógépes támogatása, illetve az automatikus gépi megértés megvalósítása. A nyelvtechnológia számos szinten és területen segítheti az idegen nyelvű szöveget olvasó embert, támogathatja az **emberi** megértést. Ide tartoznak az automatikus gépi fordítás, a fordítástámogató eszközök, a többnyelvű információkinyerés és információ-visszakeresés (pl. könyvtárakban, katalógusokban), a megértéstámogatás, a számítógéppel segített szótárkészítés, a nyelvoktatásban használható nyelvtechnológiai eszközök, illetve a beszédtechnológiával együtt az automatikus tolmácsolás, azaz a beszéd „online" fordítása is elérhető közelségbe kerül egyes alkalmazásokban.

A **gépi** megértésre irányuló kutatásokban egyrészt cél a természetes nyelvű szöveg megértésére képes technológia kifejlesztése, másrészt pedig a jelenlegi eszközökkel már automatikusan megérthető tartalom (nagy volumenű) létrehozása is: ontológiák, tudástárak építése. E két kutatási megközelítés összefonódásának eredményeképpen valósulhat meg a következő évtizedben a szemantikus web, azaz válhat géppel automatikusan értelmezhetővé az egymással szemantikus kapcsolatban álló adatok és tartalmak tömege.

4. A közeljövő kutatási területei

A fentebb ismertetett stratégiai célok elérésének érdekében részben alap- és célzott speciális kutatásokra, részben integratív, az egyes — egymástól gyakran igen távol esőnek látszó — szűkebb szakterületek kutatásait összefogó kutatásfejlesztésre van szükség. A következőkben konkrét szakmai — de reményeink szerint közérthető — javaslatokat teszünk, szem előtt tartva egyrészt a világtrendeket, másrészt a magyar nyelv egyedi jellegzetességeiből fakadó kihívásokat, az ezekkel járó előnyöket és hátrányokat.

4.1. Általános módszertani alapelvek

A kutatás magas színvonalának megtartása és biztosítása érdekében fontosnak tartjuk a Platform konszenzusán alapuló általános módszertani alapelvek megfogalmazását. Ezek egyrészt iránymutatásként szolgálhatnak a folyamatosan bővülő Platform tagjai számára, másrészt a pályázatkiírók és -értékelők munkáját is segíthetik.

Szabályalapú vagy statisztikai módszerek?. A különböző tudományterületeken gyakorta feltett kérdésre a mi válaszunk nem „vagy", hanem „és". A nyelv- és beszédtechnológiában mára a statisztikai megközelítések sokszor már megkérdőjelezik a szabályalapú megoldásokat, azonban minden statisztikai rendszernek lényegi részei egyes szabályok, tehát tisztán statisztikai rendszer nemigen létezik. Ugyanakkor a tisztán szabályalapú megoldások sem nevezhetők életképesnek a nyelv- és beszédtechnológiában, hiszen a „nyelvi helyesség" nem feltétlenül objektív fogalom, a valós nyelvhasználatot csak valamiféle statisztika képes visszaadni, tehát legalább a technológiakiértékelés szintjén a statisztika kiküszöbölhetetlen.

A szabályalapú módszerek rendkívül erőforrás-igényesek, ugyanakkor a statisztikai alapú gépi tanulás is drága, ha ún. felügyelt tanítású technikákat használunk. Márpedig ez a leginkább bevált és használt technológia szerte a világban. Ilyen például a gépi beszédfelismerés szinte egésze, ahol nagy mennyiségű pontos kézi átirat szükséges a hanganyagok mellett, de ilyen a tulajdonnév-felismerés is, amelynek során a rendszer tanításához és kiértékeléséhez is kézzel annotált korpuszokat használunk. A szükséges emberierőforrás-igény csökkentésére és a fejlesztések gyorsítására ezért előtérbe kerültek a felügyelet nélküli módszerek. Ezek azonban belátható időn belül csak kisebb részben tudják helyettesíteni a felügyelt technikákat. Fontos irányzat a részben felügyelt tanítás, ahol az ember általi ellenőrzés (hanganyag kézi leirata, címkék stb.) géppel segített módon készül a nagyobb hatékonyság érdekében.

Összehasonlíthatóság, megalapozottság. A nyelv- és beszédtechnológia alkalmazásai esetében megkerülhetetlen kérdés az egyes megoldások összehasonlíthatóvá tétele. Kívánatos, hogy a technológiakínálat sokszínű legyen, de az is, hogy sztenderd módszerek szerint összehasonlíthatók legyenek a szolgáltatók technikái. Ez mindig alkalmazásfüggő, de a hivatalos adatbázisokon, rögzített módszerekkel mért eredmények eligazítást nyújthatnak mind az alkalmazók, mind a pályázatok elbírálói számára. Az összehasonlításhoz szükséges adatbázisok elkészítése és szolgáltatása, a kiértékelési szabályrendszer kidolgozása nonprofit feladat, melyben a Platform szerepet vállalhat. Nemzeti technológiai „értékelő fórumok" nemcsak az összehasonlíthatóságot biztosíthatják, de egészséges versenyt is generálhatnak az ország és az iparág javára.

Hangsúlyozzuk, hogy kutatás-fejlesztési eredmények mindig számszerűsíthető, lehetőleg sztenderd, de statisztikailag megalapozott formában fogadhatók csak el, a szubjektív tesztek csak illusztratív jelleggel bírnak.

4.2. Infrastruktúra és erőforrások fejlesztése

A nyelv- és beszédtechnológia területén végzett érdemi innovációs tevékenység nélkülözhetetlen feltétele a korszerű nyelvi erőforrásokból, alapvető feldolgozó eszközökből álló színvonalas kutatási infrastruktúra. Ezért folyamatosan szem előtt kell tartani ezen adatbázisok és eszközök készítését és továbbfejlesztését, valamint ki kell alakítani a lehető legegységesebb feldolgozási, illetve alkalmazási protokolljukat. A legalapvetőbb nyelv- és beszédtechnológiai erőforrások közül számunkra két típus emelendő ki: egyrészt a magasabb szintű nyelvi elemzést tartalmazó lexikai erőforrások, amelyekre a legkorszerűbb szemantikus technológiák épülnek (l. Jövőkép *A szemantikus technológiák* c. fejezet); másrészt a nagyméretű, különböző nyelvi információval ellátott (*annotált*) szöveg-, illetve beszédadatbázisok (*korpuszok*), amelyek mindenfajta statisztikai alapú eljárás alapjául szolgálnak.

A magasabb szintű nyelvi elemzést, szemantikai információt tartalmazó lexikai erőforrások mindazon alkalmazásoknak az előfeltételei, amelyeknek célja (többek között) az emberi nyelv gépek általi megértése, ami a nyelv- és beszédtechnológiai kutatások egyik legfőbb stratégiai célja. Ahhoz, hogy a nyelvi információt tartalmilag megjelölt egységekbe szervezzük, olyan tudásbázisok fejlesztésére vagy magyar nyelvre való adaptálására van szükség, amelyek nyelvfüggetlen, ám a természetes nyelvűnél pontosabb definíciókat tartalmaznak és feleltetnek meg nyelvi jeleknek. Az ilyen tudásbázisok, ún. **ontológiák** a nyelvtől független, a világ jelenségeire vonatkozó tudást tartalmaznak gépi feldolgozás számára hozzáférhető, szisztematikus módon. Fontos tehát, hogy általánossá és szabványossá váljon az ontológiákban tárolt tudás reprezentációs módja (RDF (Resource Description Framework), OWL (Web Ontology Language), XML (eXtensible Markup Language)), valamint ezeknek a magyarra való honosítása.

