X. Magyar Számítógépes Nyelvészeti Konferencia

MSZNY 2014

Szerkesztette:

Tanács Attila
Varga Viktor
Vincze Veronika

Szeged, 2014. január 16-17.
http://rgai.inf.u-szeged.hu/mszny2014

ISBN: 978-963-306-246-3

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

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

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

Szeged, 2014. január

Előszó

Idén tizedik, jubileumi alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát 2014. január 16-17-én. A konferencia fő célkitűzése az elmúlt évtizedben mit sem változott: a rendezvény fő profilja a nyelv- és beszédtechnológia területén végzett legújabb, illetve folyamatban levő kutatások eredményeinek ismertetése és megvitatása, mindemellett lehetőség nyílik különféle hallgatói projektek, illetve ipari alkalmazások bemutatására is.

Nagy örömömre szolgál, hogy a hagyományoknak megfelelően a konferencia nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferenciafelhívásra szép számban beérkezett tudományos előadások közül a programbizottság 43-at fogadott el az idei évben, így 26 előadás, 10 poszter-, illetve 7 laptopos bemutató gazdagítja a konferencia programját. A programban a magyar számítógépes nyelvészet rendkívül széles skálájáról találhatunk előadásokat a számítógépes morfológia és szintaxis területétől kezdve az információkinyerésen át a klinikai szövegek számítógépes feldolgozásáig.

Nagy örömet jelent számomra az is, hogy Benczúr András, az MTA Számítástechnikai és Automatizálási Kutatóintézetének Adatbányászat és Keresés Csoportjának laborvezetője, elfogadta meghívásunkat, és *Egy Virtuális Web Obszervatórium fejlesztésének tapasztalatai* című plenáris előadása is a konferenciaprogram részét képezi.

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

Ezúton szeretném megköszönni a Neumann János Számítógép-tudományi Társaságnak szíves anyagi támogatásukat.

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

Csirik János, a rendezőbizottság elnöke

Szeged, 2014. január

Tartalomjegyzék

I. Beszédtechnológia

II. Morfológia, szintaxis

III. Szemantika, ontológia

IV. Pszichológia

V. Orvosi NLP

VI. Információkinyerés és -visszakeresés

VII. Poszterek

VIII. Laptopos bemutatók

IX. Angol nyelvű absztraktok

I. Beszédtechnológia

Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben*

Grósz Tamás[1], Kovács György[2], Tóth László[2]

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

Kivonat 2006-os megjelenésük óta egyre nagyobb népszerűségnek örvendenek az akusztikus modellezésben az ún. mély neuronhálók. A hagyományos neuronhálókkal ellentétben a mély hálók sok rejtett réteget tartalmaznak, emiatt a hagyományos módszerekkel tanítva őket nem lehet igazán jó eredményeket elérni. Cikkünkben röviden bemutatunk négy új tanítási módszert a mély neuronhálókhoz, majd a mély neuronhálókra épülő akusztikus modelleket beszédfelismerési kísérletekben értékeljük ki. A különböző módszerekkel elért eredményeket összevetjük a korábban publikált eredményeinkkel.

Kulcsszavak: mély neuronhálók, akusztikus modellezés, beszédfelismerés

1. Bevezetés

A neuronhálós beszédfelismerési technika a reneszánszát éli, köszönhetően a mély neuronhálók feltalálásának. Tavalyi cikkünkben [1] mi is bemutattuk a módszer alapötletét, és az első mély neuronhálós felismerési eredményeinket magyar nyelvű adatbázisokon. A technológia iránti érdeklődés azóta sem csökkent, példának okáért az MIT „Tech Review's" listája a mély neuronhálókat beválogatta a 2013-as év 10 legfontosabb technológiai áttörést jelentő módszere közé. Mindeközben sorra jelennek meg az újfajta mély hálózati struktúrákat vagy tanítási módszereket publikáló cikkek. Jelen anyagunkban néhány olyan új ötletet mutatunk be, amelyekkel az eredeti tanítási algoritmus eredményei még tovább javíthatók, majd a módszereket magyar nyelvű beszédfelismerési adatbázisokon értékeljük ki.

A mély neuronhálók hatékony betanításához az eredeti szerzők az ún. DBN előtanítási módszert javasolták [2], ami egy elég komplex és műveletigényes algoritmus. Egy jóval egyszerűbb alternatívaként vetették fel nemrég az ún. discriminative pre-training („diszkriminatív előtanítás") algoritmust [3]. Ezen módszer

* Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

esetén az előtanítás felügyelt módon történik: kezdetben egy hagyományos (egy rejtett réteges) hálóból indulunk ki, néhány iteráción keresztül tanítjuk, ezután egy új rejtett réteget illesztünk be a kimeneti réteg alá és a kimeneti réteget újrainicializáljuk. Az így kapott neuronhálót újra tanítjuk néhány iteráción keresztül, az új rétegek hozzáadását pedig addig ismételjük, amíg a rejtett rétegek száma el nem éri a kívánt mennyiséget. A módszer előnye, hogy a tanítás során – mindkét fázisban – csak a backpropagation algoritmust kell használunk.

Egy másik mostanában javasolt, előtanítást nem igénylő módszer az ún. rectified („egyenirányított") neuronok használata. Ezek nevüket onnan kapták, hogy egyenletükben a szokásos szigmoid aktivációs függvény le van cserélve egy olyan komponensre, amelynek működése egy egyenirányító áramkörre hasonlít (matematikailag a $max(0, x)$ függvényt valósítja meg). A rectifier neuronokra épülő mély neuronhálók használatát eredetileg képfeldolgozásban vetették fel, csoportunk az elsők között próbálta ki őket beszédfelismerésben [4]. Eredményeink egybevágnak a más kutatók által velünk párhuzamosan publikált eredményekkel: úgy tűnik, hogy az egyenirányított mély neuronhálók hasonló vagy kicsit jobb felismerési pontosságot tudnak elérni, mint hagyományos társaik, viszont a betanításuk jóval egyszerűbb és gyorsabb [5,6].

Egy harmadik nemrégen feltalált módszer a neuronháló backpropagation tanítási algoritmusát módosítja. Az ún. dropout („kiejtéses") tanulás lényege, hogy a neuronháló tanítása során minden egyes tanítópélda bevitelekor véletlenszerűen kinullázzuk („kiejtjük") a hálót alkotó neuronok kimenetének valahány (általában 10-50) százalékát [7]. Ennek az a hatása, hogy az azonos rétegbe eső neuronok kevésbé tudnak egymásra hagyatkozni, így a probléma önálló megoldására vannak kényszerítve. Ennek köszönhetően lényegesen csökken a túltanulás veszélye. A módszert eredetileg javasló cikkben kiugró, 10-20 százaléknyi relatív hibacsökkenéseket értek el képi alakfelismerési és beszédfelismerési feladatokon.

A javasolt módszerek hatékonyságát először az angol nyelvű TIMIT adatbázison szemléltetjük, mivel ezen számtalan összehasonlító eredmény áll rendelkezésre. Ezután két magyar nyelvű adatbázissal kísérletezünk. Az egyik egy híradós adatbázis, amelynek méretét tavaly óta jelentősen sikerült megnövelnünk. A másik pedig a „Szindbád történetei" című hangoskönyv hangzóanyaga, amelyen szintén publikáltunk már eredményeket korábban.

2. Mély neuronhálók

A hagyományos neuronhálók és a mély hálók között az alapvető különbség, hogy utóbbiak több (általában 3-nál több) rejtett réteggel rendelkeznek. Ezen mély struktúrájú neuronhálók használatát igazolják a legújabb matematikai érvek és empirikus kísérletek, melyek szerint adott neuronszám mellett a több rejtett réteg hatékonyabb reprezentációt tesz lehetővé. Ez indokolja tehát a sok, relatíve kisebb rejtett réteg alkalmazását egyetlen, rengeteg neuront tartalmazó réteg helyett.

A sok rejtett réteges mély neuronhálók tanítása során több olyan probléma is fellép, amelyek a hagyományos egy rejtett réteges hálók esetén nem vagy alig

megfigyelhetőek, és ezen problémák miatt a betanításuk rendkívül nehéz. A hagyományos neuronhálók tanítására általában az ún. backpropagation algoritmust szokás használni, ami tulajdonképpen a legegyszerűbb, gradiensalapú optimalizálási algoritmus neuronhálókhoz igazított változata. Több rejtett réteg esetén azonban ez az algoritmus nem hatékony. Ennek egyik oka, hogy egyre mélyebbre hatolva a gradiensek egyre kisebbek, egyre inkább „eltűnnek" (ún. „vanishing gradient" effektus), ezért az alsóbb rétegek nem fognak kellőképp tanulni [8]. Egy másik ok az ún. „explaining away" hatás, amely megnehezíti annak megtanulását, hogy melyik rejtett neuronnak mely jelenségekre kellene reagálnia [2]. Ezen problémák kiküszöbölésére találták ki az alább bemutatásra kerülő módszereket.

2.1. DBN előtanítás

A mély neuronhálók legelső tanítási módszerét 2006-ban publikálták [2], lényegében ez volt az a módszer, amely elindította a mély neuronhálók kutatását. A módszer lényege, hogy a tanítás két lépésben történik: egy felügyelet nélküli előtanítást egy felügyelt finomhangolási lépés követ. A felügyelt tanításhoz használhatjuk a backpropagation algoritmust, az előtanításhoz azonban egy új módszer szükséges: a DBN előtanítás.

A DBN előtanítással egy ún. „mély belief" hálót (Deep Belief Network, DBN) tudunk tanítani, amely rétegei korlátos Boltzmann-gépek (RBM). A korlátos Boltzmann-gépek a hagyományosaktól annyiban térnek el, hogy a neuronjaik egy páros gráfot kell hogy formázzanak. A két réteg közül a látható rétegen keresztül adhatjuk meg a bemenetet, a rejtett réteg feladata pedig az, hogy az inputnak egy jó reprezentációját tanulja meg.

Az RBM-ek tanításához a kontrasztív divergencia algoritmust (CD) használhatjuk, amely egy energiafüggvény alapú módszer. Egy RBM a következő energiát rendeli egy látható (v) és a rejtett réteg (h) állapotvektor-konfigurációhoz:

$$E(v,h;\Theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_ih_j - \sum_{i=1}^{V} b_iv_i - \sum_{j=1}^{H} a_jh_j. \tag{1}$$

Az egy lépéses kontrasztív divergencia algoritmus (CD-1) esetén a következő update szabályt alkalmazzuk a látható-rejtett súlyokra:

$$\Delta w_{ij} \propto \langle v_ih_j\rangle_{input} - \langle v_ih_j\rangle_1, \tag{2}$$

ahol $\langle . \rangle_1$ a látható és a rejtett rétegek Gibbs-mintavételezővel egy lépésben történő mintavételezése utáni kovarianciája.

Habár az RBM energiafüggvénye rendkívül jól működik bináris neuronok esetén, beszédfelismerésben azonban valós bemeneteink vannak, ennek kezelésére szükséges az energiafüggvény (1) módosítása. A valós bemenetekkel rendelkező RBM-et Gaussian-Bernoulli korlátos Boltzmann-gépnek (GRBM) nevezzük, energiafüggvénye:

$$E(v,h|\Theta) = \sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2} - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_ih_j - \sum_{j=1}^{H} a_jh_j. \tag{3}$$

Ezen új energiafüggvény esetén a CD-1 algoritmusban csupán a Gibbs-mintavételezés módját kell módosítani, a súlyok frissítése pedig továbbra is (2) szerint történik.

A DBN előtanítás során a hálót rétegpáronként tanítjuk. Az első lépésben az inputot és a legelső rejtett réteget egy GRBM-nek tekintve a CD-1 algoritmussal tanítjuk. A továbbiakban a következő RBM-nek a látható rétege az előzőleg tanított RBM rejtett rétege lesz, az új rejtett rétege pedig a következő rejtett réteg a hálóban. Az így inicializált hálók felügyelt tanulással finomhangolva lényegesen jobb eredményeket tudnak elérni, mint az előtanítás nélkül tanítottak.

2.2. Diszkriminatív előtanítás

A diszkriminatív előtanítást (Discriminative pre-training, DPT) a DBN előtanítás alternatívájaként javasolták [3]. Ahogy az elnevezésből sejthető, ez a módszer is két fázisból áll, a különbség, hogy az előtanítást is felügyelt tanítással, a backpropagation algoritmussal valósítjuk meg. Az algoritmus kezdetben egy hagyományos egy rejtett réteges neuronhálóból indul ki, amit néhány iteráción keresztül tanítunk. A következő lépésben egy új rejtett réteget illesztünk be a kimenet és a legfelső rejtett réteg közé, a kimeneti réteg súlyait újrainicializáljuk, majd az egész hálót tanítjuk néhány iteráción keresztül. Mindezt addig ismételjük, amíg a rejtett rétegek száma a kívánt mennyiséget el nem éri. A módszer előnye, hogy nem igényel külön tanítási algoritmust.

A tanítás során felmerül egy fontos kérdés, mégpedig, hogy az előtanítás során meddig tanítunk. Az eredeti cikk [3] szerint az eredmények romlanak, ha minden előtanítási lépésben a teljes konvergenciáig tanítunk. Javasolt csak néhány iterációnyit tanítani - a szerzők 1 iterációnyit javasolnak - mi a 4 iterációnyi előtanítást találtuk a legeredményesebbnek, azonban megemlítjük, hogy ha a tanító adatbázis mérete megnő, akkor az 1 iterációnyi előtanítás is elegendőnek tűnik.

2.3. Rectifier neuronhálók

Tekintve, hogy az előző két előtanításos módszernek rendkívül nagy az időigénye, sok kutató olyan módszereket próbált kidolgozni, amelyek nem igényelnek előtanítást. Az egyik ilyen javaslat nem a tanítóalgoritmust módosítja, hanem a hálót felépítő neuronokat. Az ún. rectified („egyenirányított”) neuronok nevüket onnan kapták, hogy a szokásos szigmoid aktivációs függvény le van cserélve egy olyan komponensre, amelynek működése egy egyenirányító áramkörre hasonlít (matematikailag a $max(0, x)$ függvényt valósítja meg). A rectifier neuronokra épülő mély neuronhálók használatát eredetileg képfeldolgozásban javasolták, csoportunk az elsők között próbálta ki őket beszédfelismerésben [4]. Eredményeink egybevágnak a más kutatók által velünk párhuzamosan publikált eredményekkel [5,6]: úgy tűnik, hogy az egyenirányított mély neuronhálók előtanítás nélkül is hasonló vagy kicsit jobb felismerési pontosságot tudnak elérni, mint hagyományos társaik előtanítással.

A rectifier függvény két alapvető dologban tér el a szigmoid függvénytől: az első, hogy az aktivációs érték növekedésével a neuronok nem „telítődnek”, ennek köszönhetően nem jelentkezik az eltűnő gradiens effektus. A rectifier neuronok esetén emiatt egy másik probléma jelentkezhet, mégpedig hogy a gradiens értékek „felrobbannak” (ún. „exploding gradient” effektus), azaz egyre nagyobb értékeket vesznek fel [8]. A probléma kiküszöbölése céljából a neuronok súlyait a tanítás során időről időre normalizálni szokták, mi a kettes norma szerint normalizáltunk. A másik fontos különbség, hogy negatív aktivációs értékekre 0 lesz a neuronok kimenete, aminek következtében a rejtett rétegeken belül csak a neuronoknak egy része lesz aktív adott input esetén. Ez utóbbi tulajdonságról azt is gondolhatnánk, hogy megnehezíti a tanulást, hiszen megakadályozza a gradiens visszaterjesztését, azonban a kísérleti eredmények ezt nem támasztják alá. A kísérletek azt igazolták, hogy az inaktív neuronok nem okoznak problémát mindaddig, amíg a gradiens valamilyen úton visszaterjeszthető.

Összefoglalva: a rectifier hálók nagy előnye, hogy nem igényelnek előtanítást, és a hagyományos backpropagation algoritmussal gyorsan taníthatók.

2.4. Dropout módszer

Az ún. dropout („kiejtéses”) tanulás lényege, hogy a neuronháló tanítása során minden egyes tanítópélda bevitelekor véletlenszerűen kinullázzuk („kiejtjük”) a hálót alkotó neuronok kimenetének valahány (általában 10-50) százalékát [7]. Ennek az a hatása, hogy az azonos rétegbe eső neuronok kevésbé tudnak egymásra hagyatkozni, így a probléma önálló megoldására vannak kényszerítve. Ennek köszönhetően lényegesen csökken a túltanulás veszélye. A módszert eredetileg javasló cikkben kiugró, 10-20 százaléknyi relatív hibacsökkenéseket értek el képi alakfelismerési és beszédfelismerési feladatokon.

A dropout technika előnye, hogy roppant egyszerűen implementálható, és elvileg minden esetben kombinálható a backpropagation algoritmussal. Az eredeti cikkben előtanított szigmoid hálók finomhangolása során alkalmazták, de azóta többen megmutatták, hogy rectifier neuronhálók tanításával kombinálva is remekül működnek [5]. További javulás érhető el az eredményekben, ha a tanítás során minden inputvektort többször (2-3-szor) is felhasználunk egy iteráción belül, különböző neuronkieséssel. Ugyan ez némileg javít az eredményeken, de az algoritmus futásidejét sokszorosára növeli, ezért mi csak egyszer használtunk fel minden inputvektort egy tanítási iterációban.

3. Kísérleti eredmények

A továbbiakban kísérleti úton vizsgáljuk meg, hogy a mély neuronhálók különböző tanítási módszerekkel milyen pontosságú beszédfelismerést tesznek lehetővé. Az akusztikus modellek készítése az ún. hibrid HMM/ANN sémát követi [9], azaz a neuronhálók feladata az akusztikus vektorok alapján megbecsülni a rejtett Markov-modell állapotainak valószínűségét, majd ezek alapján a teljes

megfigyeléssorozathoz a rejtett Markov-modell a megszokott módon rendel valószínűségeket. Mivel a neuronhálóknak állapotvalószínűségeket kell visszaadniuk, ezért minden esetben első lépésben egy rejtett Markov-modellt tanítottunk be a HTK programcsomag használatával [10], majd ezt kényszerített illesztés üzemmódban futtatva kaptunk állapotcímkéket minden egyes spektrális vektorhoz. Ezeket a címkéket kellett a neuronhálónak megtanulnia, amihez inputként az aktuális akusztikus megfigyelést, plusz annak 7-7 szomszédját kapta meg.

A modellek kiértékelését háromféle adatbázison végeztük el. Mindhárom esetben azonos volt az előfeldolgozás: e célra a jól bevált mel-kepsztrális együtthatókat (MFCC) használtuk, egész pontosan 13 együtthatót (a nulladikat is beleértve) és az első-második deriváltjaikat. A híradós adatbázis esetében szószintű nyelvi modellt is használtunk, a többi adatbázis esetén pusztán egy beszédhang bigram támogatta a beszédhang szintű felismerést.

Minden módszer esetében 128-as batch-eken tanítottunk, a momentumot 0.9-re állítottuk és backpropagation algoritmus esetén a korai leállást használtuk, a betanított mély hálók minden rejtett rétege 1024 neuronból állt.

A DBN előtanítás esetén a paraméterezés annyiban változott a tavalyi cikkünkben közölthöz képest, hogy lényegesen kevesebb epochon keresztül futtattuk a kontrasztív divergencia algoritmust az egyes RBM-ekre: 5 epoch a GRBM esetén és 3 a többi esetén a tavalyi 50-20 helyett. Tapasztalataink szerint ez volt az az iterációszám, amely során a rekonstrukciós hiba lényegesen csökkent, az ezt követő epochokban a súlyok is már csak minimálisan változtak. Az epochszám jelentős csökkentésével a tanításhoz szükséges idő is számottevően csökkent.

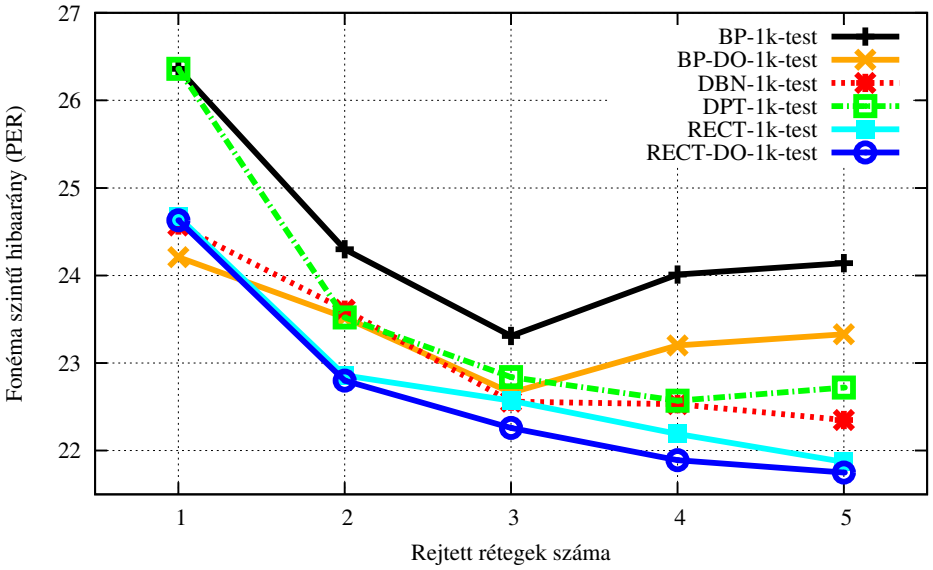
A diszkriminatív előtanítás esetén minden új rejtett réteg hozzáadása után 4 iteráción keresztül előtanítottunk 0.01-es fix tanulási rátával.

A rectifier neuronhálók estében a tanulási ráta 0.001 volt, illetve minden iteráció végén a súlyokat normalizáltuk, hogy az egy neuronhoz tartozó súlyok 2-es normája 1 legyen.

A dropout módszer esetén szigmoid hálókra a 10%-os neuronkiesési valószínűséget találtuk a legjobbnak, rectifier hálók estén pedig a 20%-ot. A tanítási iterációk végén a [7]-ben javasolt módon a súlyokat csökkentjük 10, illetve 20%-kal (a neuronkiesési valószínűséggel), hogy kompenzáljuk az a tényt, hogy tesztelés során a neuronok nem „esnek ki” véletlenszerűen.

3.1. TIMIT

A TIMIT adatbázis a legismertebb angol nyelvű beszédadatbázis [11]. Habár mai szemmel nézve már egyértelműen kicsinek számít, a nagy előnye, hogy rengeteg eredményt közöltek rajta, továbbá a mérete miatt viszonylag gyorsan lehet kísérletezni vele, ezért továbbra is népszerű, főleg ha újszerű modellek első kiértékeléséről van szó. Esetünkben azért esett rá a választás, mert több mély neuronhálós módszer eredményeit is a TIMIT-en közölték, így kézenfekvőnek tűnt a használata az implementációnk helyességének igazolására. A TIMIT adatbázis felosztására és címkézésére a tavalyi cikkünkben [1] ismertetett (és amúgy sztenderdnek számító) módszert használtuk. A továbbiakban csak monofón eredményeket közlünk.



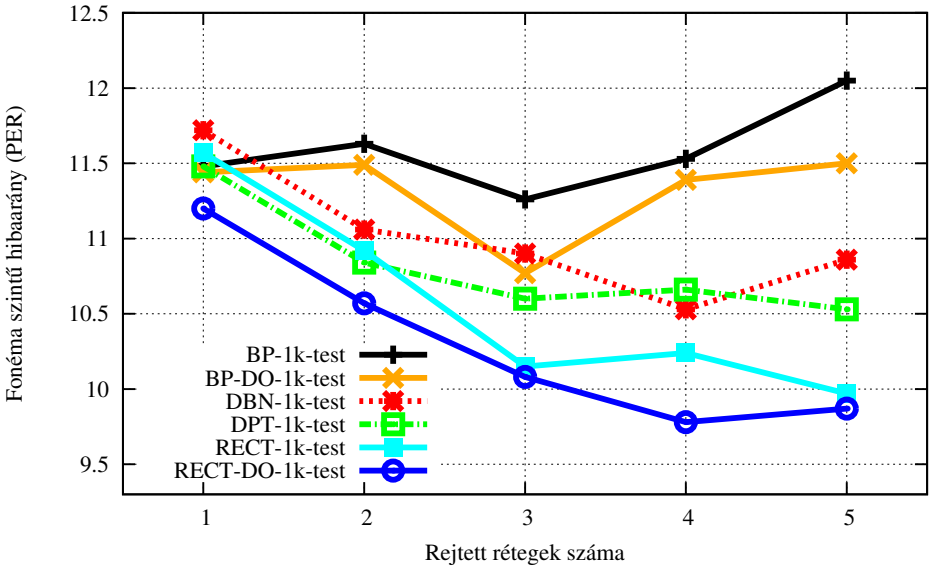
1. ábra. A különböző módszerek eredményei[4]a TIMIT core teszt halmazon a rejtett rétegek számának függvényében

A TIMIT adatbázison elért beszédhang szintű eredményeket láthatjuk a 1. ábrán. Jól látható, hogy a hagyományos backpropagation tanítóalgoritmusnál mindegyik ismertetett módszer jobban teljesített, ezen felül az is megfigyelhető, hogy a két előtanításos módszer nagyjából azonos eredményeket ért el. A legjobb eredményeket a rectifier hálókkal tudtuk kihozni: 21.75%, ami nagyjából 3%-os relatív javulás az előtanításos módszerekhez képest, illetve 7%-os relatív javulás a legjobb hagyományos (előtanítás nélküli) módszerhez képest. A korábbi cikkünkben közölt legjobb monofón eredményünkhöz (22.8%) képest a legjobb módszerrel több mint 1%-os javulást sikerült elérnünk, azonos módszerrel pedig 22.35%-ot, ami igazolja, hogy célszerű kevesebb előtanítást alkalmazni.

Az 1. ábrán megfigyelhető a dropout módszer hatékonysága is: míg szigmoid hálók esetén átlagosan 1%-os javulást hozott, ami 4%-os relatív javulásnak felel meg, addig a rectifier hálók esetén lényegesen kisebb a javulás. Ez utóbbinak az oka abban keresendő, hogy megfigyeléseink szerint a rectifier hálók neuronjainak átlagosan 70%-a inaktív tanítás során, ezt a dropout módszerrel kb. 75%-ra tudtuk növelni, ami nem hozott jelentős javulást az eredményekben.

Megvizsgáltuk továbbá, hogy a legjobban teljesítő mély neuronhálónkkal megegyező paraméterszámú hagyományos, egy rejtett réteges háló milyen eredményeket tud elérni. Az így kapott 23.5% lényegesen rosszabb mint a mély struk-

[4] Jelmagyarázat: **BP**: backpropagation, **BP-DO**: backpropagation+dropout, **DBN**: DBN előtanítás, **DPT**: diszkriminatív előtanítás, **RECT**: rectifier háló, **RECT-DO**: rectifier háló+dropout



2. ábra. A különböző módszerek eredményei a hangoskönyv adatbázis teszthalmazán a rejtett rétegek számának függvényében

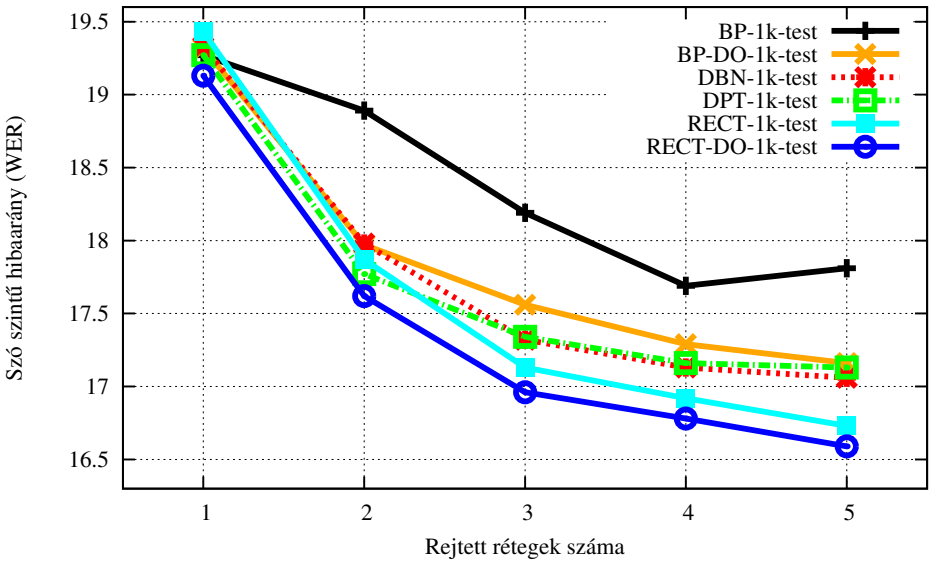
túrával elérhető eredmények, ami igazolja, hogy célszerű azonos paraméterszám esetén a mély struktúrájú hálót választani.

3.2. Hangoskönyv

A hangoskönyv adatbázisunk megegyezik a tavaly használttal. A 2. ábrán a különböző rétegszámmal elért eredményeket láthatjuk.

Megfigyelhető, hogy a különböző tanítási módszerek eredményei már jobban eltérnek, mint a korábbi adatbázison, viszont továbbra is megállapíthatjuk, hogy ha 2 vagy annál több rejtett réteget használunk, akkor a hagyományos módszer mindig a legrosszabb. A TIMIT-en elért eredményekhez hasonlóan itt is a rectifier hálók teljesítettek a legjobban, a legjobb eredményt (9.78%-ot) 4 rejtett réteggel dropout módszerrel tanítva értük el, ez több mint 13%-os hibacsökkenést jelent.

Az adatbázis sajátosságai miatt az figyelhető meg, hogy a hagyományosan tanított hálók esetén a rétegszám növelésével nem tudunk jelentős javulást elérni. A dropout módszer szigmoid hálók esetén 3 rejtett réteggel teljesített a legjobban – nagyjából 0.5%-kal jobb eredményt ért el –, rectifier háló esetén pedig 4 rejtett réteg esetén javított jelentősebben. A hagyományosan (azaz csak backpropagation algoritmussal) tanított, illetve a backpropagation+dropout módszerrel tanított szigmoid hálók kivételével mindegyik módszer esetén 4 vagy 5 rejtett réteggel értük el a legjobb eredményt. A tavalyi legjobb monofón eredményhez (10.62%) képest idén jelentős javulást tudtunk elérni (9.78%).



3. ábra. A különböző módszerek eredményei a híradós adatbázis teszthalmazán a rejtett rétegek számának függvényében

3.3. Híradós adatbázis

A híradós adatbázis, amely méretét tavaly óta sikerült jelentősen megnövelnünk, nagyjából 28 órányi hanganyagot tartalmaz. Az adatbázis felosztása: 22 órányi anyag a betanítási rész, 2 órányi a fejlesztési halmaz és a maradék 4 órányi hanganyag pedig a tesztelő blokk.