Az ontológiák egyrészről általános tudást tartalmaznak, másrészről egy-egy szakterület specifikus tudásanyagát is reprezentálhatják – utóbbi típusú ontológiák a szakontológiák. Nyilvánvaló, hogy az utóbbiak megfelelő kialakításához az adott szakterület magas szintű ismeretére van szükség. Mivel a stratégiailag fontos tudományterületek (orvostudomány, jogtudomány, mérnöki tudományok) szakemberei általában nehezen elérhetők és idejük nehezen megfizethető, különösen fontos lenne megfelelő anyagi forrásokat találni a közös munkához és erősíteni az együttműködést ezen területek képviselőivel.

Az erőforrások kifejlesztése mellett fontos az erőforrások feldolgozása, megosztása és elérhetővé tétele is. Az erőforrások egységes megjelentetése, hozzáférhetővé tétele, valamint a nemzetközi nyelvtechnológiához való kapcsolódásunk szempontjából sarkalatos kérdés **a magyar BLARK (Basic Language Resource Kit) nyelvtechnológiai alapeszközkészlet** kifejlesztése és közzététele. Ennek fontos hozadéka lesz, hogy a már rendelkezésre álló elemzési megoldások használható, szabványos formában elérhetőek lesznek mind a magyar nyelvtechnológusok, mind a magyarral foglalkozó külföldiek számára.

Az alábbiakban felsoroljuk, mely új nyelvi erőforrások előállítását tartjuk kiemelkedően fontosnak:

1. Magyar nyelvű beszélt nyelvi adatbázisok
   Kiemelt jelentőségű, hogy nagyméretű, szöveges leirattal rendelkező különféle beszédstílusú beszédadatbázisok készüljenek magyar nyelven is. Noha számos jó minőségű tervezett (olvasott) beszédet tartalmazó adatbázis készült el a Platform tagjainak a közreműködésével is, a nemzetközi szinten elfogadott adatbázisméretektől általában egy-két nagyságrend lemaradás tapasztalható. Elsősorban a gépi beszédfelismerésnél lényeges, hogy nagyobb méretű adatbázisok szülessenek a statisztikai nyelvi és akusztikai modellek jobb becsülhetősége és így a nagyobb felismerési pontosság érdekében. A méret mellett ugyanakkor nagyon fontos, hogy ne csak döntően olvasott, hanem inkább kevésbé tervezetten előállított, de spontán vagy ahhoz közeli beszéd kerüljön rögzítésre. Hiszen természetszerűleg az ilyen jellegű beszéd (beszélgetés ember-gép, ember-ember között) szöveggé alakítása a tipikus, élet- és alkalmazásközeli feladat. Ilyen esetekben a hangkapcsolateloszlást nem lehet előre tervezni, ezért csak a jelentős (tipikusan több mint 100 órás) adatbázisméret tesz lehetővé reprezentatív mintavételt. Lényeges, hogy a beszélők száma, kora, neme stb. is jól kövesse a megcélzott réteget. Megjegyezzük, hogy a gépi beszédfelismerés mellett beszélőazonosításra, dialógusmodellezésre és általános fonetikai, morfológiai, korpusznyelvészeti kutatásokra is rendkívül jól használhatók az ilyen nyelvi erőforrások. A következő típusú beszédadatbázisok elkészítését javasoljuk elsősorban:
   - Spontán monológok (pl. diktálási alkalmazáshoz).
   - Spontán beszélgetések (pl. banki ügyfélszolgálati beszélgetések monitorozásához).
   - Több résztvevős megbeszélések (pl. üzleti, szakmai megbeszélések automatizált lejegyzéséhez).
   - Telefonos üzenetek (pl. automatikus hangpostaátíráshoz).
   - Telefonos beszélgetések, telekonferenciák (pl. telefonos ügyfélszolgálatok minőségbiztosításához).
   - Multimédia híranyagok és beszélgetések (pl. a beszédtartalom szerinti kereshetőség biztosításához, automata feliratozáshoz).
   - Magyar (hangzó) nyelvváltozatok digitális rögzítése.
2. Idegen nyelvű beszélt nyelvi erőforrások
   Ezek fontosságát, jelentőségét az adja, hogy ma a beszédtechnológia nagy része (és a nyelvtechnológia mind nagyobb része) algoritmikusan nyelvfüggetlen, tehát a magyar nyelvű tapasztalatok adott esetekben nagyon jól kiterjeszthetők más nyelvekre. Elsősorban a közép-kelet-európai nyelvek jönnek számításba részint a kulturális és egyéb hasonlóságok, részint a piac nyitottsága miatt.
   - Közép-kelet-európai nyelvekre a fenti típusú adatbázisok előállítása.
   - A nagy nyugati és esetleg keleti nyelvekre (FIGS, JCK) az adatbázis-beszerzés segítése.
   - Párhuzamos beszélt nyelvi korpuszok kialakítása beszédfordítás céljára.
   - Kihalóban levő uráli nyelvrokonaink hangzó nyelvi anyagainak rögzítése és digitalizálása.

   - A magyarországi kisebbségek hangzó nyelvi anyagainak rögzítése és digitalizálása.
3. Magyar nyelvű írott nyelvi erőforrások
   Fontos hangsúlyozni, hogy bár írott nyelvű tartalom egyre nagyobb mennyiségben érhető el a weben, — éppen ezért — folyamatosan nő az igény az intenzív feldolgozással (különféle szinten történő címkézés, strukturálás stb.) jelentős hozzáadott értéket hordozó tartalmakra, melyek alapvető erőforrásként szolgálnak az információkinyerési, -visszakeresési és számos további nyelvtechnológiai alkalmazás számára. Ezen felül a nyelvi örökség megőrzése (l. 3.3.) és hozzáférhetőségének biztosítása szempontjából is kívánatos az alábbi adatbázisok létrehozása.
   - A magyar nyelv különböző nyelvváltozatainak írott korpusza.
   - Az egyes mondatrészek közötti függőségek teljes annotációját tartalmazó korpusz (ún. dependency bank) a mélyebb szintaktikai elemzés megvalósításához.
   - Az egyes szaknyelvek (jogi, orvosi stb.) korpuszai és az ezekhez tartozó szakontológiák.
   - Szemantikai információt tartalmazó lexikai erőforrások előállítása.
   - Változatos szövegtípusokból álló, kézzel tulajdonnév-annotált referenciakorpusz.
   - Megfelelő lefedettségű, **a magyar WordNettel** (l. Jelenkép *Magyar WordNet* rész) is összekapcsolódó tanulói szótár, a magyar nyelv népszerűsítése érdekében.
4. Idegen nyelvű írott nyelvi erőforrások
   A következőkben csak azokat az idegen nyelvű korpuszokat említjük, melyeknek alighanem az egyetlen esélye a digitális fennmaradásra és hasznosításra, ha magyarországi kezdeményezés karolja fel az ügyüket.
   - Kihalóban levő rokon nyelvek korpuszai.
   - Magyarországi kisebbségi nyelvi írott korpuszok.
   - Párhuzamos írott nyelvi korpuszok építése automatikus szótárgenerálás, illetve gépi fordítás céljára; elsősorban kevésbé kutatott közép-kelet-európai nyelvekre, ahol nagyobb magyar kisebbség él.