A híradós adatbázison szószintű felismerést tudtunk végezni, az ehhez szükséges nyelvi modellt az origo (www.origo.hu) hírportál szövegei alapján készítettük. Az így előálló korpusz nagyjából 50 millió szavas, mivel a magyar nyelv agglutináló (toldalékoló) nyelv. A korpusz lecsökkentése érdekében csak azokat a szavakat használtuk, amelyek legalább kétszer előfordultak a híranyagban, így 486982 szó maradt. A szavak kiejtését a Magyar Kiejtési Szótárból [12] vettük. A trigram nyelvi modellünket a HTK [10] nyelvi modellező eszközei segítségével hoztuk létre.

Ezen adatbázis esetén környezetfüggő (trifón) modelleket használtunk, ennek eredményeképp az adatbázis mérete miatt 2348 állapot adódott, azaz ennyi osztályon tanítottuk a neuronhálókat.

A 3. ábrán láthatóak az elért szószintű eredmények különböző rejtett rétegszám mellett. Ezen adatbázis esetén is elmondható, hogy a hagyományos módszer adja a legrosszabb eredményt, továbbá az is megfigyelhető, hogy a tanító adatbázis megnövekedése miatt a különböző tanítási módszerek eredményei jóval kevésbé térnek el. Továbbra is a rectifier hálók adják a legjobb eredményt (16.6%), ez a hagyományos módszerrel elérhető legjobb eredményhez (17.7%) képest 6%-os relatív hibacsökkenés.

1. táblázat. Az 5 rejtett réteges háló különböző módszerekkel történő tanításához szükséges idők

Módszer	Előtanítási idő	Finomhangolási idő
Hagyományos	0 óra	4.5 óra
Dropout	0 óra	5.5 óra
DBN előtanítás	1 óra	4 óra
Diszkriminatív előtanítás	2.5 óra	3 óra
Rectifier háló	0 óra	4 óra
Rectifier háló + Dropout	0 óra	4.5 óra

A híradós adatbázishoz közölt korábbi legjobb eredményünkhöz (16.9%) [13] képest is sikerült javítanunk, pedig a rejtett rétegek neuronszáma 2048-ról 1024-re csökkent.

Végül megvizsgáltuk az egyes módszerek időigényét: a 1. táblázatban az 5 rejtett réteges mély hálók különböző módszerekkel történő betanításához szükséges időket láthatjuk a híradós adatbázisra, GeForce GTX 560 Ti grafikus kártyát használva. Megállapítható, hogy a rectifier hálók nem csak jobb eredményeket adnak, de a betanításukhoz is kevesebb idő szükséges, mint a többi módszer esetén.

4. Konklúzió

Cikkünkben bemutattuk a mély neuronhálókra épülő akusztikus modelleket, illetve a betanításukhoz legújabban javasolt algoritmusokat. A kísérleti eredmények egyértelműen igazolják, hogy az új algoritmusok jobb eredményeket tudnak adni, miközben egyszerűbbek és/vagy kisebb időigényűek, mint az eredeti DBN előtanításra alapuló megoldás. Az eredményeket és a tanítási időket figyelembe véve megállapíthatjuk, hogy a legjobb módszer – az itt ismertetettek közül – a rectifier hálók dropout módszerrel történő tanítása.

Hivatkozások

1. Grósz T., Tóth, L.: Mély neuronhálók az akusztikus modellezésben. In: Proc. MSZNY. (2013) 3–12
2. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
3. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. ASRU. (2011) 24–29
4. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proc. ICASSP. (2013) 6985–6989
5. Dahl, G.E., Sainath, T.N., Hinton, G.: Improving deep neural networks for lvcsr using rectified linear units and dropout. In: Proc. ICASSP. (2013) 8609–8613

6. Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G.: On rectified linear units for speech processing. In: Proc. ICASSP. (2013) 3517–3521
7. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. In: CoRR. Volume abs/1207.0580. (2012)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proc. AISTATS. (2010) 249–256
9. Bourlard, H., Morgan, N.: Connectionist speech recognition: a hybrid approach. Kluwer Academic (1994)
10. Young, S., et al.: The HTK book. Cambridge Univ. Engineering Department (2005)
11. Lamel L., Kassel R., S.S.: Speech database development: Design and analysis of the acoustic-phonetic corpus. In: DARPA Speech Recognition Workshop. (1986) 121–124
12. Abari, K., Olaszy, G., Zainkó, C., Kiss, G.: Hungarian pronunciation dictionary on Internet (in Hungarian). In: Proc. MSZNY. (2006) 223–230
13. Tóth, L., Grósz, T.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: TSD. (2013) 36–43

Lexikai modellezés a közlés tervezettségének függvényében magyar nyelvű beszédfelismerésnél

Tarján Balázs[1], Fegyó Tibor[1,2], Mihajlik Péter[1,3]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
[2] AITIA International Zrt.
[3] THINKTech Kutatási Központ Nonprofit Kft.
{tarjanb, mihajlik, fegyo}@tmit.bme.hu

Kivonat: A morfémákban gazdag nyelvek nagyszótáras, gépi beszédfelismerésénél gyakran használnak szónál kisebb elemekre, ún. morfokra épülő nyelvi modelleket. Ezek alkalmazása azonban többletmunkát, magasabb rendszerkomplexitást igényel, ugyanakkor a javulás mértéke változó. Cikkünkben a morfalapú nyelvi modellezéssel elérhető hibacsökkenés előrejelzésére teszünk kísérletet. Ehhez először azonosítjuk a hibacsökkenést befolyásoló tényezőket, majd kísérleti úton megvizsgáljuk pontos hatásukat. Eredményeink alapján elmondható, hogy a morfalapú modellek alkalmazása kisméretű tanítószövegek, illetve korlátozott szótárméret mellett járhat jelentős előnnyel. Előnyös még a kevésbé spontán, tervezettebb beszédet tartalmazó adatbázisok esetén, míg a jel-zaj viszony romlása csökkenti a hibacsökkenés mértékét, csakúgy, mint az abszolút hibát. Az utolsó fejezetben bemutatunk egy mérőszámot, mely erős összefüggést mutat a kísérleti adatbázisainkon mérhető morfalapú hibacsökkenéssel. Ez a mérőszám nem csak a feladat tervezettségét, hanem a tanítószöveg mennyiségét is figyelembe veszi.

1 Bevezetés

Gépi beszédfelismeréssel, vagyis az automatikus beszéd-szöveg átalakítást lehetővé tevő megoldásokkal korábban csak a modern technológiák iránt elszántan érdeklődő kevesek találkozhattak. Túlzás lenne azt állítani, hogy azóta mindennapjaink része, tény viszont, hogy az okostelefonok terjedésével immáron **rengeteg felhasználó számára vált elérhetővé** egy-egy a technológia élvonalába tartozó megoldás, akár magyar nyelven is. Adja magát a következtetés, hogy ezek szerint a technológia beérett, és innentől kezdve csak apró finomításokra van szükség. A szakirodalmat tanulmányozva, vagy akár csak egy korszerű rendszert huzamosabb ideig tesztelve azonban láthatjuk, hogy ez távolról sincs így. A legelterjedtebb rejtett Markov-modell (Hidden Markov Modell – **HMM**) alapú statisztikai felismerők teljesítménye továbbra is durván leromlik zajos környezetben a humán észlelőkkel szemben, illetve akusztikus és nyelvi modelljeink továbbra is csak egyelőre meghatározott felismerési feladatra működnek optimálisan.

A fent vázolt problémák okainak jobb megértését tűzte ki célul az ún. OUCH (Outing Unfortunate Characteristics of HMMs) projekt. A kutatást összefoglaló ta-

nulmány [1] szerzői egyrészt rámutatnak a jelenleg használt technológiák hiányosságaira, majd bemutatják a technológia legfontosabb szereplőivel készített interjúik eredményét is. Ez alapján szakmai konszenzus van azzal kapcsolatban, hogy se az akusztikus és nyelvi modellek, se a beszédjellemzők kinyerését célzó technikák **nem tekinthetőek érettnek**, ugyanis működésük nem elég robosztus, még akkor sem, ha nagyon sok pénzt fektetnek a fejlesztésükbe. A tanulmány egyik fontos végkövetkeztetése, hogy beszédfelismerés területén dolgozó kutatóknak több energiát kellene fordítaniuk a felismerési hibák okainak mélyebb megértésére.

Cikkünk célja ezzel összhangban, hogy jobban megismerjük a szó- és morfalapú nyelvi modellezés hibaarányai közötti összefüggéseket. Számos vizsgálat bizonyítja [2]–[4], hogy morfémákban gazdag nyelveken a folyamatos, nagyszótáras gépi beszédfelismerő rendszerek **hibája csökkenthető**, ha szavak helyett statisztikai úton nyert morfémákat (ún. morfokat) [5] alkalmazunk a nyelvi modellben. Semmi nem garantálja azonban, hogy ez a hibacsökkenés jelentős mértékű lesz, sőt azt sem, hogy nem növekszik a hiba [6]. Figyelembe véve a többletmunkát és komplexitás növekedést, amivel a morfalapú rendszerek tanítása jár, felmerül az igény a **várható hibacsökkenés előrejelzésére**.

Korábbi munkáinkban megvizsgáltuk a szöveges tanítóadat mennyiségének, az akusztikus modell illeszkedésének és a felismerési feladat tervezettségének kapcsolatát az elérhető hibacsökkenéssel [7], [8]. Mostani munkákban egyrészt szeretnénk korábbi megállapításainkat új adatbázisokon is tesztelni, valamint kiterjeszteni az ún. **követőmorfos** [4] nyelvi modellekre is. Ezenkívül új szempontként megvizsgáljuk a feldolgozandó hanganyag jel-zaj viszonyának illetve a felismerő rendszer szótárméretének hatását is. Morfalapú nyelvi modellezés esetén sajnos elkerülhetetlen, hogy valamilyen típusú speciális jelölést vezessünk be a szóhatár későbbi visszaállíthatósága érdekében. Érdekes kérdés, hogy mennyivel lehetne pontosabb egy olyan morfalapú rendszer, melyben **eltekintünk a szóhatár-visszaállítástól**. Cikkünkben ennek a meghatározására is kísérletet teszünk.

A következőkben először a televíziós híradók felvételeit tartalmazó tanító- és tesztadatbázist ismertetjük, majd kitérünk a modellek tanításnál és kísérleteinknél alkalmazott módszerekre. A felismerési feladat és módszertan bemutatása után ismertetjük a híradó adatbázison kapott eredményeket, majd az utolsó előtti fejezetben a hibacsökkenés előrejelzésére teszünk kísérletet. Végül összefoglalását adjuk vizsgálataink legfontosabb eredményeinek.

2 Tanító és tesztadatbázis

Kísérleteink döntő többségét egy **televíziós híradófelvételeket** tartalmazó adatbázison végeztük. Ilyen típusú – az angol terminológia szerint *broadcast speech*nek nevezett – adatbázison már korábban is kísérleteztünk [6], [7], [9], melyek tapasztalatait cikkünkben a vonatkozó részeknél felidézzük majd. Hasonló magyar nyelvű felismerési feladaton két további munkát fontos megemlíteni. Az első, mély neuronhálók tanítási módszereit veti össze, mely technika segítségével meg is javítja a HMM alapú akusztikus modell eredményét híradós felvételek felhasználásával [10]. Míg egy másik a témában született cikkben elsősorban a kézi leiratok felhasználása nélkül törté-

nő, felügyelet nélküli tanítási módszereken van a hangsúly [11]. A fenti két cikk egyike sem alkalmaz azonban morfalapú lexikai modelleket.

2.1 Akusztikus tanító- és tesztanyagok

Összesen 50 órányi televíziós híradó kézi leiratát készítettük el és használtuk fel a felismerő akusztikus tanításához. Tesztelési célokra 6 teljes híradó felvételét, összesen 155 perc hanganyagot különítettünk el. A 6 híradó mindegyike 2012 januárjában került adásba a TV2, a Duna TV és az MTV műsorán. A tesztanyagot két részre bontottuk: 2 híradót a fejlesztés során szükség paraméterek hangolására (**Dev**), míg a maradék 4-et a rendszer kiértékeléséhez (**Eval**) használtunk fel. A tesztanyag kézi átírásakor az egyes szegmenseket akusztikai tulajdonságaik szerint különböző csoportokba soroltuk. Az egyes csoportok jelentését és méretének eloszlását az **1. táblázatban** foglaljuk össze. Fontos megjegyezni, abban az esetben, ha egy szegmensre többféle kategória leírása is illett, akkor mindig a nagyobb sorszámú kategóriába soroltuk. Ebből következik például, hogy az F4 kategóriájú szegmensekben a zaj nem biztos, hogy a szegmens teljes hosszára kiterjed. A jel-zaj viszony (Signal-to-Noise Ratio – **SNR**) hozzávetőleges meghatározásához a NIST STNR[1] és WADA SNR [12] algoritmusokat alkalmaztuk. Ahol kevés adat állt rendelkezésre ott nem adtuk meg az SNR-t, mivel az algoritmusok nem szolgáltak megbízható értékkel.

1. táblázat: A tesztadatbázis eloszlása akusztikai kategóriák szerint

Jelölés	Jelentés	Hossz [perc]	SNR [dB]
F0	Tervezett beszéd csendes környezetben	38	20-25
F1	Spontán beszélgetés csendes környezetben	18	20-25
F2	Telefonos beszélgetés	2	-
F3	Beszéd háttérzenével	10	8-10
F4	Beszéd zajos környezetben	84	10-15
F5	Nem anyanyelvi beszélő	3	-

2.2 Szöveges tanítóanyagok

A felismerő nyelvi modelljének tanításához szükséges szövegkorpuszokat több forrásból gyűjtöttük össze. Egyrészt felhasználtuk az akusztikus modell tanításához használt 50 órányi hanganyag kézi leiratát (**TRS**). Ez önmagában túl kevés lett volna egy hatékony felismerő tanításához, így különböző webes híroldalakról is gyűjtöttünk további adatokat. A webes szövegek gyűjtésével, tárolásával és feldolgozásával kapcsolatos részletek [9]-ben találhatóak meg. Az ott bemutatott rendszerhez hasonlóan most is két részre bontottuk a webes tanítószöveget. Az első a tesztanyag előtti 30 napban (2011. december 1-31.) közölt híreket tartalmazza (**WEB 30D+**), míg a második minden anyagot, ami korábban keletkezett (**WEB 30D-**). Ennek a szétválasztásnak az a célja, hogy a nyelvi modellek interpolációja során a tesztelés időpontjához közelebb eső hírek nagyobb súlyt kaphassanak, mint a régebbiek. Részletes adatok a

[1] http://labrosa.ee.columbia.edu/~dpwe/tmp/nist/doc/stnr.txt

2. táblázatban találhatóak. A korábbi rendszerhez képest lényeges eltérés azonban, hogy a tanítószöveg normalizálása során a kivételes kiejtéssel rendelkező szavakat (tipikusan nevezett entitásokat) a kiejtett alakjuknak megfelelően írtuk át. Ennek a változtatásnak a célja az volt, hogy a szó- és morfalapú eredmények összehasonlíthatóságát ne befolyásolják a kivételes írásmódú szavak szegmentálási nehézségei.

2. táblázat: A tanítószövegek részletes adatai

Tanítószöveg	Összes szó [millió szó]	Szótárméret [ezer szó]	Eval PPL [-]	Eval OOV [%]
TRS	0,53	74	598	10.0
WEB 30D+	1,2	115	761	8.3
WEB 30D-	50,1	955	551	1.5

3 Tanítási és kísérleti módszerek

Ebben a fejezetben a tanítás során alkalmazott modellezési lépéseket és a kísérleti körülményeket ismertetjük.

3.1 Akusztikus modell

A híradó felismerési feladathoz tartozó akusztikus modell tanításához az erre a célra elkülönített 50 óra hanganyagot használtuk fel. Az annotált felvételek segítségével háromállapotú, balról-jobbra struktúrájú, környezetfüggő rejtett Markov-modelleket tanítottunk a Hidden Markov Model Toolkit [13] eszközeinek segítségével. A létrejött akusztikus modell 4630 egyenként 15 Gauss-függvényből álló állapotot tartalmaz. Híranyag felismerési kísérleteink során minden esetben ezt az akusztikus modellt használtuk. A lexikai elemek fonetikus átírását a magyar nyelv hasonulási tulajdonságait figyelembe vevő, automatikus eljárással készítettük.

3.2 Nyelvi modellek

A felismerési tesztjeinkben használt összes nyelvi modell módosított Kneser-Ney simítás használatával készült az SRI Language Modeling Toolkit (SRILM) [14] segítségével. Az interpolált nyelvi modellek készítéséhez és optimalizálásához az SRILM beépített lineáris interpolációs és perplexitás számító eljárásait használtuk. A nyelvi modellek fokszámát minden modell esetén egyedileg optimalizáltuk.

3.2.1 Morfszegmentálás

A morfémákban gazdag nyelvek esetén (pl. magyar, finn, török, észt, stb.) visszatérő probléma szóalapú nyelvi modellezés (**WORD**) esetén a nagy szótárméret, az ebből fakadó adatelégtelenségi problémák, illetve az a tény, hogy még sok tanítóadat esetén is nagy lehet a szótáron kívüli szavak (Out of Vocabulary – **OOV**) aránya a felisme-

résí feladatban. Gyakori megközelítés a probléma enyhítésére, hogy a szótárat kisebb, ám gyakrabban előforduló elemekre darabolják fel, így csökkentve a szótárméretet és növelve a tanítóminták számát. Kísérleteinkben a szótári elemeket egy elterjedten alkalmazott felügyelet nélküli szegmentáló eljárással [5] ún. **morfokra** bontjuk:

„utalt az okra az uniós alapelv ekre amely eket a tagállam oknak tisztelet ben kell tartani uk”

3.2.2 Szóhatár-visszaállítás

A gépi beszédfelismerő kimenetén szóhatárok mentén szegmentált szöveget várunk, ezért a morfalapú rendszer tanítószövegében valamilyen módon jelölnünk kell azokat. Erre kétféle technikát alkalmazunk. Az első ún. **szóhatár-jelölő morfos** (word boundary tag – **WB**) megközelítésnél egy dedikált szimbólumot használunk a különböző szavakhoz tartozó morfok elválasztásához:

„utalt <w> az okra <w> az <w> uniós <w> alapelv ekre <w> és <w> jogszabály okra <w> amely eket <w> a <w> tagállam oknak <w> tisztelet ben <w> kell <w> tartani uk”

A módszer előnye, hogy mindössze egyetlen plusz szótári elem bevételével kezelni tudjuk a szóhatár-visszaállítás problémáját. Hátrány viszont, hogy ez a szimbólum nagyon gyakori elemé válik, és így rontja az n-gram modell predikciós képességét. A másik megoldás az ún. **követőmorfos** (Non-initial morph – **NI**) jelölés, azaz amikor minden olyan morfot megjelölünk a szövegben, mely nem az első tagja egy szónak:

„utalt az -okra az uniós alapelv -ekre és jogszabály -okra amely -eket a tagállam -oknak tisztelet -ben kell tartani –uk”

Előnyös tulajdonsága a módszernek, hogy kevesebb jelölésre van szükség, mivel a szavak többsége egyetlen morfból áll. Hátrány azonban, hogy jelentős mértékben megnőhet a szótárméret a szóhatár-jelöléses módszerhez képest, hiszen az egyes morfok követőmorfos és szó eleji alakjait meg kell különböztetni egymástól. Igaz, a szóalapú rendszerhez képest még így is jelentős lehet a szótárméret csökkenése.

3.3 Hálózatépítés és dekódolás

A 16 kHz-en mintavételezett felvételek lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre, és ún. vak csatornakiegyenlítő eljárást is alkalmaztunk. A súlyozott véges állapotú átalakítókra (Weighted Finite State Transducer – **WFST**) épülő felismerő hálózatok generálását és optimalizálását az Mtool keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztéshez a VOXerver [6] nevű WFST dekódert használtuk. A cikkünkben összehasonlított szó- és morfalapú rendszerek futásidejében keletkező különbségeket minden esetben kiegyenlítettük a keresési szélesség hangolásával. A felismerő rendszerek teljesítményének értékeléséhez szóhiba-arányt (Word Error Rate – **WER**) illetve néhány esetben betűhiba-arányt (Letter Error Rate – **LER**) számoltunk.

4 Kísérleti eredmények a híradó adatbázison

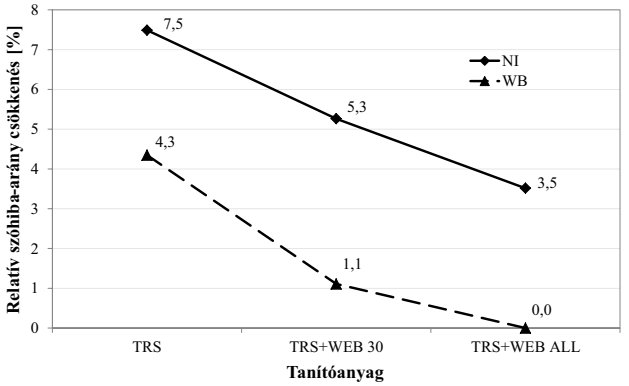
4.1 Morfalapú hibacsökkenés a tanítószöveg méretének függvényében

Első kísérletünk célja a morfalapú nyelvi modellekkel elérhető felismerési hibaarány-csökkenés és a tanítóadat-mennyiség közötti összefüggés vizsgálata volt. Korábbi munkáinkban [7], [8] arra jutottunk, hogy a morfalapú módszerek előnye a tanítószöveg méretének növekedésével egyre csökken, sőt bizonyos méret fölött teljesen el is tűnik [6]. Fontos megjegyezni ugyanakkor, hogy a fent idézett cikkeinkben csak a szóhatár-jelöléses (WB) megközelítést vizsgáltuk. Mostani összevetésünkben három mérési pontot alkalmazunk. Az első esetben csak a kézi leiratok (**TRS**) alapján tanítjuk a modelleket. Második esetben a TRS és WEB 30+ korpuszok alapján (**TRS+WEB 30**), míg a harmadik esetben a TRS, a WEB 30+ és WEB 30- korpuszokat is felhasználjuk a tanításhoz (**TRS+WEB ALL**). A nyelvi modellek interpolációs súlyát a tesztanyag Dev halmazán optimalizáltuk.

3. táblázat: Felismerési eredmények különböző méretű tanítószövegekkel

Tanító-anyag	Lexikai modell	N-gram fokszám	Szótár-méret [ezer szó]	Dev WER [%]	Eval WER [%]	Rel. WER csök. [%]
TRS	WORD	3	74	38,2	41,4	
	WB	4	12	35,2	39,6	*-4.3*
	NI	4	16	34,8	38,3	*-7.5*
TRS+ WEB 30	WORD	3	151	32,4	36,1	
	WB	4	24	31,1	35,7	*-1.1*
	NI	4	31	30,9	34,2	*-5.3*
TRS+ WEB ALL	WORD	4	978	23,0	25,6	
	WB	5	175	23,1	25,6	*0.0*
	NI	4	204	22,3	24,7	*-3.5*

A felismerési eredményeket a **3. táblázatban** foglaltuk össze. A morfalapú rend-

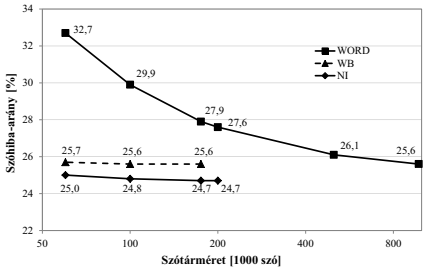


1. ábra. A különböző méretű tanítószövegek alapján tanított morf nyelvi modellekkel elérhető relatív szóhiba-arány csökkenés

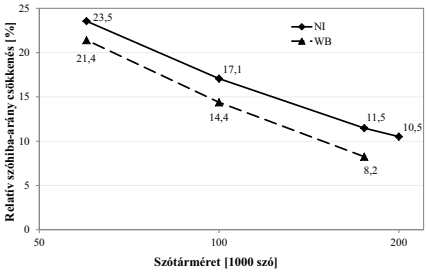
szerekkel a szóalapú rendszerhez képest elérhető relatív szóhiba-arány csökkenéseket az **1. ábrán** mutatjuk be. A WB típusú morfmodell – korábbi eredményeinkkel összhangban – a tanítószöveg méretének növekedésével elveszti az előnyét a szóalapú rendszerhez képest. Ellenben az NI típusú modell még a legnagyobb modellméret mellett is 3,5%-kal jobban teljesít. Igaz a tendencia itt is arra utal, hogy idővel elvész az előny. Érdemes lehet a jövőben még az eddigieknél is nagyobb tanítószöveget bevonni a vizsgálatainkba.

4.2 Morfalapú hibacsökkenés a szótárméret függvényében

Minden korábbi vizsgálatunkban teljes szótárméret mellett hasonlítottuk össze a szó- és morfalapú nyelvi modelleket. Döntésünknek az volt az oka, hogy úgy éreztük, aránytalanul nagy előnyt élveznének a morfalapú megközelítések, ha szóalapú rendszer szótárméretét velük megegyező szintre csökkentenénk. Ezzel szemben számos tanulmányban [15], [16] kiegyenlített szótárméret mellett mérik a hibaarány csökkenést. Itt két szemlélet ütközik. Az egyik szerint a rendszereket nem csak azonos számításigény, de azonos memóriaigény mellett kell vizsgálni. A memóriaigényt azonban csupán a szótárérettel nem lehet kézben tartani, így adott tanítószöveg esetén érdemesebb megpróbálni kihozni a maximumot a vizsgált modellezési technikából. Elfogadva mindkét szemlélet létjogosultságát célunk az volt, hogy kimutassuk a kettő



2. ábra. A különböző szótárméretű nyelvi modellekkel mért felismerési hibák



3. ábra. Relatív szóhiba-arány csökkenés a szótárméret függvényében

közötti különbséget. Méréseinket 60, 100, 175, 200, 500 és 978 ezres szótárméret mellett végeztük a TRS+WEB ALL nyelvi modellek felhasználásával.

A **2. ábrán** jól kivehető, hogy míg a szóalapú nyelvi modellek felismerési hibája erősen függ a szótárérettől, addig a morfalapú modellek az általunk vizsgált 60 ezres határig nagyjából érzéketlenek rá. Ebből következik az is, hogy a szótárméret csökkentésével szignifikánsan nagyobb morfalapú hibaarány-csökkenést mérhetünk, mintha az egyes modelleket teljes szótárméret mellett hasonlítanánk össze (**3. ábra**). Nem meglepő módon a követőmorfos technika ebben a kísérletben is őrizte az előnyét a szóhatár-jelöléses megközelítéshez képest.

4.3 Morfalapú hibacsökkenés a tervezettség és a jel-zaj viszony függvényében

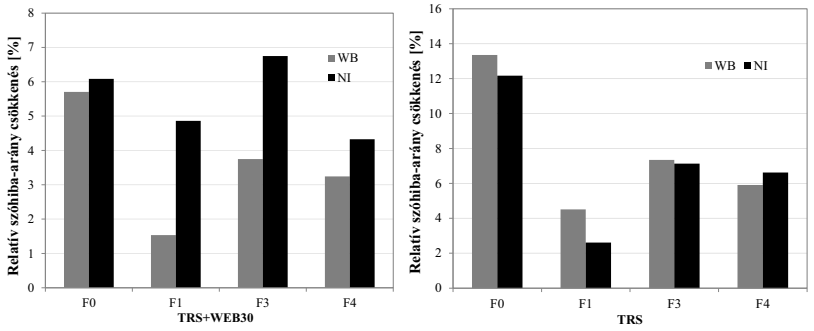
Mint a 2.2-es fejezetben ismertettük, rendelkezésünkre áll a tesztanyag akusztikai viszonyok és tervezettség szerinti felbontása is, melynek köszönhetően a morfalapú

hibacsökkenés mértékét vizsgálhatjuk e paraméterek tekintetében is (**4. táblázat**). Az F0, F2, F3 és F4 kategóriák összehasonlításával képet kaphatunk arról a jelenségről, melyre már a bevezetőben is utaltunk. A tiszta, tervezett beszéd (F0) felismerése még kevés szöveges tanítóadat esetén is tolerálható hibával jár (~30%), növelve a tanítószöveg mennyiségét viszont nagyon alacsony szinte is csökkenthető (~16%). Ez a hibaarány például már lehetővé teszi egy jól érhető felirat készítését a televíziós anyagokhoz. Látható azonban, hogy a jel-zaj viszony csökkenésével (F3, F4) jelentősen megnő a hibák száma. 10-15 dB-es átlagos jel-zaj viszony mellett már 10-15%-kal magasabb hibát kapunk. Telefonos beszéd esetén (F2) nehézséget jelent az alacsony spektrális sávszélesség, így nem véletlen, hogy erre a feladatra külön akusztikus modellt szokás tanítani [17]. Nem csak az akusztikus körülmények játszanak azonban fontos szerepet. A hibaarány még magas jel-zaj viszony esetén is megnőhet, ha tervezett helyett spontán vagy félig spontán (F1) beszédet írunk át, melynek oka a szavak nehezebb predikálhatóságában és a lazább artikulációban keresendő [7].

4. táblázat: Felismerési eredmények F-kategóriák szerinti felbontásban

Tanító-anyag	Lexikai modell	Dev+Eval WER [%] F0	F1	F2	F3	F4	F5
TRS	WORD	33.7	42.1	76.0	46.3	42.3	55.4
	WB	29.2	40.2	67.4	42.9	39.8	54.7
	NI	29.6	41.0	67.1	43.0	39.5	52.2
TRS+ WEB30	WORD	26.3	39.1	67.1	40.0	37.0	49.5
	WB	24.8	38.5	60.5	38.5	35.8	48.7
	NI	24.7	37.2	64.7	37.3	35.4	47.3
TRS+ WEB ALL	WORD	15.6	31.5	56.2	28.7	27.5	38.2
	WB	15.9	29.7	53.5	30.7	27.4	35.8
	NI	15.6	29.5	56.6	28.1	26.6	37.6

A számszerű felismerési hibák mellett érdemes megvizsgálni a morfalapú módszerekkel nyerhető hibacsökkenést is. Ennél a vizsgálatnál az F2 és F5 kategóriákat figyelmen kívül hagytuk, ugyanis a tesztanyag csak nagyon kis része tartalmaz ilyen mintákat (**1. táblázat**). Érdemi következtetéseket csak az F0 és F4 kategória eredményeiből érdemes levonni, mivel ezek a kategóriák képezték a tesztanyag legnagyobb

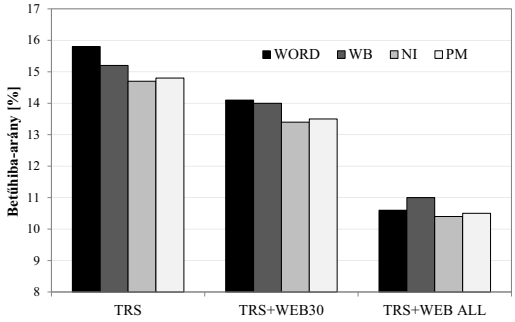


4. ábra. A morfalapú módszerekkel nyerhető relatív szóhiba-arány csökkenés a tervezettség és a jel-zaj viszony függvényében

részét. A **4. ábrán** megfigyelhető, hogy a tiszta, tervezett felvételeken (F0) jelentősen nagyobb a morfalapú hibacsökkenés, mint az alacsony jel-zaj viszonnyal rendelkezőkön (F4). Hasonló eredményre jutottunk korábban, amikor az akusztikus modell illeszkedésének hatását vizsgáltuk [8]. A jelenség oka abban keresendő, hogy morfok jellemzően a szavaknál kevesebb fonémából állnak, így akusztikailag könnyebben összetéveszthetőek. A jobb jel-zaj viszony és akusztikus modell segíti kiemelni a morfalapú nyelvi modell előnyeit. Részben ugyanerre vezethető vissza az F0 és F1 kategóriák közötti különbség is. Itt a jel-zaj viszony megegyezik, azonban a tervezett beszédhez jobban illeszkedik az akusztikus modell, illetve a morfalapú nyelvi modell.