4.3. A gépi beszédfelismerés kutatási irányai

A legtermészetesebb emberi kommunikáció a beszéd, ezért a beszéd szövegtartalmának automatikus felismerése a modern kor egyik legjobban áhított eszköze. Az emberi hatékonyságot elérő beszéd-szöveg átalakítás, dallamfelismerés stb. azonban a korábban elképzeltnél sokkal nehezebb feladatnak mutatkozik, ezért azt gondoljuk, hogy e cél elérésének dátumát bölcsebb nem előrevetíteni. Ugyanakkor biztosak vagyunk benne, hogy koncentrált erőfeszítésekkel folyamatos haladást lehet elérni a gépi beszédfelismerés majd minden területén.

A jelenlegi technológiai szint — ahogy a Jelenképben is bemutatjuk — számos gyakorlati alkalmazáshoz szolgálhat alapul. Ugyanakkor, hazai és nemzetközi tekintetben is, az alább felsorolt kutatási irányokban történő előrelépés exponenciálisan tágíthatja az új szolgáltatások, termékek körét.

Robusztus beszédfelismerési technikák. Adott témakörre és beszélőre specializált, közelbeszélő mikrofon melletti beszédfelismerés pontossága igen magas is lehet — ameddig a háttérzaj nem hallható, vagy lényegesen alacsonyabb a szintje, mint a felismerendő beszédé. Amint a zavaró jel szintje emelkedik, a szófelismerési pontosság rohamosan — az emberi felismerési teljesítménytől gyorsan és jelentősen leszakadva — csökken. Ennek egyik alapvető oka az, hogy az alkalmazott jelfeldolgozás, mely a hangnyomás-idő függvényből állapítja meg a beszéd akusztikai lényegét, meg sem közelíti az emberi hallás lényegkiemelési képességeit. Ehhez kapcsolódóan a másik fő problémát ott találjuk, hogy a beszédfelismerés elemi akusztikus egységeinek modelljei is túlegyszerűsítettek, és a gépi modellezési és lényegkiemelési fázisok az emberi feldolgozással ellentétben teljesen különválnak.

A téma hosszabb ideje folyamatos kutatás tárgya, azonban a zajrobusztusság tekintetében igazán jelentős előrelépés az elmúlt évtizedekben nemigen mutatkozott, mivel sokáig nem volt világos, hogy ez a terület képezi a beszédfelismerés szűk keresztmetszetét. Másrészt az emberi hallásról is nagyon keveset tudunk: sem a fizikája, fiziológiája, sem a kognitív, neurológiai vonatkozásai nincsenek kellő mértékben feltérképezve. További nehézséget jelentett az, hogy az összetett pszichofizikai-matematikai modellek olyan nagy számításigényűek, hogy néhány évvel ezelőttig nem is volt reális esélye kivitelezésüknek.

A probléma nehézségét reálisan látva, a területen folyó kutatások kiemelt gyakorlati jelentőségére tekintettel feltétlen hangsúlyoznunk kell annak szükségét, hogy e terület az eddigieknél jóval nagyobb támogatásban részesüljön. Mivel ez a beszédfelismerés hatékonyságát legjobban korlátozó szűk keresztmetszet, ha ezen a területen sikerül előrelépni, az a beszédfelismerés minden ágában azonnali pozitív hatással mutatkozik. Másrészt a zajrezisztencia kialakítása nyelvfüggetlen, tehát nincsenek előnyben az adatbázisokkal jobban ellátott nemzetközi kutatóműhelyek. Harmadrészt azért is alkalmas lehet a magyar kutatóközösség a feladatra, mert nemcsak hagyományokkal és tapasztalatokkal rendelkezik e téren, de a magyar orvoslás, biológiai-fiziológiai kutatások is igen magas színvonalúak, illetve a nemzetközi kapcsolatrendszerünk is segítheti az ilyen irányú eredmények hatékony elérését.

A téma hatékony műveléséhez kislétszámú elkötelezett és magasan kvalifikált kutatócsoport(ok) hosszabb távú (5-10 év) állandó és motiváló támogatása szükséges. Ennek várható költsége nemzetgazdasági szempontból elhanyagolható, haszna viszont igen jelentős lehet.

Spontán társalgási beszéd felismerése. A legjobb akusztikai lényegkiemelés esetén is problémát jelenthet a laza artikuláció és a spontán beszédben tipikus gyors beszédtempó. További nehézség, hogy a szöveges tartalmat gyakran bennfentes téma határozza meg, azaz a lexikon és a nyelvi modell nem lehet elég felkészült az ilyen esetekre. Ugyanakkor a természetes kommunikáció jelentős része ebbe a kategóriába esik, tehát a gyakorlati alkalmazások szempontjából kiemelt fontosságú a terület.

Ehhez egyrészt a témához illeszkedő adatbázisok használata, másrészt a kiejtési modellek beszédstílusra, tempóra való specializálása szükséges. Ezeken felül várhatóan a beszélőváltások vizsgálata, az automatikus beszélőadaptáció, valamint a lexikális és nyelvi adaptáció segíthet sokat a felismerési pontosság érdemi növelésében.

Nagyszótáras folyamatos beszédfelismerés gazdag morfológiájú nyelvekre. Az ilyen nyelvek — köztük a magyar, finn, török, arab — ma a beszédfelismerési kutatások egyik kiemelt helyén szerepelnek. Itt az okoz problémát, hogy míg a beszédfelismerés kimenetén szavak sorozatát várjuk, az ilyen nyelveket szavakkal és azok kapcsolataival közvetlenül modellezni szinte lehetetlen. Míg angolra 60.000 szavas szótárral szinte minden beszédfelismerési alkalmazás jól elboldogul, magyarra hasonló lefedettséghez akár milliónál is több szót tartalmazó szótár kellene. Az igazi probléma azonban a szókapcsolatok modellezésénél következik, a tipikus modellezési megközelítésnél két szó alapján következtetünk a harmadik valószínűségére, azaz a szókapcsolatok száma köbösen emelkedik. Végül terabájtos memóriaigények lépnének fel az „egyszerű" szöveg-beszéd átalakítási feladatoknál.

A probléma kezelésében már jelentős eredmények születtek elsősorban finn kutatók munkája alapján, és a magyar nyelv tekintetében elértekre is büszkék lehetünk. A probléma azonban még korántsem tekinthető megoldottnak: a jelenlegi eljárások főleg tervezett beszéd esetén hatékonyak, valamint egyes nyelvekre (mint a török és arab) jelenleg még nem sikerült áttörést elérni. A magyar kutatóknak tehát más nyelvű nemzetközi kutatásokba is érdemes lehet bekapcsolódniuk, hiszen egyrészt a kutatási tapasztalatokat is kamatoztathatják, másrészt a magyar anyanyelv is sok segítséget jelenthet.

Nyelvfüggetlen beszédfelismerő módszerek kialakítása, célcsoport: a közép-kelet-európai nyelvek. Ma már nem csak az fontos, hogy egy adott nyelven minél nagyobb beszédfelismerési pontosságot érjünk el, hanem az is, hogy milyen gyorsan sikerül a technológiát az adott nyelvre adaptálni. A Platform kutatói ezen a területen is tettek fontos előrelépéseket: a nyelvi sajátságok ismeretét nélkülöző beszédfelismerési technológiáról mutatták meg, hogy a magyar nyelv esetén sem marad el szignifikánsan a sztenderd módszerekkel elért eredményektől.