4.4 Szóhatár-jelölés hatása a felismerési hibára

A híradó felismerő rendszerünk kiértékelésének utolsó lépésében azt vizsgáltuk, hogy a morfalapú nyelvi modellekben használt szóhatár-jelölésnek milyen hatása van a felismerés pontosságára. Hipotézisünk szerint ezek csökkentik a morfalapú modellezés hatékonyságát, ám használatuk elkerülhetetlen a szóhatár-visszaállítás miatt. A rendszerek összehasonlításához itt természetesen szóhiba-arány mérése nem jöhetett szóba, ezért betűhiba-arányokat mértünk (**5. ábra**). Meglepő módon előzetes feltevé-



5. ábra. Betűhiba-arányok a híradó Eval tesztadatbázison szóalapú, morfalapú és szóhatár-jelölés nélküli morfalapú nyelvi modellekkel

sünkkel ellentétes eredményt kaptunk. Bár a szóhatár-jelölés nélküli morfalapú rendszer (Pure morph – **PM**) minden tanítókorpusz méret mellett jobban teljesített, mint a szóalapú és WB morf modellezés, az NI morf megközelítést azonban nem tudta túlszárnyalni. Ebből azt a következtetést vonhatjuk le, hogy a szóhatár-jelölés csupán a WB technika esetén tekinthető szükséges rossz megoldásnak, az NI esetén még javítja is a modellezés pontosságát. Ez magyarázhatja tehát, hogy az NI megközelítés minden esetben alacsonyabb felismerési hibát eredményez, mint a WB. Meg kell jegyeznünk azonban, hogy a morfalapú modellek összevethetőségét valamelyest nehezíti, hogy a szóhatár-jelölés nélküli morfmodellben minden morf elejére helyeztünk egy opcionális szünetmodellt a kiejtési modellben, míg a szóhatár-jelöléses modelleknél (WB, NI) csak a lehetséges szóhatárokra.

5 Kísérletek a morfalapú hibacsökkenés előrejelzésére

Az előző fejezetben híradó felismerési feladaton mutattuk be a tanítókorpusz és a szótár méretének, valamint a jel-zaj viszonynak és a beszéd tervezettségének hatását a morfalapú hibacsökkenésre. Ebben a fejezetben az eddigi eredményeket kiegészítjük három különböző ügyfélszolgálati adatbázis eredményeivel, és kísérletet teszünk a hibacsökkenés mértékét előre jelezni. Egy korábbi cikkünkben [7] abból a feltételezésből indult ki, hogy szóalakok száma összefügg a feladat tervezettségével. Ezzel a megközelítéssel összefüggést találtunk egy adott méretű tanítószövegben előforduló szóalakok száma és a relatív szóhiba-arány csökkenés között. Láthattuk azonban, hogy a tervezettségen kívül más tényezők is szerepet játszanak. Jelenlegi vizsgálatunkban a tanítószöveg méretének hatását is megpróbáltuk figyelembe venni. Minden nyelvi modellben teljes szótárat használtunk, míg az akusztikai viszonyok hatását úgy igyekeztünk kiküszöbölni, hogy a kísérleteket beszélőfüggetlen akusztikus modellel végeztük. Az ügyfélszolgálati felismerések eredményei a Magyar Telefonos Ügyfélszolgálati Beszédadatbázisok (**MTUBA**) kiértékelésből származnak. Az MTUBA II. [18] biztosítási, az MTUBA III. távközlési, míg az MTUBA IV. [17] banki ügyfélszolgálati telefonbeszélgetések rögzítéséből származik. Az adatbázisokkal kapcsolatban bővebb információk a megjelölt hivatkozásokban találhatóak. A tanítószövegek adatait és a felismerési eredményeket az **5. táblázatban** ismertetjük.

5. táblázat: Ügyfélszolgálati felismerési eredmények

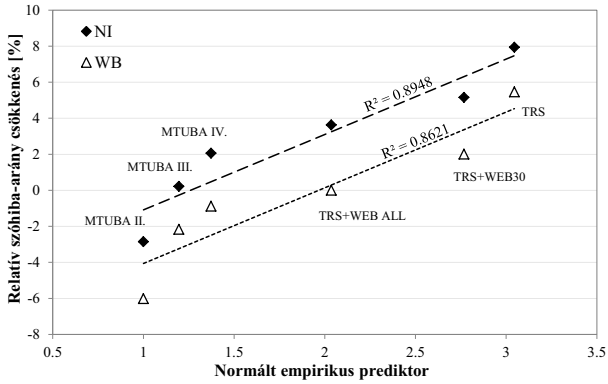
Tanító-anyag	Lexikai modell	$Tokens(T)$	$\frac{1}{n}\sum_{i=1}^{n} Types(T_i)$	WER [%]	Rel. WER csök. [%]
MTUBA II.	WORD	487	10938	31,6	
	WB			33,5	*-6.0*
	NI			32,5	*-2.8*
MTUBA III.	WORD	1313	14064	46,1	
	WB			47,1	*-2.2*
	NI			46,0	*0.2*
MTUBA IV.	WORD	794	15576	34,0	
	WB			34,3	*-0.9*
	NI			33,3	*2.1*

A morfalapú hibacsökkenés előrejelzéséhez először felbontjuk a tanítószövegeket *n* db, diszjunkt, egyenként *k* szót tartalmazó részkorpuszra (T_i). A részkorpuszok mérete kísérletünkben $k = 160000$.

$$\bigcap_{i=1}^{n} T_i = \emptyset \qquad Tokens(T_i, 0 \le i \le n) = k$$

Az empirikus prediktort ez alapján úgy számoljuk, hogy meghatározzuk az egyes részkorpuszokon mért szóalakszámok átlagát és elosztjuk a teljes tanítószöveg (T) szószámának logaritmusával.

$$Emprikus\ prediktor = \frac{\frac{1}{n}\sum_{i=1}^{n} Types(T_i)}{\log_{10} Tokens(T)}$$



6. ábra. Összefüggés a morfalapú rendszerekkel kapható szóhiba-arány csökkenés és a csökkenés mértékét előrejelző empirikus mérőszám között

A képletben új elemként a nevezőt azonosíthatjuk, melytől azt várjuk, hogy képes kompenzálni a hibacsökkenés függését a tanítószöveg méretétől. Az ügyfélszolgálati és híradós adatbázisokon mérhető összefüggést a prediktor és a morfalapú hibacsökkenés között a **6. ábrán** mutatjuk be. Mint az ábrán látható sikerült egy olyan immáron a tanítószöveg méretet is figyelembe vevő mértéket bevezetnünk, mely erősen korrelál a WB (R^2=0.86) és az NI morfalapú (R^2=0.89) megközelítésekkel kapható relatív szóhiba-arány csökkenéssel.

6 Összefoglalás

Cikkünkben arra kerestük a választ, hogy milyen tényezőktől függ a morfalapú nyelvi modellezéssel, a szóalapú rendszerekkel szemben elérhető hibacsökkenés a nagyszótáras gépi beszédfelismerésben. Televíziós híradók adatbázisán végzett kísérleteink rámutattak, hogy a tanítószöveg méretének növekedésével csökken az adatelégtelenség, így a morfalapú rendszerekkel nyerhető előny is. Megállapítottuk, hogy a morfalapú nyelvi modellek kisebb, tömörebb szótáruk miatt keveset veszítenek pontosságukból még akkor is, ha erősen korlátozzuk a szótár méretét. Azaz a morfalapú rendszerek használata kevés és korlátozott szótárméretű tanítószöveg esetén lehet különösen előnyös. Számszerűsítettük továbbá a jel-zaj viszony romlásának hatását, illetve azt is megmutattuk, hogy a közlés tervezettségnek növekedése, hogyan növeli a várható hibacsökkenést.

Az utolsó fejezetben bevezettünk egy mérőszámot, mely erős összefüggést mutat három telefonos ügyfélszolgálati és a híradó adatbázison mérhető morfalapú hibacsökkenésekkel. Ez mérőszám jelenleg csak a feladat tervezettségét és tanítószöveg mennyiségét veszi figyelembe, így a jövőben szeretnénk továbbfejleszteni oly módon, hogy az akusztikus körülményeket is számításba vegye. Emellett, de ettől nem teljesen függetlenül elméleti magyarázatot is szeretnénk találni a morfalapú hibacsökkenést befolyásoló egyes tényezők közötti összefüggésekre.

Köszönetnyilvánítás

Köszönettel tartozunk Balog Andrásnak a híradó tesztanyag F-kategória szerinti válogatásáért, illetve Sárosi Gellértnek az ügyfélszolgálati rendszerek kiértékeléséért. Kutatásunkat a Mindroom (KMOP-1.1.3-08/A-2009-0006), WEBRA TIME SAVE (GOP-1.1.1-11-2012-0377), FuturICT.hu (TÁMOP-4.2.2.C-11/1/KONV-2012-0013) és DIANA (KMR_12-1-2012-0207) projektek támogatták.

Hivatkozások

1. N. Morgan, J. Cohen, S. H. Krishnan, S. Chang, and S. Wegmann, “Final Report : OUCH Project (Outing Unfortunate Characteristics of HMMs),” 2013.
2. P. Mihajlik, Z. Tuske, B. Tarján, B. Németh, and T. Fegyó, “Improved Recognition of Spontaneous Hungarian Speech—Morphological and Acoustic Modeling Techniques for a Less Resourced Task,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 18, no. 6, pp. 1588–1600, Aug. 2010.
3. M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, “Unlimited vocabulary speech recognition for agglutinative languages,” in HLT-NAACL 2006, 2006, pp. 487–494.
4. E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, “Turkish Broadcast News Transcription and Retrieval,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 17, no. 5, pp. 874–883, Jul. 2009.
5. M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0,” in Publications in Computer and Information Science, Report A81, 2005.
6. B. Tarján, P. Mihajlik, A. Balog, and T. Fegyó, “Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection,” in 2nd International Conference on Cognitive Infocommunications (CogInfoCom), 2011, pp. 1–5.
7. B. Tarján and P. Mihajlik, “On morph-based LVCSR improvements,” in Spoken Language Technologies for Under-Resourced Languages (SLTU-2010), 2010, pp. 10–16.
8. L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, “Speech Recognition Experiments with Audiobooks,” Acta Cybern., vol. 19, no. 4, pp. 695–713, 2010.
9. B. Tarjan, T. Mozsolics, A. Balog, D. Halmos, T. Fegyo, and P. Mihajlik, “Broadcast news transcription in Central-East European languages,” in IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), 2012, pp. 59–64.
10. L. Tóth and T. Grósz, “A Comparison of Deep Neural Network Training Methods for Large Vocabulary Speech Recognition,” pp. 1–8, 2012.
11. A. Roy, L. Lamel, T. Fraga, J. Gauvain, and I. Oparin, “Some Issues affecting the Transcription of Hungarian Broadcast Audio,” in 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), 2013, no. August, pp. 3102–3106.
12. C. Kim and R. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in INTERSPEECH, 2008, pp. 2598–2601.
13. S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, The {HTK} Book, version 3.4. Cambridge, UK: Cambridge University Engineering Department, 2006.
14. A. Stolcke, “SRILM – an extensible language modeling toolkit,” in Proceedings International Conference on Spoken Language Processing, 2002, pp. 901–904.

15. V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in Proc. Eurospeech’03, 2003, pp. 2293–2296.
16. M. A. Basha Shaik, A. El-Desoky Mousa, R. Schlüter, and H. Ney, “Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR,” in Interspeech, 2011, pp. 1441–1444.
17. B. Tarján, G. Sárosi, T. Fegyó, and P. Mihajlik, “Improved Recognition of Hungarian Call Center Conversations,” in The 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), 2013, pp. 65–70.
18. G. Sárosi, B. Tarján, T. Fegyó, and P. Mihajlik, “Automated Transcription of Conversational Call Center Speech – with Respect to Non-verbal Acoustic Events,” Intell. Decis. Technol.

A HuComTech audio adatbázis szintaktikai szintjének multimodális vizsgálata

Kiss Hermina

HuComTech Group, Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
4032 Debrecen, Egyetem tér 1.
kissh3@gmail.com

Kivonat: A HuComTech spontán nyelvi adatbázis szintaktikai szintjének elveiről és szabályrendszerének újdonságairól a VIII. Magyar Számítógépes Nyelvészeti Konferencián számoltam be 2011-ben [4]. Bemutattam többek között a mondat fogalmának újszerű értelmezését, a saját elméleti rendszerünkön belül létrejött tagmondatok közötti hierarchiák jelentőségét, az implicit nyelvi elemek kimutatásának elvi és gyakorlati lehetőségeit, valamint ismertettem a legfontosabb alapelveket (szabályrendszerünk legyen preteoretikus, tükrözze a különböző tudományok közötti konszenzust, valamint legyen alulspecifikált). Idén az azóta megszületett kutatási eredményeket és az annotálási munka folyamatát szeretném felvázolni. Különös hangsúlyt fektettünk arra, hogy egy speciális kódolási rendszer kifejlesztésével a spontánbeszéd jellegzetességeit kezelhetővé tegyük a magyar nyelv mondattani keretei között, mint ami rendszerében nem, csak megvalósulásában különbözik attól. A spontánbeszéd-kutatás szintaktikai elemzésének speciális jellegét azzal lehet leginkább kiemelni, ha különös figyelmet fordítunk az implicit nyelvi elemek összegyűjtésére és rendszerezésére, valamint a tagmondatok hierarchiájának jellemzésére. Mivel jelen esetben két személy közötti kommunikáció szintaktikai elemzéséről van szó, minden esetben az egyes beszélők által megvalósított egyes fordulókat tekintjük az elemzés tárgyának. Az egyes fordulókon belül azonosítjuk a szintaktikai struktúrát.

1 Bevezetés

Munkánk célja az, hogy egy olyan szintaktikai elemzési módszert teszteljünk, amely lehetővé teszi a korpuszon belül a multimodális együttállások vizsgálatát. Azt kutatjuk, hogy a szintaktikai annotálás mint a multimodális annotálás része hozzájárulhat-e ahhoz, hogy az egyik modalitásból hiányzó elemet egy másik modalitás ugyanazon időpillanatában kutathassuk, tehát egy befejezetlen mondat kézmozdulatokkal, mimikával való lezárását nyomon követhetjük-e szintaktikai szinten is. Azt feltételezzük, hogy a nyelvi kompetencia jelen van a kognícióban, és a felszínre kerülő, az egyidőben sokféle folyamatot magába foglaló tudati tevékenység bizonyos (nem teljes) lenyomatát képező nyelvi elemek mellett a nem nyelvi formák is szerepet játszanak a szintaktikai szerkezet kialakításában. Tehát a konverzációban megjelenő grammatikai hiányok gesztusokkal és mimikákkal egészülnek ki, teljesebbé téve a grammatikai szerkezetet.

A beszélt nyelvi szintaxist úgy közelítjük meg, hogy nagy jelentőséget tulajdonítunk az implicit nyelvi elemeknek. A hiánya felől ábrázoljuk a nyelvi struktúrákat, mert kutatásunk alapját az explicit módon jelen nem lévő nyelvi elemek által létrejött információhiány feldolgozása képezi.

Előadásom első felében szeretném csoportosítani az annotálási folyamat problémaköreit a fent említett cél tükrében. Kidolgoztunk számos problémakört, amelyek a hiány kategóriáinak átgondolását segítették. Vizsgáltuk például az egyedi szó- és nyelvhasználatból adódó jelenségeket, sajátosságokat, vagy a töltelékszavak, indulatszavak spontán beszédbe illeszkedő rendszerét, illetve a pongyola nyelvhasználat következményeként létrejövő szintaktikai problémákat. (Mint például az abszolút és relatív főnév elhelyezkedése a mondat hierarchiájában, a kötőszóval kezdődő mondatok kérdése, a főnevesült jelző mondattani szerepköre, a függő beszédben jelen lévő implicit elemek, az ellipszis számos kérdésköre, illetőleg a dialógus másik szereplőjének a vizsgált személy grammatikai szerkezetére tett hatása).

Az előadás második felében pedig a – fentebb felvázolt – hiányok felől megközelített szintaktikai struktúrák multimodális együttállásai által megszületett kutatási eredményeket és magát a kutatási folyamatot mutatom be.

2 A HuComTech korpusz szintaktikai szintjének legfontosabb annotálási problémái

A 2011-es MSZNY Konferencián csoportosítottuk az annotálási folyamat problémaköreit a fent említett cél tükrében. Később kidolgoztunk számos problémakört, amelyek a hiány kategóriáinak átgondolását segítették. Vizsgáltuk például az egyedi szó- és nyelvhasználatból adódó jelenségeket, sajátosságokat, vagy a töltelékszavak, indulatszavak spontán beszédbe illeszkedő rendszerét, illetve a pongyola nyelvhasználat következményeként létrejövő szintaktikai problémákat. (Mint például az abszolút és relatív főnév elhelyezkedése a mondat hierarchiájában, a kötőszóval kezdődő mondatok kérdése, a főnevesült jelző mondattani szerepköre, a függő beszédben jelen lévő implicit elemek, az ellipszis számos kérdésköre, illetőleg a dialógus másik szereplőjének a vizsgált személy grammatikájára tett hatása). A szintaktikai szabályrendszer a beszélt nyelv mondattani jellegzetességeinek kutatása céljából jött létre. Ezek a jellegzetességek a szabályrendszer alapján a hiány kategóriáiból vezethetők le. A hiány kategóriáit vizsgálva hat fő probléma köré csoportosíthatjuk a spontán beszélt nyelv mondattani karakterét. A problémák egy része és azok feloldásai már elhangzottak a 2011-es MSZNY Konferencián, ezért csak utalásszerűen jegyezzük meg. Átgondoltuk a bevezetett fogalmak definícióit: a tagmondat fogalmát, a minimális mondat fogalmát, illetve a minimális mondat bevezetésének problémakörét, a teljes és hiányos tagmondat fogalmát, a befejezetlen mondat fogalmát, a hiány kategóriáit, azaz milyen szintaktikai elemek hiányozhatnak egy tagmondatból, valamint a felszínen jelen lévő hiány szempontjából nem releváns

tagmondat fogalmát. Vizsgáltuk, hogy a szerkezetek időbeli elrendezésének figyelembevétele, valamint a töltelékszavak, hezitációk, újraindítások, megakadások, iterációk hogy befolyásolják a tagmondatok számát, továbbá azt, hogy a kialakított új definícióink tükrében az egyszerű és összetett mondatok határsávja hol helyezkedjen el (halmozott mondatrészes mondatok esetén, a mondathoz lazán kapcsolódó részeket tartalmazó mondatok esetén, mondatszót és megszólítást tartalmazó mondatok esetén, igeneves szerkezeteket tartalmazó mondatok esetén, módosító mondatrészeket tartalmazó mondatok esetén, „mint"-es szerkezeteket tartalmazó mondatok esetén, bevezető szavakat és kifejezéseket, valamint előrevetett propozítumokat tartalmazó mondatok esetén) [1].

3 A szintaktikai struktúrák multimodális együttállásának problémafelvetése, annak indoklása és a kutatás célktűzései

A fentebb felvázolt hiányok felől megközelített szintaktikai struktúrák multimodális együttállásai által megszületett kutatási eredményeket és magát a kutatási folyamatot mutatom be. Multimodális és szintaktikai elemzési szempontokat alkalmazva azt feltételezzük, hogy a spontán beszélt nyelv mondattani szerkezeteit nem csupán szintaktikai eszközökkel lehet feltárni, hiszen a hiányzó mondatrészek kiegészülhetnek multimodális elemekkel szintaktikailag tágabb értelemben véve teljesebbé téve a mondatot.

A multimodális elemzés és az annotációs folyamat egységes strukturális szempontok alapján azt ígéri, hogy tükrözi a nyelv beszéd közben kialakuló szerkesztését, és lehetővé teszi a spontán beszéd szintaxisának többirányú kutatását.

Kutatásunk célja az, hogy a multimodális HuComTech-korpusz szintaktikailag annotált dialógusainak adatait felhasználva [2] összefüggéseket találjunk a szintaktikai elemek és a multimodális jegyek között. Jelen előadásban a multimodális jegyek közül a deiktikus gesztusok vizsgálat tesszük vizsgálatunk középpontjává. A deiktikus gesztusoknak azt a funkcióját figyeljük, amelyek az adott kontextusban rámutatásszerűen utalnak tárgyakra, személyekre vagy a kontextus egyéb vonatkozásaira, tényezőire. Ezek többnyire a helyet (itt, ott, innen, onnan), az időt (akkoriban, ma, mesélték, utánajárok), vagy a személyt (én, mi, veletek) jelölik. A hiányok elemzését a befejezetlen mondatok kommunikatív funkciókkal való összevetésével kezdtük [3]. Előadásomban a spontán beszélt nyelvi korpusz hiányos tagmondatai közül azokat a szintaktikai eseteket vizsgálom, amikor a tárgy hiányzik. A hipotézisünk tehát az, hogy a grammatikai szerkezetből hiányzó tárgyi mondatrész hiánya esetén a szintaktikai szerkezet kiegészülhet a multimodális rendszerben, jelen esetben a deiktikus gesztusok által, teljesebbé téve a mondatrészek között kialakuló szerkezetet. Jelenleg a kutatásnak a kezdeti szakaszában vagyunk, ezért az eredmények közlésén és elemzésén kívül kérdésfelvetésekkel mutatjuk meg a további kutatási irányokat ebben a témakörben.

4 Annotációs sémák bemutatása

A kidolgozásra került szintaktikai elemzési rendszert szinkronba hoztuk a HuComTech kutatócsoport által tervezett videó-annotáló program képi feldolgozásának annotációs sémájával. A kutatócsoport által alkalmazott angol nyelvű videóannotációs terminológia a következő:

1. táblázat: A HuComTech korpusz videóannotációs sémája

Group	event
Communcation	Start, end
Facial expression	Natural, happy, surprised, sad, recalling, tensed
gaze	Blink, orientation (up, down, left etc.)
eyebrows	Up, scowl
Head movement	Nod, shake, turn, sideways etc.
Hand shape	Open, half-open, fist, index-out, thumb-out, spread
touch motion	Tap, scratch
posture	Upright, lean, rotate, crossing arm, holding head, shoulder up
deictic	Addressee, self, shape, object, measure
emotions	Natural, happy, surprised, sad, recalling, tensed
emblems	Attention, agree, disagree, refusal, doubt, numbers etc.

Ebből a rendszerből a deiktikus gesztusok annotációs címkéit használtuk fel: addressee, self, shape, object, measure. Az addressee címkével ellátott gesztusok azt jelentik, hogy a riportban részt vevő személy a riporterre mutat, a self címke azt jelenti, hogy a riportalany önmagára mutat, az object címkével ellátott gesztusok olyan kézjelek, amelyek a riportszobában található tárgyakra mutatnak rá, a measure címke pedig azt jelenti, hogy a beszélő a levegőben egy tárgyat formáz.

A szintaktikai adatok kinyerése az ELAN program segítségével történt. A kutatási témánkhoz kapcsolódó szintjei és címkéi (fully aligned, overlap, left overlap, right overlap, surrounding, within, no overlap) lehetővé tették újabb szempontok beemelését a problémakörbe. A fully aligned funkció lehetővé teszi számunkra, hogy olyan gesztusokat keressünk, amelyek teljes egészében fedik az adott tagmondatot. Az overlap funkció abban segít, hogy egy tagmondatban megtaláljuk a különböző gesztusokkal való átfedéseket. Ennek a funkciónak az alfunkciója a left overlap és a right overlap. A left overlap funkcióban a gesztus a mondat felétől számítva balra kezdődik el, és átívelve a tagmondaton a következő tagmondatban fejeződik be. A right overlap funkcióban a gesztus a mondat felétől számítva jobbra kezdődik el, és tovább folytatódik, átnyúlva a következő tagmondatba. A surrounding azt jelenti, hogy a gesztus már tart a tagmondat megkezdésénél, és tovább ível a tagmondaton. A

within funkció pedig azt mutatja meg, amikor tagmondatnak a belsejében történik a gesztus, egy rövid szakaszon.

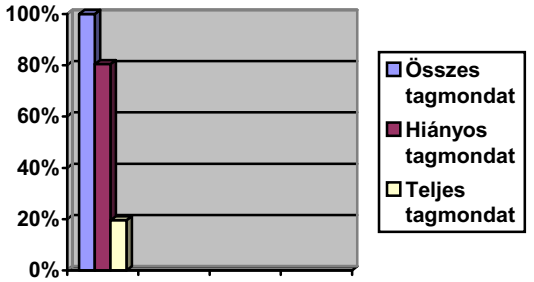
A HuComTech csoport által tervezett videóannotáló program, a szintaktikai kézi annotálás szabályai, valamint az ELAN program szolgáltatja számunkra az elméleti alapot és a terminológiai bázist.

De az *így/úgy* töltelékszavakat nem soroljuk ide, mivel grammatikailag (határozóként) kapcsolódnak a tagmondathoz.

5 A multimodális vizsgálat eredményei

Az adatbázis jelenlegi készültségi állapotában 1 132 165 annotáció van a különböző szinteken. Ebből 9435 annotáció vonatkozik a szintaktikai szintre, azaz 9435 mondat van szintaktikailag elemezve. Összesen 17 516 tagmondat van. 17 516 tagmondat biztosítja tehát a viszonyítási alapot a statisztikai adatokhoz és azok értékeléséhez.

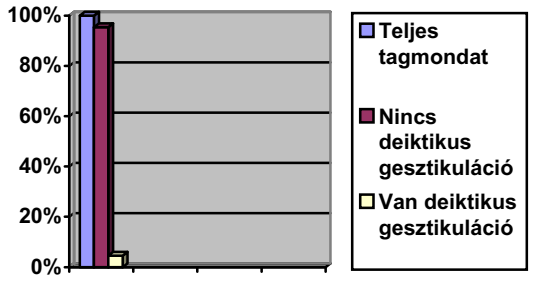
Az összes tagmondat 19,53 %- a, azaz 3422 a teljes tagmondat. Ebből következik, hogy a tagmondatok 80,47 %-ából, 14094 tagmondatból hiányzik valami:



1. ábra: A hiányos és teljes tagmondatok aránya az összes tagmondathoz képest

3422 teljes tagmondatból 156-ban található valamilyen deiktikus gesztikuláció (DeicticClass: addressee, self, shape, object, measure). Ez 4,55%-a a 3422 teljes tagmondatnak. A 3422 teljes tagmondat 95,45 %-ában, 3266 tagmondatban nem található egyetlen deictikus gesztikuláció (DeicticClas) sem.

A teljes tagmondatok több mint 95,45%-ában tehát nincs deiktikus gesztikuláció:



2. ábra: A deiktikus gesztikuláció jelenléte és hiánya a teljes tagmondatokban

3. ábra: A deiktikus gesztikuláció hiánya a tárgyhiányos tagmondatokban

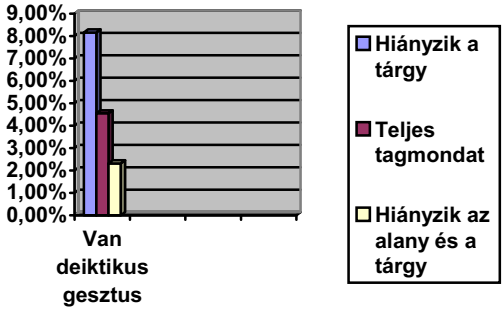
147 tárgyhiányos tagmondatból 12 tagmondatban található valamilyen deiktikus gesztus (DeicticClass: addressee, self, shape, object, measure), tehát a tárgyhiányos mondatoknak 8,16%-ában. 135 tárgyhiányos tagmondatban nincs deiktikus gesztus (DeicticClass), ami a tárgyhiányos mondatok 91,84 %-a.

Ha egy tagmondatból hiányzik a tárgy, akkor a várakozással ellentétben azt találtuk, hogy az esetek több mint 90%-ában nincs deiktikus gesztus sem. Arra számítottunk, hogy abban az esetben, ha hiányzik a tagmondatból a tárgy, a mondat grammatikailag kiegészülhet a kéz deiktikus gesztusaival, ezért a tárgyhiányos mondatok nagy részében deiktikus gesztusokat találhatunk. Feltevésünk ennek következtében fordítva sem teljesülhet, amely így hangzott: ha a kéz deiktikus módon gesztikulál, akkor hiányozhat a tagmondatból a tárgy.

Azt a feltevést is megfogalmaztuk, hogy ha nem hiányzik a tagmondatból a tárgy, és a mondat teljes, kevesebb gesztikulációt tapasztalunk, mint amikor fellép a tárgy hiánya, hiszen ez bizonyíthatja azt, hogy a tárgy grammatikus hiánya kiegészülhet valamelyik deiktikus gesztikulációval: a tárgyhiányos tagmondat jelentősen több gesztikulációra számítottunk, mint a teljes mondatban. Ez a feltevésünk

beigazolódott, hiszen a teljes tagmondatoknak 4,55%-ában található valamilyen deiktikus gesztikuláció, a tárgyhiányos mondatoknak pedig majdnem duplája, 8,16%-a tartalmazza azt.