Kihasználva, hogy a nagyobb nemzetközi beszédtechnológiai cégek a költséghatékonyság miatt (a sztenderd technikák drágasága és a kisebb populáció miatt) a közép-kelet-európai régiót nem tekintették célcsoportjuknak, a magyar kutatók és fejlesztők számára különleges lehetőség mutatkozik. A világszínvonalhoz közeli alaptechnológia, a helyismeret és a már elért eredmények gyors és olcsóbb beszédfelismerő rendszerek kialakítását teszik lehetővé a környező országok nyelveire.

Itt elsősorban alkalmazott kutatásra és kísérleti fejlesztésre van szükség. A feladat nagy, de elég jól átlátható, ütemezhető, ami tehát gazdasági szempontból jól kezelhető.

4.4. A gépi beszédelőállítás kutatási irányai

A gépi beszédelőállítást sokan megoldott problémának tekintik, ám az emberével minden körülmény között összetéveszthető gépi beszéd előállítása még mindig távoli cél. Egyes szűkebb témakörökben és sok kézi munka árán megtévesztően élethű beszéd állítható elő, azonban az általános és hibátlan témafüggetlen automatikus szöveg-beszéd átalakítás még utópia. Továbbá az általános célú szövegfelolvasó szoftver is a nehezen elérhető célok közé tartozik, mivel a felolvasási technológiákat témához, célközönséghez, műfajhoz kell kötni, és kevés az olyan terület, ahol le lehet mondani az automatikus felolvasók folyamatos emberi tanításáról, támogatásáról. A bemenő szövegekben mindig lehetnek olyan részek, amelyeknek a kiejtését eddig még nem rögzítették elektronikusan: ezek a kivételes írásmódú és kiejtésű szavak. Mindezekből következik, hogy a hibamentes automatikus szövegfelolvasás eléréséhez némi emberi támogatásra sokáig szükség lesz, ennek csökkentésére átfogó kutatásra van szükség. A korszerű megoldásokhoz itt is nagyméretű és több szinten pontosan címkézett beszédadatbázisokat kell felépíteni.

A gépi szövegfelolvasás megítélésének három fő kritériuma van: helyes-e a kiejtés (szegmentális szint), helyes-e a hangsúlyozás, a beszéddallam és a ritmus (szupraszegmentális szint), valamint hogy emberi hangszínezete van-e a szintetizátornak. Az ebbe a kritériumrendszerbe illeszkedő, általunk fontosnak tartott fejlesztési területeket vázoljuk fel a következőkben.

Skálázható kiejtésátíró szoftver és kiejtési szótárak fejlesztése. A fenti kritériumrendszer első elemét érinti a korrekt hangátírás. Magyar nyelvre jelenleg még nem létezik olyan szoftver, amely tesztelt és minősített kiejtési átírást valósít meg, esetleg hangolható, skálázható (minden kutatóközösség a saját szempontjai szerint alakít ki nem teljes megoldásokat). Emellett az egyes szakmákat érintő szakszavak kiejtési szótárait kell elektronikus, egységes, szabványosított formában elkészíteni. Ezzel a munkával csak csökkenteni lehet a jövőbeni emberi támogatás nagyságát, azt teljesen kiküszöbölni nem lehet, mert mindig lesznek olyan szavak, kifejezések, amelyeknek a kiejtését legalább egy alkalommal meg kell határozni. Javasoljuk egy központi kiejtési adatbank létrehozását, ahonnan a jövő nyelv- és beszédtechnológiai rendszerei lekérdezhetik a szükséges adatokat.

Hangsúlykijelölés szöveganalízis alapján. A helyes hangsúlyozás megvalósítása az automatikus szövegfelolvasás lényeges eleme. Kezdeti sikereket elkönyvelhetünk ezen a téren, de az átfogó megoldáshoz nagyobb erőforrásokat kell mobilizálni mind nyelvészeti, mind informatikai területről. Magyar nyelvre jelenleg nem létezik sem szabály-, sem statisztikai alapú szoftver, amely a szöveg elemzése alapján képes lenne a mondat szavaira a helyes hangsúlykiosztást teljes komplexitásában elvégezni. Megjegyezzük, hogy az automatikus hangsúlykijelölés hiánya kihat a szövegkivonatolási technológiák teljesítőképességére is, hiszen nehéz a lényeget kiemelni egy szövegből, ha nem tudjuk, hogy mely szavak a hangsúlyosak.

Az emberi hangszínezet közelítése. A hangkarakter-transzformáció a kifejezésforma bővítését teszi lehetővé. Adott egy általános paraméterhalmaz a beszédszintézishez (a hétköznapi beszéd általános alapjellemzői). Pótlólagos jellemzők hozzáadásával elérhető, hogy a szintetizált szöveg érdes, bársonyos, rekedt, suttogó, levegős hangszínezettel szólaljon meg.

A kiejtés stílusára (parancsoló, leíró, határozott, magyarázó stb.) jellemző paramétercsoportok kutatása még gyermekcipőben jár. Az emberek közötti párbeszédben fontosak az ilyen kiejtési stílusok, amelyek természetesen összekapcsolhatók a kimondandó szöveg tartalmával.

A kiejtési formák fontos csoportját alkotják az érzelmi töltést kifejező beszédformák (pl. mérges, bosszús, álmodozó, szomorú, vidám stb.). Az érzelem kifejezésének akusztikai fogódzóit már világszerte kutatják; magyar vonatkozásban a kezdeti kutatások pár éve indultak el. A jövő beszédszintetizátoraival szemben támasztott alapvető követelmény lesz, hogy érzelmeket hangban ki tudjanak fejezni.

Hasonlóan a jövő egyik ígéretes kutatási iránya a spontán beszédstílus megvalósítása. A szituációhoz illő gépi hang jellegzetességeinek kutatása még csak csírájában lelhető fel mind a magyar, mind más nyelvek vonatkozásában. Az adott személy hangjára való transzformáció (hangutánzás) is fontos eleme lesz a következő évtized beszédtechnológiájának. A megrendelő felolvas egy adott szöveget, és az általa megvásárolandó beszédszintetizátor hangját a gyártó a megrendelő hangjára hangolja. Így minden embernek lehet majd egy saját hangú szövegfelolvasója. Ez komoly piaci érdeklődésre tarthat számot.

Többnyelvű szintézist támogató keretrendszer fejlesztése. A statisztikai és fonetikai módszerek ötvözésével, valamint a megfelelő nyelvi modulok kialakításával olyan általános keretrendszerek fejleszthetők ki, amelyekkel más nyelvekre is ki lehet terjeszteni a szövegfelolvasást (például e-mailfelolvasóban az idegen nyelvű szót vagy esetleg teljes levelet a beszédszintetizátor nyelvváltással tudja felolvasni). Az ilyen kutatás kétirányú lehet. Kívánatosak olyan megoldások, amelyekben a magyar nyelvű beszédszintetizátor más nyelven is meg tud szólalni (érezhetően magyar akcentussal, de helyes kiejtéssel). A másik irány, amikor nem magyar nyelvű területre szánják az idegen nyelvű szintetizátort, hanem saját nyelvterületére. Ilyenkor nem magyar akcentusra kell tervezni a rendszert.

4.5. A gépi fordítás és fordítástámogatás kutatási irányai

A gépi fordító rendszerek speciális helyet foglalnak el a nyelvi rendszerek között. Az első számítógépek megjelenése után sokan úgy gondolták, hogy a gépi fordítás lényegében egy (át)kódolási feladat, ami rövid időn belül megvalósítható lesz. A kezdeti lelkesedést kudarcok követték; rájöttek, hogy a feladat sokkal összetettebb az eredetileg vártnál. Ma már kimondhatjuk, hogy a nyelvtechnológia egyik legnehezebb feladatáról van szó. A terület háttérbe szorult, hogy aztán a számítási kapacitás rohamos fejlődése nyomán a 80-as években újraéledjen. Mára

világossá vált, hogy a gépi fordítás nem tudja helyettesíteni az emberi fordítói munkát. Nem reális cél az emberi fordítás minőségének elérése, de a fordítás sebességének és a megértésben nyújtott segítségnek az arányát figyelembe véve megtérülő befektetés a gépi fordításba invesztálni. A gépi fordító eszközök legkézenfekvőbb haszna, hogy az idegen nyelvet nem ismerő, esetleg elolvasni sem tudó befogadó részére képes a szöveget nyersfordításban anyanyelvén prezentálni — másodperceken belül. A gépi fordításhoz szükséges a teljes nyelvtechnológiai feldolgozó lánc elemző és generáló oldalon is. A magyar BLARK (l. 4.2.) elkészülése e szempontból is rendkívül fontos.

A gépi fordítás szakmai diskurzusát napjainkban is meghatározza a 4.1. részben említett statisztikai, illetve szabályalapú rendszerek (látszólagos) ellentéte. A gazdag morfológiájú nyelvekre, így a magyarra is, nagyobb hatékonysággal működnek a szabályalapú fordítók, amelyeket ezért szükséges nagy erőkkel továbbfejleszteni. A közeljövő feladatai közé tartozik — így kimondottan a Platformon belül a korábbi évek során fejlesztett magyar-angol, angol-magyar gépi fordító rendszer esetében is — a meglévő fordítás minőségének javítása, illetve a magyart is tartalmazó nyelvpárok körének szélesítése. A minőségjavítás történhet akár a statisztikai és szabályalapú módszerek integratív alkalmazásával, akár fordítási minták szabályalapú rendszerekbe való beépítésével. Bármelyik módszert alkalmazzuk is, nagyméretű párhuzamos korpuszok építése (vö. 4.2.) elengedhetetlen a fejlesztéshez. Mivel a nyelvi többértelműség miatt a mai fordítórendszerek nagy többsége több fordítási alternatívát is generál egy forrásnyelvi mondathoz, a további feladatok között kell megemlíteni a szintaktikai és szemantikai egyértelműsítést, szemantikai információk használatával (l. 4.6. rész).

Fordítástámogatás. A fordítástámogatás területén a Platform jelenleg is világszínvonalú megoldásokat szállít a professzionális fordítók részére (l. Jelenkép *Gépi fordítás és fordítástámogatás*). A közeljövő feladata a gépi fordítás integrálása a fordítást támogató rendszerekbe, illetve a diktálórendszerek és a fordítástámogatás összekapcsolása.

Megértéstámogatás. A megértéstámogató eszközök használata azt a befogadót segíti, aki ismeretlen nyelven írt szöveget akar közvetlenül megérteni, belőle a lényeget kihámozni, de nem igényli a szöveg pontos lefordítását. Ide tartozik a terminológiakivonatolás, szólistakészítés és automatikus szótárazás, a szöveg szempontjából releváns alapvető nyelvtani információk kompakt formában való megjelenítése, a szöveg automatikus összegzése is. A megértéstámogató eszközök segítségével tájékozódni tudunk a szövegben, illetve az anyanyelvünktől idegen nyelvtani jelenségeket is kezelhetjük. Ez a megközelítés hasznos lehet idegen nyelvű menetrendek, sajtótermékek böngészésekor, vagy abban az esetben, ha fontos, hogy valóban az eredeti (nyelvű) dokumentum tartalmához férjünk hozzá. A megértéstámogató eszközök azokra a kisebb nyelvekre is létrehozhatók, melyekre a gépi fordítás megvalósítása nem kifizetődő.

Szótárak, számítógépes lexikográfia. Mindig szükség lesz a nyelvek változó szókincsét követő és bemutató újabb és újabb szótárakra. Fontos a már ma is folyó (automatikus/félautomatikus/hagyományos) szótárépítő munkálatok egységes szervezeti keretbe foglalása. A lexikográfiailag releváns információk szövegekből való kinyerését célzó nyelvtechnológiai algoritmusok kutatása közelebb visz a szótárkészítés automatizálásához. A következő évtizedben várható olyan nyelvfüggetlen korpuszalapú automatikus szótárépítő eljárások megjelenése, melyek segítségével *dinamikusan* készíthetünk szótárt az aktuális célra kialakított korpusz alapján, legyen az speciális szaknyelvi vagy akár idegen nyelvű korpusz. A gyors és rugalmas automatikus módszerek segítségével a szótárak naprakészebbé és teljesebbé válhatnak.

4.6. Az információkinyerés és -visszakeresés kutatási irányai

A nyelvi alapú tudástárak létrehozásához Magyarországon nemcsak a kulcsszavas keresés infrastruktúrájának javítására és a magyar nyelvű információkinyerés fejlesztésére, hanem a nyelvek közötti információkinyerés jelentős erősítésére is szükség van, mind a szöveges, mind a beszéd-, illetve multimédia-tartalmú adatbázisokban. Célunk részint az információkinyeréshez elengedhetetlen nyelvi modalitások (tagadás, spekuláció, időbeliség stb.) automatikus felismerése, másrészt az emberi kommunikációt átható érzelmi hozzáállás elemzése.

A ***hangzó anyagokból történő információkinyerés***, -visszakeresés első lépése a beszéd és nem beszéd részek szétválasztása, kategorizálása, majd a beszéd-szöveg átalakítás. Ez utóbbi feladat még egyetlen nyelvre sem megoldott feladat, ezért a „beszédbányászat" szűk keresztmetszete a gépi beszédfelismerés. Azonban, mint ahogy a szövegalapú keresésnél sem várhatunk 100%-os pontosságot, akár csak 50%-os szófelismerési pontosságú beszéd-szöveg átalakítás is már gyakorlatilag hasznosítható beszédinformáció-kinyerő rendszert adhat. Természetesen a használhatóságot a pontosabb beszéd-szöveg átalakítás nagyban segíti, ez irányban célzott kutatások szükségesek. A további lépések általában megegyeznek a szövegből történő információkivonatolás lépéseivel.

A ***szövegből történő információkinyerés*** egyes lépései során a szöveget alkotóelemeire bontjuk, majd a speciális jelentőséggel bíró elemeket lokalizáljuk és azonosítjuk. Az alkotóelemekre bontás különböző lépései (tokenizálás, mondatra bontás, morfológiai elemzés, szófaji egyértelműsítés) már tulajdonképpen megoldottnak tekinthetők a magyar nyelvre — annál nagyobb kihívást jelent a mélyebb mondaton belüli összefüggések és a mondatok közti összefüggések automatikus felismerése.