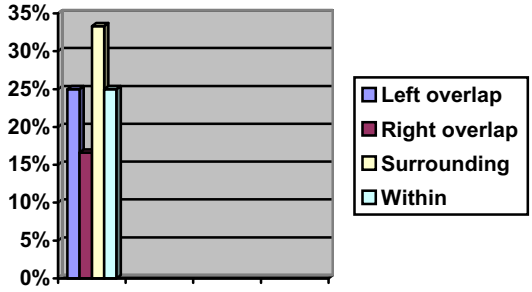
Ha hiányzik a tárgy, akkor 3,6%-kal több deiktikus gesztikuláció történik, mint teljes mondatok esetén. Legkevesebb deiktikus gesztus akkor történik, amikor az alany és a tárgy egyszerre hiányzik. Tárgyhiányos tagmondatokban valamivel kevesebb deiktikus gesztikuláció történik, mint a teljes tagmondatokban:



4. ábra: A deiktikus gesztus arányai a teljes, a tárgyhiányos és az alany-tárgy hiányos tagmondatok esetén

Ennek kapcsán feltettük azt a kérdést, hogy mi lehet az oka annak, hogy jelentősen több deiktikus gesztikuláció fordul elő a tárgyhiányos mondatokban, mint a teljes tagmondatokban, mégis csupán 8,16%-a tartalmaz legalább egy deiktikus gesztust a tárgyhiányos tagmondatoknak. A kérdésfelvetést azzal a megfigyeléssel pontosítottuk, hogy a teljes mondatokból vagy azért hiányzik a tárgy, mert az ige vonzatköre meg sem kívánja azt, vagy azért, mert az ige tárgyi vonzata grammatikusan megjelenik a mondatban. Jóval több azoknak az igéknek a száma, amelyek a teljes mondatokban előfordulhatnak, mint azoknak, amelyek a tárgyhiányos mondatokban szerepelhetnek. Abban az esetben van jelölve egy tagmondaton, hogy hiányzik belőle a tárgy, ha az ige vonzatköre megkívánná, de valamilyen okból nem jelenik meg lexikálisan a tagmondatban. Ha az ige vonzatköréhez hozzátartozik a tárgy, abban az esetben valóban több deiktikus gesztikuláció történik a tagmondatban.

Közelebb juthatunk a válaszhoz, ha megvizsgáljuk, hogy pontosan milyen típusú deiktikus elemek hiányoznak a tárgyhiányos tagmondatokból, és azok a tagmondatnak melyik részén helyezkednek el:



5. ábra: A gesztusok jelenléte a tagmondatokban

147 tárgyhiányos mondatból 12 tartalmaz deiktikus gesztust, ebből 3 tagmondatban a gesztus a mondat felétől számítva balra kezdődik el, és átívelve a tagmondaton a következő tagmondatban fejeződik be (left overlap), 2 tagmondatban a gesztus a mondat felétől számítva jobbra kezdődik el, és tovább folytatódik, átnyúlva a következő tagmondatba (right overlap), 4 tagmondatban a gesztus már tart a tagmondat megkezdésénél, és tovább ível a tagmondaton (surrounding), és 3 tagmondatnak a belsejében történik a gesztus, egy rövid szakaszon (within). Ebből az látszik, hogy ha a gesztus átível a tagmondaton, ha előbb kezdődik el, illetve ha később fejeződik be, akkor vagy nem követi pontosan a mondat szerkezetét, vagy olyan tárgyat kísérhet, amelyik több tagmondaton át jelen van a kommunikációban, és megjelenik az igék vonzatkörében is. Előfordulhat az is, hogy a mondat átkötésében van funkciója egy-egy gesztusnak, így a kötőszó mint szintaktikai elem kapcsolatban állhat a gesztusrendszerrel.

Továbbá feltételezhetjük azt is, hogy az alárendelő tagmondatok esetén a főmondat alatt létrejött gesztus átívelése az alárendelő tagmondatba utalhat arra, hogy a főmondatban létrejött hiány az alárendelésben kerül kifejtésre, s ezt a grammatikai kapcsolatot egy gesztikuláció is kísérheti, mivel a tárgyi alárendelés esetén a főmondatból hiányzó tárgy kiegészül a mellékmondatban.

Az összetett mondatokat vizsgálva megállapítottuk, hogy 4 olyan tagmondat van, amelyben hiányzik a tárgy, mellérendelő tagmondat, s egyben a deiktikus gesztust is tartalmaz. Ez a deiktikus gesztust tartalmazó összes mondatnak csupán 0,06 %-a. 1 olyan tárgyhiányos tagmondat van, aminek alárendeltje van, és deiktikus gesztus is található a mondatban. Ez a deiktikus gesztust tartalmazó mondatoknak 0,05%-a. 1 olyan tárgyhiányos tagmondat van, ami alá van rendelve egy másik tagmondatnak, hiányzik belőle a tárgy, és deiktikus gesztust tartalmaz, ami az összes deiktikus gesztikulációt tartalmazó tagmondatoknak 0,05%-a. Ezért a deiktikus gesztusok mint mondatrészt pótló elemek nincsenek jelen számottevő mértékben az összetett tagmondatokban, hogy vizsgálható legyen a deiktikus gesztus és a főmondat hiányának mellékmondatban való kiegészülése. Végső megállapítást akkor tehetünk majd, ha az összes gesztust megvizsgáljuk ebből a szempontból. Olyan kérdéseket tehetünk majd fel az összes gesztus értelmezése közben, hogy ha az alárendelő mondat főmondatából hiányzik a tárgy, akkor biztos-e minden esetben az, hogy az

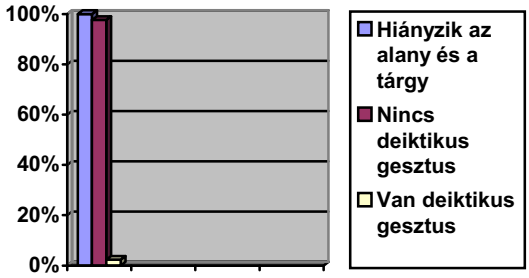
alárendelése tárgyi alárendelés, és ha nem, akkor mi a különbség ebben az esetben a jelenséget kísérő gesztusrendszerben; valamint a főmondatból hiányzó tárgy kiegészül-e minden esetben az alárendelt tagmondatban a szintaktikai elemek vagy gesztusok segítségével, és ha nem, annak van-e valamilyen gesztikulációs következménye.

Az adatok alapján az is látható, hogy a deiktikus gesztusok időbeli elhelyezkedése a mondaton belül és a magyar szórend nem függnek össze egymással. A tagmondatnak nem azon a részén történik a gesztikuláció, ahol a szerkezeti hiányt megállapíthatjuk, s így a hiány feltételezett helye és a gesztikuláció nem fedik egymást az időben. További kutatási célunk az, hogy a hiány helyének lokalizálhatóságát behatároljuk, illetve ha nincs ilyen, akkor ezt a megállapítást kimondhassuk.

Olyan deiktikus gesztusokat érdemes tehát keresni, amelynek nem csupán kommunikatív és pragmatikai funkciója van, hanem egy-egy szóhoz köthető annak jelenléte. Ez bizonyíthatná azt, hogy a gesztus valóban egy adott mondatrészhez kapcsolható. Ennek feltárásához nem csupán a deiktikus gesztusokat kell figyelni, hanem az összes gesztust. Lehet-e olyan megállapítást tenni egyes tagmondatokban, hogy annyi mondatrész van, amennyi gesztus zajlott a tagmondatban? Ha csak egy gesztus van az adott tagmondatban, akkor egyetlen mondatész hangzott el, vagy egyetlen mondatrész hiányzik? Van- e olyan mondatrész, amihez akár több gesztus is társul általában? Az, hogy mi számít egyetlen gesztusnak, azaz hol vannak a gesztusok határai, a videoannotációs elemzés kapcsán eldöntésre került, ezért azt elfogadva nem tartjuk problémának, hogy hol van két gesztus között a határ, határsáv.

A lexikális elemekhez köthető gesztusok vizsgálatakor a következő kérdéseket gondoltuk át: Ha egy mondatban több gesztus jelenik meg, akkor megállapítható-e, hogy melyik vonatkozik egy adott mondatrészre. Honnan tudhatjuk, hogy melyik gesztus hiányzik, tehát lehet-e olyan megállapítást tenni, hogy ha például nincs deiktikus gesztus, akkor hiányzik az a mondatrész, amely ehhez a gesztushoz köthető? Van-e mondatrész-specifikus multimodális tartomány, azaz van-e olyan mondatrész, amit dominánsan kísér egy adott gesztus? Ezek a kérdések a következő mondattípusból alakultak ki: *„Add ide a …”* A beszélő nem mondja ki, de rámutat arra a tárgyra, ami hiányzik a tagmondatból.

646 olyan tagmondat van, amelyből egyszerre hiányzik az alany és a tárgy. Ez 3,68 %-a az összes tagmondatnak (17516). 646 alany- és tárgyhiányos tagmondatból 15 tagmondatban van deiktikus gesztikuláció (DeicticClass), ami 2,32 %. Összesen 631 olyan alany- és tárgyhányos tagmondat van, amelyben egyáltalán nincs deiktikus gesztikuláció (DeicticClass), ez pedig 97,68%. 17516 tagmondatból 147 esetben hiányzik a tárgy, ami a hiányos tagmondatok 0,83 %-a. Az alany és a tárgy együtt 646 mondatból hiányzik, ami a hiányos mondatoknak (14094) összesen 4,58 %-a:



6. ábra: A deiktikus gesztusok jelenléte és hiánya az alany-tárgy hiányos tagmondatok esetén

Megállapíthatjuk, hogy több esetben hiányzik a tárgy az alannyal együtt, mint csak a tárgy önmagában, hiszen az alany és a tárgy együtt 646 esetben hiányzik, a tárgy önmagában viszont csak 147-szer. Ebből a spontán nyelvben előforduló alanyi és tárgyas ragozású igék arányára következtethetünk. A hiányok együttes jelenlétében nem találtunk szabályszerűséget és gyakoriságot, mivel gyakori együttállásokat nem tudtunk szabályszerűsíteni. Az állítmány (igei és névszói-igei állítmány) és a tárgy együtt például csupán 4 alkalommal hiányzik, ez a hiányos tagmondatoknak csupán 0,028%-a. A tárgy és a határozó együtt 8 alkalommal szerepel. Ez 0,056%-a a hiányos mondatoknak.

2594 olyan eset van, amikor egymást követi két egyszerű mondat. 6841 olyan mondat van, ami egy tagmondatból áll. Ez 39,05%-a az összes tagmondatnak (17516), és 72,5%-a az összes mondatnak (9435). 371 két tagmondatból álló mondat van, ami 3,93%-a az összes mondatnak. 187 három tagmondatból álló mondat van, ami 1,98%-a az összes mondatnak. 118 olyan mondat van, ami 4 tagmondatból áll, ez 1,25%-a az összes mondatnak. 53 olyan mondat van, ami 5 tagmondatból áll, ez 0,56%-a az összes mondatnak. Táblázatosan ezt így láthatjuk:

2. táblázat: Az összetett mondatok százalékos adatai

Tagmondatok száma	Előfordulások száma	Hány százaléka ez az összes mondatnak
1	6841	72,5%
2	371	3,93%
3	187	1,98%
4	118	1,25%
5	53	0,56%
6	46	0,48%
7	26	0,27%
8	26	0,27%
9	17	0,18%
10	7	0,07%

Azt láthatjuk tehát, hogy a legtöbb esetben egyszerű mondatokat használunk. Egyszerű és összetett mondatokat vizsgálva olyan kérdések merülnek föl, hogy ha egyszerű mondat esetén hiányzik a tárgy, akkor a körülötte lévő tagmondatokból hiányzik-e a tárgy, jelzi-e gesztikuláció, hogy a tárgy már korábban ki lett mondva, illetve később lesz kimondva. Összetett mondatok esetében is felmerül a kérdés, hogy valamelyik tagmondatban ki van-e mondva a tárgy, illetve jelzi-e gesztus, hogy a tárgy már korábban ki lett mondva, illetve később lesz kimondva. A kontextusban ki lett-e mondva, ki lesz-e mondva egyáltalán a hiányzó mondatrész, és ha igen, hány mondatrésszel korábban vagy később. Megfigyelés tárgyává válhat, hogy van-e különbség a két problémakör eredményei között, és mi a különbség oka.

6 Következtetés és összegzés

Annak, hogy nincs pontos időbeli egyezés a gesztusok és valamely hiányzó elem között, annak az is lehet az oka, hogy a gesztusok valószínűleg megelőzik az adott szintaktikai egységet, a nyelvi funkció és a gesztusrendszer működése aszinkronban van egymással. Ennek pedig az lehet az oka, hogy a gesztusok inkább a mondatalkotás feltételét, a kognitív tervezést kísérik, az pedig megelőzi a gondolat nyelvi formában való kifejezését[1].Ez magyarázhatja azt is, hogy miért nyúlnak át a tagmondathatáron egyes gesztusok: a tervezés hosszabb távú, mint magának a gondolatnak egy-egy adott, lebontott szintaktikai egysége. Valószínűsíthetjük, hogy ha a tervezés és a kimondott mondatrész időben elcsúszik, akkor lehet, hogy az előző tagmondat gesztusai már a következő tagmondat mondatrészét tartalmazzák Ha a tervezés csupán egy szónyi, akkor valószínűsíthetjük, hogy a tervezést követően rövidebb időn belül megjelenik a kimondott szó, mint ha mondat nagyságú tervezésről van szó.

Ennek tükrében a kérdéseink úgy módosulnak, hogy a tervezés és a megszületŏ nyelvi forma között lehet-e kapcsolatot feltételezni, azaz a tervezésnek vannak-e olyan gesztusai, amelyeket egy adott mondatrészhez tudunk kapcsolni. Ezért a vártnál kevésbé fontos az, hogy a gesztus hol helyezkedik el a mondatrészekhez képest, inkább a gesztusok karaktere, jellemző adottsága válik hangsúlyossá. Mindez felveti azt a kérdést is, hogy ez a jelenség mennyire egyénspecifikus, valamint azt is, hogy a tanult gesztusok, szemben az ösztönös gesztusokkal mennyire befolyásolják akár a tervezést, akár a mondatrészek kimondását, és fordítva, a tervezés mennyire befolyásolja azt, hogy a gesztusok tanult vagy ösztönös formában jelennek meg. Előfordulhat-e az, hogy a gesztus a tervezés szakaszában jön létre, tehát egyidejűség lép fel, vagy pedig a gesztus a tervezéshez képest mindig utóidejű.

További célunk az, hogy a felmerülő kérdésekre az alapkoncepciónknak megfelelően az egész gesztusrendszert kutatva megfelelő válaszokat kapjunk. Tervezzük azt is, hogy a gesztusokat külön-külön is vizsgálat tárgyává tesszük, és az együttállásokat is figyeljük. Összességében már most elmondhatjuk, hogy a nem nyelvi módon kifejezett szintaktikai elemeknek a multimodális ember-gép kommunikáció beszédfeldolgozásában betöltött szerepét leginkább azzal lehet

[1] Prof. Dr. Hunyadi László témavezető személyes közlése.

hangsúlyozni, hogy figyelembe vesszük, hogy a grammatikai információ és a modalitások együttes jelenléte valóban összefügg egymással, mert mindez együtt szükséges ahhoz, hogy a közvetített információ a birtokunkba kerülhessen.