Mivel a nevek a szövegekben található tartalom lényeges és jól elkülöníthető tulajdonságokkal rendelkező elemei, az információkinyerés egyik legfontosabb lépése a *tulajdonnév-felismerés* (named entity recognition), amelynek célja a szövegben található tulajdonnevek felismerése és szemantikai kategorizációja (pl. személynév, földrajzi név, intézménynév stb.). A következő feladat a *referenciafeloldás* (reference resolution), amelynek során megállapítjuk, hogy a felismert nevek közül melyek jelölik ugyanazt az entitást, majd az ezen entitások közötti

szemantikai viszonyokat kell feltérképezni. Ezt követően fel kell ismerni a szövegben található eseményeket, ezek szemantikai osztályát, valamint azt, hogy a szöveg által meghatározott entitások milyen szerepet töltenek be az eseményben. Majd következik a szöveg idői szerkezetének feltárása, végül a *keretillesztés* (template filling), melynek során sztereotipikus mintákat készítünk a lefedendő területre, és az ezen mintákban található üres helyeket feltöltjük a szövegekből kinyert információkkal. Az információkinyerés ezen bonyolultabb lépéseinek megoldása még előttünk áll.

Azok a technológiák, amelyek a célzott *webbányászatot* segítik, mint — a fentiek mellett — a tartalom/téma szerinti osztályozás vagy a különféle megjelölési (markup) megoldások, különösen fontosak. A hagyományos információs tárak, a könyvtáraktól az adatbázisokig, csak annyira fognak túlélni, amennyire a web részeivé válnak. Ez az a széles sodrású folyamat, ami a *szemantikus web* létrejöttét kikerülhetetlenné teszi. A szöveg egyes részeinek megjelölése, elemekre bontása csupán eszköz a nagyobb cél: a szöveg megértése, a tudás kinyerése eléréséhez. A szemantikus web akár úgy is felfogható, mint egyszerűsített, ma még a szövegértésben az emberi képességektől messze elmaradó algoritmusok számára is érthető tartalom.

Az információkinyeréshez és -visszakereséshez elengedhetetlen az a feldolgozási lépés, amely a természetes nyelvi kifejezéseket megfelelő fogalmakhoz köti — például a 4.2. részben említett ontológiák megfelelő fogalmaihoz. Ez a lépés a **jelentésegyértelműsítés**, mely az egyik legnagyobb kihívás a nyelvtechnológia számára, mivel egy adott szó vagy kifejezés szövegkörnyezettől függően jelenthet mást és mást. Könnyebb kezelni az olyan eseteket, amikor az azonos alakú szavak más szófajúak (pl. *nyúl*, *vár*), nehezebben detektálhatók az egy szófajba tartozó azonos alakú szavak jelentései (pl. *egér*: állat vs. számítógép-tartozék), még nehezebb egy ige különböző jelentéseit, jelentésárnyalatait automatikusan felismerni. Fontossága ellenére a jelentésegyértelműsítés feladata még az angol nyelvre sem megoldott, vagyis nem létezik általánosan elfogadott, hatékony módszer, de az eredmények — nem utolsósorban az egyre jobb minőségű ontológiák megjelenésének köszönhetően — sokat javultak az elmúlt években. Tekintve e kutatási irány sokrétű felhasználhatóságát, várható, hogy a témában végzett kutatások az elkövetkezendő évek egyik legmeghatározóbb irányát fogják adni nemzetközi és hazai szinten egyaránt. A jelentésegyértelműsítés fejlődésével egyre nagyobb teret nyerhetnek azok az alkalmazások, amelyek nemcsak szavak, hanem egyre teljesebb szövegek számítógépes megértését tűzik ki célul.

4.7. Integratív kutatási irányok

Ma az egyik legnagyobb kihívás a teljesen eltérő tudományos hátterű kutatók és műhelyek már meglévő eredményeinek, folyamatban lévő kutatásainak összehangolása a közös célok érdekében. Különösen igaz ez a nyelv- és beszédtechnológiára, ahol a szűkebben vett technológusok is különböző: nyelvészi, mérnöki, informatikusi alapképzettségűek, ugyanakkor a nyelv- és beszédtechnológia ezer szállal kötődik a matematikához, a fizikához, a biológiához, az orvostudomá-

nyokhoz, a pszichológiához és ezek határterületeihez, a neurolingvisztikához, a pszichoakusztikához stb.

Beszédfordítás, automatikus tolmácsolás. A nyelv- és beszédtechnológia talán legjobban várt alkalmazása a beszédfordító gép. A feladat egyben az egyik legnagyobb technológiai kihívás is: önmagában a témafüggetlen gépi beszédfelismerés és a szövegfordítás is hatalmas kihívás, ezek kombinációja pedig hatványozott nehézséget jelent. A témakör szűkítésével viszont igenis lehetséges gyakorlatban is használható beszédfordítókat készíteni (lásd pl. a BBN és az IBM által fejlesztett arab-amerikai katonai célú alkalmazások, vagy a TC-STAR projektben az európai parlamenti beszédek online fordítása). Így reményteljes vállalkozás a magyar-angol, -német stb. nyelvpárokra is restriktív célú beszédfordító rendszereket készíteni. Arra különösen ügyelni kell, hogy a fordítás- és a beszédtechnológia nem lehet független, szoros kollaborációra van szükség. A fordító dolgát nagyban megkönnyítheti, ha egyrészt nemcsak a szószintű felismerési kimenetet kapja meg, hanem a morfémaszintűt is (lehetőleg ugyanazon morfológiai rendszerben, mint amivel maga a fordító dolgozik), valamint nemcsak a legvalószínűbb morfémasorozatot kapja meg a szövegfordító, hanem pl. az első 10 legvalószínűbbet.

Összefoglalva tehát a nyelv- és beszédtechnológiai kutató-fejlesztő műhelyek minden eddiginél szorosabb együttműködésére van szükség, ami kitartó munka árán nagy bizonyossággal meghozza gyümölcsét. Rövid és középtávon a szűkebb területekre specializált kutatás-fejlesztés lehet sikeres (egyes kórházi alkalmazások, idegenforgalmi megoldások merülnek fel például), a témakör általánosabb szintű megoldása csak ezek után, a távolabbi jövőben hozhat a felhasználók számára hasznosítható eredményeket.

Beszédterápiai és diagnosztikai kutatások. A beszédoktató rendszerek megoldási lehetőségei egyre nőnek. A kifejlesztésre kerülő multimodális eszközök (az auditív mellett a látási és érzékelési csatorna aktiválásával) számos beszédsérülés gyógyításának segédeszközei. A technológia magába foglalja a beszédfelismerés, -szintézis, -elemzés és vizuális megjelenítés legújabb kutatási eredményeit és eszközrendszerét. Ezek a rendszerek alkalmat adnak a hallássérültek beszédfejlesztésére, artikulációs hibák korrekciójára (sziszegők, magánhangzók), megkésett beszédfejlődés terápiájára, cochleáris implantátummal rendelkezők rehabilitációjára, fonológiai problémák javítására, idegen akcentus csökkentésére. A beszédhibás és hallássérült emberek beszédoktatásán kívül, vagy inkább azt háttérbe szorítva, egy új irányzat annak vizsgálata, hogyan lehetne az idegennyelv-oktatásban hatékonyan hasznosítani a számítógépes rendszereket (Computer Aided Language Learning, CALL).

Ide tartozik még a beszédalapú diagnosztika is, mely lehetővé teszi a hangképzési rendellenességek vizsgálatát és automatikus diagnosztizálását. Emellett a beszéd részletes vizsgálatával sok egyéb betegség is előre jelezhető (pl. az Alzheimer-kór).

Multimodális dialógusrendszerek. Az emberi kommunikáció tipikusan kétoldalú, és a gyakorlati problémák során viszonylag nagy arányban tipikus kérdésekre tipikus válaszok születnek (pl. ügyfélszolgálati rendszerek). Ezért gyakran merül fel, hogy a gépies emberi munkát emberies gépi munkával váltsuk ki, azaz automatizáljuk a válaszadást a tipikus kérdések kategorizálása után. Az ügyfélszolgálati munka sokszor megoldható gépies válasszal, ugyanakkor sokan emberi megnyilvánulásokat is elvárnának a géptől. Ezt érdemben megvalósítani jelenleg reménytelen vállalkozás. Számos esetben nem is lenne szükséges a gépet valódi emberi intelligenciával felvértezni, sokszor apróbb „emberi jellegű" megnyilvánulások, melyek a felhasználó viselkedéséhez adaptálhatók, jelentősen javíthatnak az adott szolgáltatás tetszési indexén. Ilyen lehet például, ha a gépi beszéd sebessége, esetleg stílusa alkalmazkodik a beszélőéhez — ehhez mind az input, mind az output tekintetében szükséges a jelenlegi technológiák továbbfejlesztése.

A teljes értékű emberi intelligencia ugyanakkor utópisztikus cél lenne, ezért a modern kutatások egy része egyfajta „állati intelligencia" alkalmazását tartja célravezetőnek a hatékony ember-gép kommunikációban. Különösen a multimodális felületek esetén van lehetőség ilyennek az alkalmazására, ahol a bemeneti oldalon vizuális gesztusfelismerés alapján akár egy rajzfilmfigura a beszédkapcsolat kiegészítőjeként metakommunikációval — pl. szemhunyorgatással — jelez vissza egyszerűen és hatékonyan. Indult már ilyen témában kutatás, de az ember-ember, ember-gép, ember-(házi)állat kommunikáció kimeríthetetlen területek, melyek kutatása újszerű, emberbarát gyakorlati megoldásokhoz vezethet.

Nyelvi tartalom megértése, beszédfelismerés, beszédelőállítás, fordítás. Jól ismert, hogy a megértett beszéd felismerése sokkal pontosabb, mint a nem tudatosult közlésé, valamint az átélt szövegtartalmat sokkal kifejezőbben és helyesebben tudjuk felolvasni, mint a szolgai módon felolvasott szöveget. Ez azt mutatja, hogy a kogníció, a nyelvi tartalom valódi megértése fontos szerepet játszik a beszédpercepcióban és -artikulációban is. Fokozottan igaz ez az emberi fordításra is. Tehát a beszéd-szöveg, szöveg-fogalom, szöveg-beszéd átalakítás nem különülnek el egymástól az emberi beszédértés folyamatában. Ugyanakkor a nyelv- és beszédtechnológiában e rendszerkomponensek csak névlegesen kapcsolódnak össze, valódi megértésről nem beszélhetünk. Így viszont a gépi beszédfelismerés, -szintézis és fordítás olyan távol marad az emberi műveletektől, amit nem biztos, hogy az egyéb technológiák javításával be lehet hozni. Stratégiailag fontos cél tehát bekapcsolódni az emberi beszédértés, a kogníció nemzetközi szinten előrehaladott kutatásaiba, az eredményeket alkalmazásra kész technológiába beépíteni, hogy az egyelőre még majdnem utópisztikus távlati cél, az emberéhez hasonló képességű gépi fordítás, beszédfelismerés és -szintézis előállhasson. Ez egyelőre kifejezetten hosszú távú kutatási feladat, azonban már ma látszik, hogy a világ erre halad, és kimaradni súlyos vétek lenne.

5. Alkalmazási területek

A nyelv- és beszédtechnológia nemzetgazdasági hasznosíthatósága a természetes nyelven történő kommunikáció alapvető fontossága miatt rendkívül sokrétű. Az alábbiakban olyan gyakorlati alkalmazási területeket emelünk ki, ahol akár rövid távon is sikeresen bevezethetők a nyelv- és beszédtechnológiai fejlesztések a gazdasági szféra, az állam- és közigazgatás, az egészségügy vagy az oktatás bizonyos területein. További jövőbeli gyakorlati alkalmazási lehetőségekről ld. a Jövőkép *Kiemelt alkalmazások* fejezetét.

5.1. A kutatás-fejlesztési eredmények gyakorlati felhasználása

Ipari alkalmazások. A vállalatok számára a **hatékony tudásbeszerzés**, **konkurencia-** és **trendanalízis**, a nyelvalapú **multimédia- és tudásmenedzsment** milliárdokban mérhető realizált hasznot jelenthet. Ezek az alkalmazások mind intenzíven építenek a nyelv- és beszédtechnológiai fejlesztésekre.

Ma még a **hatékony internetes keresés**hez tapasztalatra, időre és gyakran némi szakmai jártasságra van szükség. A jövőben a robbanásszerű mértékben növekvő webtartalom megköveteli a gyorsabb, pontosabb és laikusok számára is könnyen használható keresést, melyet a továbbfejlesztett információ-visszakereső és szemantikus technológiák tesznek lehetővé. Hasonlóak mondhatók el az **üzleti intelligenciát**, **döntéshozást támogató szoftverek** területén is.

A nyelvtechnológia lehetővé teszi az adott nyelven elérhető információk más nyelvekre való gyors és költséghatékony átültetését. Arra számítunk, hogy egy évtizeden belül elkövetkezik az az ideális állapot, amikor az interneten található idegen nyelvű honlapok böngészése nem fog problémát okozni: az **automatikus gépi fordítási** megoldások segítségével saját anyanyelvünkön, nagyjából érthető módon olvashatjuk a különböző tartalmakat, és mint felhasználók jelentős segítséget kapunk például az e-kereskedelemben.

A beszédtechnológia lehetővé teszi a bármikor és bárhonnan történő **telefonos ügyintézés** általános elterjedését, illetve segít minden olyan esetben, ahol a vizuális információ nem adható át hatékonyan.

A nagyméretű **multimédia-adatbázisok** tartalmi kereshetőségét a beszédfelismerési és információ-visszakeresési technológiák teszik lehetővé, így válnak ezek a multimédia-archívumok szélesebb körben és hatékonyabban hasznosíthatóvá.

Információkinyerő alkalmazásokat használhatnak a **sajtófigyelő** cégek, a webes szolgáltatásokat nyújtó kis- és középvállalkozások, amivel emberi munkaerőt, időt és pénzt takaríthatnak meg.

Az információkinyerés hatékonyságának növelése kedvező innovációs hatással jár, hiszen például a szabadalmak, tudományos közlemények automatikus feldolgozása felgyorsítja az információ áramlását az akadémiai, innovációs és ipari szféra között.

Állami, közigazgatási alkalmazások. A nyelv- és beszédtechnológia az állampolgárok számára alapjaiban változtathatja meg a mindennapi ügyintézést. Gépi

beszédfelismerésen alapuló telefontudakozók, **beszédalapú** *call centerek*, **komplex** (mobil)telefonos **ügyfélkiszolgáló rendszerek**, **természetes nyelvi interfészek** válthatják fel a humán operátorokat, és könnyíthetik meg a formanyomtatvány-kitöltésen alapuló jelenleg kezdetleges elektronikus ügyintézést.

A minisztériumok, az államigazgatási szervek, a nemzetvédelem és a rendőrség munkáját segíthetik a különböző **információkinyerő eszközök**, a nyelvtechnológia eredményeit felhasználó alkalmazások (pl. automatikus anonimizáló rendszerek vagy intelligens keresőeszközök).

A **természetes nyelvi** alapon történő hatékony **tudásszerzés** jelentős társadalmi hatással bírhat például a jogalkalmazás területén, hiszen a különböző jogszabály-gyűjteményekből hatékonyan visszakereshető információ nemcsak a szakemberek munkáját könnyíti meg, hanem a jogi információ előzetes feldolgozásával, kategorizálásával hozzáférhetőbbé, könnyebben értelmezhetővé teszi a jogszabályokat, és így közvetlenül is hozzájárulhat a jogbiztonság növekedéséhez Magyarországon.

A **gépi fordítástámogatás** segítségével radikálisan csökkenthetők az emberi tolmácsolás és fordítás által igényelt költségek, így például az Európai Parlament működési költségeinek tetemes része. Ez összeurópai érdek, vagyis a gépi fordítás és fordítástámogatás területén hosszú távú, folyamatos fejlesztésekre van szükség.

Egészségügyi alkalmazások. A nyelv- és beszédtechnológián alapuló **orvosdiagnosztikai eszközök** fejlesztése egyre szélesebb körben jellemző, ilyen segédeszközök használatára az orvostársadalomban kifejezett igény van.[3]

A nagy tömegű **orvosi információ nyelvtechnológiával támogatott feldolgozása** kiemelt fontosságú a diagnosztikában, a gyógyszerkutatásban, az információ rendszerezésében és kategorizálásában (pl. leletek automatikus osztályozása, géppel segített diagnózisok felállítása, orvosi utasítások ellenőrzése). **Terápiás** és **rehabilitációs** alkalmazások (egyedi beszélőre adaptált beszédszintézis, beszédterápiás, beszélni tanító szoftverek), a mindennapi munkavégzést támogató **segédeszközök** (pl. orvosi diktáló rendszerek), az egészségügyben használható hasonló alkalmazások, illetve az ezekhez szükséges alacsonyabb szintű feladatokat megoldó rendszerek kifejlesztése olyan célok, melyek megfelelő nyelvtechnológiai befektetéssel rövid időn belül megvalósíthatók.

A megváltozott munkaképességű személyek társadalmi integrációja. A nyelv- és beszédtechnológia hozzájárul a megváltozott munkaképességű személyek társadalmi integrációjához is. Az egyik legelemibb igény mind látás-, mind hallássérültek számára a vizuális, illetve auditív információ más médiumon keresztül való elérhetővé tétele. A beszédfelismerés és -szintézis technológiái már

[3] Ilyen például a hangképző szervi megbetegedések (pl. gégerák) kimutatására szolgáló beszédakusztikai-számítástechnikai diagnosztikai eljárás kidolgozása. Továbbá a beszédinformáció agyi feldolgozási zavarainak kimutatásában is egyre több beszéd- és nyelvi diagnosztikára és terápiára használatos eszköz jelenik meg, és válik egyre népszerűbbé.

lehetővé teszik azt, hogy az ember-gép kommunikációban olyan ember is részt vehessen, akinek a beszédészlelés vagy a gépelés nehézséget okoz, vagy lehetetlen. Folyamatos fejlesztés alatt állnak a **mindennapi életet jelentősen megkönnyítő alkalmazások** (pl. gépi felolvasó rendszer, hangos információs tábla, környezeti intelligenciával felruházott eszközök, automatikus beszédfeliratozó). Hosszútávú és komplex fejlesztést igényelne egy jelnyelvfelismerő és a jeleket írott vagy beszélt formába átalakító rendszer.

Az oktatás nyelv- és beszédtechnológiai támogatása. A nyelv- és beszédtechnológia fontos szerepet kaphat az **oktatás hatékonyságának növelésé**ben. A nyelvi erőforrások (l. 4.2.), szöveges adatbázisok új eszközzel gazdagítják a nyelvoktatás módszertanát, használatukkal valódi környezetbe ágyazott élőnyelvi példaanyag áll a nyelvtanuló rendelkezésére, segítségükkel a nyelvi jelenségek egzakt, empirikus módszerekkel tanulmányozhatók. Ide tartoznak az **intelligens nyelvoktató szoftverek**, minden a nyelvi tudatosságot elősegítő alkalmazás, a valamilyen nyelvi kihívással küzdő emberek segítését célzó rendszerek, tanító gépek (pl. a beszédkorrektor rendszerek). Egyre nagyobb az igény olyan mérési módszerek fejlesztésére is, amelyekkel a beszédterápia javító hatása objektív módon kiértékelhető, a fejlődési lépték összehasonlítható.

5.2. Alkalmazási példák

A következőkben szemléltető jelleggel a 4. fejezet egyes speciális kutatási irányaihoz társítunk egy-egy gyakorlati alkalmazási példát.

- Robusztus beszédfelismerési technikák: autóban és tömegközlekedési eszközökön használható navigációs rendszerek vezérlése.
- Spontán társalgási beszéd felismerése: ügyfélszolgálatok minőségbiztosítása.
- Nagyszótáras folyamatos beszédfelismerési technikák gazdag morfológiájú nyelvekre: híradók automatikus feliratozása.
- Nyelvfüggetlen beszédfelismerő módszerek kialakítása: horvát, román stb. nyelvű multimédia-menedzsment.
- Hangkarakter- és kiejtésistílus-transzformáció: emberközeli automatikus regényfelolvasás vakoknak, gyerekeknek.
- Érzelem kifejezése gépi beszéddel és spontán beszédstílus megvalósítása: barátságos, emberi érzetű gépi ügyfélszolgálat.
- Többnyelvű szintézist támogató keretrendszer fejlesztése: turisztikai információs rendszer telefonon.
- Fordítástámogatás: fordítóiroda munkáját megkönnyítő megoldások.
- Megértéstámogatás: az e-kereskedelem nyelvi támogatása.
- Szövegből történő információkinyerés: webalapú piacelemzés, konkurenciaanalízis.
- Hangzó anyagokból történő információkinyerés: automata telefonos ügyfélszolgálat.
- Információ-visszakeresés: webes keresés, mélyebb tartalmi összefüggések kinyerése.

– Beszédfordítás: kórházi, biztosítási sürgősségi esetekre szabott alkalmazások.
– Beszédterápiai és -diagnosztikai kutatások: logopédiai tanítóeszközök, gégerák-diagnosztika.
– Multimodális dialógusrendszerek: navigációs és jegyautomata tömegközlekedésben.

Ahogy láthattuk, a nyelv- és beszédtechnológia jelen és főleg jövőbeli alkalmazási lehetőségei igen széleskörűek, közvetlen vagy közvetett használatuk egészen bizonyosan beépül mindennapi életünkbe. A Platform véleménye szerint nemcsak szakmai szempontból vonzó a kihívás, de a nemzetgazdaság szempontjából is kedvező lehet olyan technológiába fektetni, melynek potenciális napi felhasználója gyakorlatilag a teljes lakosság, és amely ilyen átfogó mértékben fokozza a nemzetgazdaság versenyképességét.

Szerzői index, névmutató