Hivatkozások

1. Keszler B.: Mondattan. In: Keszler B. (szerk.): Magyar Grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 461—471
2. Pápay K., Szeghalmy Sz., Szekrényes I.: HuComTech Multimodal Corpus Annotation. Argumentum 7. Debreceni Egyetemi Kiadó Kiadó, Debrecen (2011) 330—347
3. Hunyadi L.: Possible communicative cues to syntactic incompleteness in spoken dialogues. Argumentum 9 (2013)
4. Kiss H.: A HuComTech audio adatbázis szintaktikai szintjének elvei és szabályrendszerének újdonságai. In: MSzNy 2011 (2011) 199—216

II. Morfológia, szintaxis

HuLaPos2 – Fordítsunk morfológiát

Laki László[1,2], Orosz György[1,2]

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

Kivonat Jelen munkánkban bemutatunk egy gépi fordításon alapuló nyelvfüggetlen teljes morfológiai egyértelműsítő rendszert, ami egyidejűleg végzi a szótövesítést és a morfológiai egyértelműsítést. Annak érdekében, hogy demonstráljuk a módszer hatékonyságát, több különböző nyelv legjobb rendszerével hasonlítottuk össze. A legtöbb nyelv esetén rendszerünk jobban teljesít szófaji egyértelműsítés tekintetében, valamint a szótövesítés pontossága hasonló az általunk összehasonlított rendszerekével.

1. Bevezetés

A nyelvtechnológiai feldolgozási lánc fontos elemei a morfológiai elemzés és egyértelműsítés. Az utóbbi komponens feladata, hogy egyértelműen meghatározza a szavak szótövét, és megállapítsa azok morfoszintaktikai (PoS) címkéit. Az első, erre a célra létrehozott eszközök angol nyelvű szövegek elemzésére szolgáltak, melyek azonban egymást követően végezték a PoS címkézést és a szótövesítést. Így az ezek alapján létrehozott újabb rendszerek is ezt a sémát követték. Következésképp kevés olyan eszköz létezik, amely teljes morfológiai egyértelműsítést végez, ami elengedhetetlen morfológiailag gazdag nyelvek elemzése esetén. Továbbá csak néhány olyan eljárás létezik, amely grammatikailag nagyon különböző nyelvek esetében is ugyanolyan magas pontossággal képes működni. Bár az egyes nyelvspecifikus eszközök sokszor magas pontosságot produkálnak, de a legtöbbször csak egy-egy nyelv nagy teljesítményű elemzésére korlátozódik a tudásuk.

A tanulmány célja egy Moses SMT[3] rendszeren alapuló nyelvfüggetlen morfológiai elemző rendszer bemutatása, amely különböző típusú nyelvek esetén végez teljes morfológiai egyértelműsítést úgy, hogy pontossága felveszi a versenyt a nyelvfüggő társai eredményeivel.

Dolgozatunk első részében ismertetjük a létrehozott rendszer (HuLaPos2) felépítését, majd bemutatjuk az általa elért eredményeket összehasonlítva azokat hat különböző nyelv state-of-the-art egyértelműsítő eredményeivel.

[3] Statisztikai gépi fordító

2. Kapcsolódó munkák

Az első általánosan elterjedt statisztikai taggerek rejtett Markov-modellen alapultak, úgymint a TnT [1] vagy a HunPos [2]. Ezzel párhuzamosan Ratnaparkhi [3] bemutatott egy maximum entrópián alapuló megközelítést, amit számos nyelv esetében sikerrel alkalmaztak (pl. a Stanford tagger [4] különböző adaptációi, vagy a `magyarlanc` [5]). Ezeken kívül számos más felügyelt tanulásos módszer is jól teljesít különböző nyelvek esetében: úgymint Brill transzformációalapú módszere [6], az SVMTool [7] Support Vector Machine alapú modellje, vagy a TreeTagger [8] döntési fákat használó algoritmusa.

Mora és Sánchez [9] voltak az elsők, akik SMT módszert használtak szófaji egyértelműsítésre, de ők a rendszert csak az angol nyelv PoS taggelésére tervezték, lemmatizálásra nem. Munkájukban a tanítóanyagban nem előforduló szavak (OOV) kezelésére egy szógyakoriságon alapuló modellt és egy 11 elemből álló szuffixum listát alkalmaztak.

Hasonló megközelítést használtunk egy korábbi munkánkban [10], ahol a fenti metódust magyar nyelvre alkalmaztuk. A Mora és Sánchez által angol nyelvre optimalizált algoritmus jelentős mértékben alulmaradt a legjobb magyar elemzőkhöz képest (pl. a morfológiai elemzővel kiegészített PurePos [11]). Ez többek között azzal is magyarázható, hogy a magyar nyelv agglutináló tulajdonságaiból adódóan fejlettebb módszerek szükségesek a jelentős számú OOV tokenek kezelésére. Ebben a tanulmányban a Laki-rendszer továbbfejlesztett változatát mutatjuk be.

3. Elméleti háttér

3.1. Kifejezésalapú statisztikai gépi fordítás

A gépi fordítórendszer leképezést biztosít két nyelv között függetlenül attól, hogy ezek természetes vagy mesterséges nyelvek. A statisztikai gépi fordító algoritmusok párhuzamos kétnyelvű korpuszokból gépi tanulásos módszerek segítségével tanulják meg a transzformációhoz szükséges modelleket.

Ha W egy mondat a forrásnyelvi szövegből, melynek a helyes fordítása $\hat{T}$, akkor a fordítási feladat a következőképpen formalizálható:

$$\hat{T} = \underset{T}{argmax}\ P(T|W) = \underset{T}{argmax}\ P(W|T)P(T) \tag{1}$$

ahol $P(T)$ a nyelvi modell és $P(W|T)$ fordítási modell. Míg az első modell a lefordított szöveg olvashatóságára (folyékonyságára) ad becslést, addig a második modell a fordítás minőségét becsüli. A statisztikai gépi fordítás egyik gyakran használt változata a kifejezésalapú fordítás, melynek alapja, hogy a fordítandó W mondatot kifejezésekre bontjuk $W = w_1 w_2 \dots w_N = w_0^N$, amiket különkülön lefordítunk. A lefordított részek legjobb kombinációját véve kapjuk a célnyelvi mondatot ($T = t_0^N$). A kifejezések fordítását a párhuzamos tanító anyagból számolt $\phi(w_i^{i+k_1}|t_i^{i+k_2})$ valószínűségi eloszlás alapján végzi a rendszer. Ezek használatával a (1) a következőképpen fejthető ki:

$$\underset{T}{argmax}\ P(W|T)P(T) = \underset{T}{argmax}\left[\prod_{i=0}^{N}\ \phi(w_i^{i+k_1}|t_i^{i+k_2})P(t_i|t_{i-1}^{i-j})\right] \qquad (2)$$

3.2. Morfológiai egyértelműsítés mint gépi fordítási feladat

A szófaji címkézés feladatára számos módszer létezik, melyek közül a legelterjedtebbek a rejtett Markov-modellezésen (HMM) alapulók. Ennek működése a következőképpen (vö. (3)) írható le formálisan: ha W az elemzendő szöveg egy mondata, mely helyes elemzésének címkesorozata $\hat{T}$, akkor ennek valószínűsége maximális a címkeátmenet-modell $P(T)$ és a lexikai-modell $P(W|T)$ szorzatát tekintve. A legtöbb rendszer (így pl. a TnT és a HunPos is) az első valószínűségi értéket egy másodrendű modellel becsli, ami lényegében egy címkékre épülő trigram modell: $P(t_i|t_{i-1}^{i-2})$. A lexikai-modell becslésére pedig legtöbbször maximum likelihood becslést alkalmaznak, ami a szavakhoz rendelt morfoszintaktikai címkék relatív gyakoriságából tevődik össze: $P(w_i|t_i)$.

$$\hat{T} = \underset{T}{argmax}\ P(W|T)P(T) = \underset{T}{argmax}\left[\prod_{i=0}^{N}\ P(w_i|t_i)P(t_i|t_{i-1}^{i-2})\right] \qquad (3)$$

Összevetve a (1) és (3) egyenleteket láthatjuk, hogy a statisztikai gépi fordítás feladata könnyen megfeleltethető a morfológiai címkézés HMM módszerének. A megfeleltetés lépései: az SMT nyelvi modellje a címkeátmenet-valószínűség modell, míg a fordítási modell a lexikai modellnek felelnek meg. A leképezésen túl az is megfigyelhető még, hogy az SMT-n alapú megközelítés egy általánosabb keretrendszert biztosít a feladat megoldására

Motivációnk a nyílt forráskódú Moses SMT toolkit [12] keretrendszert használatára a következők voltak:

1. A Moses tanítási lánca gyors a valószínűségi modellek létrehozását illetően.
2. A leggyakrabban alkalmazott HMM alapú elemzőkkel szemben a Moses rendszer által létrehozott fordítási modell nemcsak egy-egy szó lehetséges elemzését tartalmazza, hanem a hosszabb kifejezésekéit is, ami lehetővé teszi az elemző számára, hogy a szöveg hosszabb részeit is egy egységként kezelje.
3. A címkeátmenet-valószínűség modell (a nyelvmodell) építése során nemcsak az azt megelőző két szó elemzését veszi figyelembe, hanem akár a mondatban szereplő összes megelőzőét, valamint a létező egyik legjobb simító algoritmust, a módosított Kneser-Ney simítást [13] használja.
4. A dekóder a beam-search algoritmus egy hatékony és gyors változatát az úgynevezett verem dekódolást alkalmazza. A módszer legnagyobb előnye, hogy az elemzést a dekódoló működésének köszönhetően a szavak tetszőleges sorrendjében végezheti, szemben a HMM-alapú elemzők szigorúan balról jobbra történő működésével.
5. A dekódolás folyamatába egyszerűen integrálható morfológiai guesser vagy elemző.

4. A rendszer bemutatása

Ebben a fejezetben áttekintjük azokat a legfontosabb módosításokat, amelyek megkülönböztetik az eredeti SMT rendszert a morfológiai egyértelműsítőtől (egy részletesebb leírás a [14] cikkünkben olvasható).

A szuffixumokat használó ragozó nyelvek esetén (mint például a magyar vagy a török) a szótövek egyszerűen leírhatók olyan rekordokkal, melyek megadják azt a szükséges transzformációt, amit el kell végezni egy adott szón, hogy megkapjuk annak szótövét. Egy ilyen rekord: $\langle cut, paste \rangle$, ahol a *cut* a sztringről eltávolítandó karakterek számát adja meg, a *paste* pedig az a karaktersorozat, amit illeszteni kell a „csonka szó" végére, hogy megkapjuk a szótövet. Ezt az ötletet használva az elemzőnk a morfoszintaktikai címkék mellett képes még reprezentálni a szótöveket is.

Másrészt természetes nyelvek esetében az SMT rendszer szóösszekötője gépi tanulásos algoritmusokat használ a fordítási frázispárok meghatározásához. Ez a mi esetünkben a feladat felesleges bonyolítása, mivel a morfológiai egyértelműsítéshez egy egyértelmű monoton megfeleltetésre van szükség, mely a tokeneket az elemzéseikhez rendeli. Ezért a HuLaPos2 rendszerben a Giza++ algoritmust monoton leképezéssel helyettesítettük.

Harmadrészt, a Moses dekóder legnagyobb előnye, hogy hosszabb kifejezéseket is képes egy egységként fordítani, de itt a frázisok maximális hossza és a nyelvi modell mérete nagyban befolyásolja a rendszer minőségét. Ezért szükséges ezen paramétereinek finomhangolása, amihez az optimális beállításokat – minden nyelvre külön-külön – empirikusan határoztuk meg.

Végül az adathiány által okozott problémák elkerülése érdekében a számjegyek generikus szimbólumokkal lettek helyettesítve a tanítóhalmazban és a bemeneti szövegben egyaránt. Az SMT rendszer legnagyobb hiányossága, hogy a tanítóhalmazban nem szereplő szavakat figyelmen kívül hagyja, és semmilyen elemzést sem ad hozzájuk. Ennek kiküszöbölésére rendszerünkbe – a PurePos és HunPos rendszerekhez hasonlóan – egy trie-alapú suffix-guessert építettünk, amely elemzési javaslatokat ad az OOV szavakra. Ez az algoritmus a tanítóhalmazban ritkán előforduló szavak végződései alapján képes megbecsülni, hogy egy szó az egyes *(szótő-transzformáció; címke)* elemzésekkel milyen valószínűséggel címkézhető. Ennek a módszernek további előnye, hogy az elemzések valószínűségének számítása – a TnT-hez hasonlóan – különböző hosszúságú toldalékok simított interpolált modellje alapján történik. Ráadásul ez az algoritmus megoldást nyújt az SMT rendszer azon gyengeségére, miszerint az OOV szavakat tartalmazó szegmensek elemzése során a dekódoló csak az unigram modelleket használhatja. Mivel ez a modul arra hivatott, hogy a ritkán előforduló szavakat kezelje, ezért ilyen tulajdonságú szavakon kell betanítani. A ritka szavak esetén a használt küszöbértéket empirikusan határoztuk meg: a legmagasabb pontosságot általában akkor értük el, amikor ez az érték 2 volt, azaz a guesser csak hapaxokon volt tanítva. A javasló komponens a következő módon lett a dekódolóba integrálva: A Moses képes a kifejezések fordítása közben előre definiált fordítási javaslatokat is figyelembe venni. Ezzel az egyszerű módszerrel a tanítóhalmazban nem szereplő szavakhoz hozzárendeljük a guesser javaslatait, mint előfordítás.

5. Eredmények

A HuLaPos2 rendszert több különböző nyelvhez (magyar, szerb, horvát, bolgár, portugál és angol) elérhető legjobb pontossággal teljesítő egyértelműsítő rendszerekkel hasonlítottuk össze. A tanító- és a teszthalmazt a kapcsolódó publikációkban leírt módon (részletesen lentebb) definiáltuk. A rendszerek pontosságának részletes összehasonlítását a 1-es és 2-es táblázatokban foglaltuk össze, ahol az első táblázatba gyűjtöttük össze azokat a rendszereket, amelyek teljes morfológiai egyértelműsítést csinálnak, míg a második táblázatban szereplők csak morfológiai egyértelműsítést végeznek.

1. táblázat. A HuLaPos2 rendszer minőségének összehasonlítása más rendszerekével a szófaji egyértelműsítés, szótövesítés, valamint a teljes morfológiai egyértelműsítés tekintetében

Nyelv	Rendszer	Szószíntű pontosság		
		címkézés	szótövesítés	teljes
magyar (MSD)	HuLaPos2	**99,57%**	**97,24%**	96,84%
	PurePos	96,74%	96,35%	94,76%
magyar (HUMor)	HuLaPos2	**99,18%**	98,23%	97,62%
	PurePos	96,50%	96,27%	94,53%
	PurePos+MA	98,96%	**99,53%**	98,77%
horvát	HuLaPos2	**93,25%**	96,21%	90,77%
	HunPos+CST	87,11%	**97,78%**	–
szerb	HuLaPos2	**92,28%**	92,72%	86,51%
	HunPos+CST	85,00%	**95,95%**	–

Magyar nyelv esetében a legjobb egyértelműsítő rendszer a PurePos [11], ami egy HMM-alapú teljes morfológiai egyértelműsítő, melybe morfológiai elemző van integrálva. Az eredmények összehasonlításához a Szeged Korpuszt [15] választottuk, melynek két változatán teszteltük rendszerünket: az eredeti MSD-kódolással készültet, és egy HuMor [16] címkékre automatikusan átírtat. A HuLaPos2 rendszert a PurePos rendszer morfológiai elemzőt használó, valamint anélkül működő (tehát nyelvfüggetlen) változataival hasonlítottuk össze. Az eredmények megmutatták, hogy a HuLaPos2 az összes mért esetben jobb eredményt ért el a PurePos morfológiai elemző nélküli változatával szemben, és szófaji címkézés esetén pontossága meghaladja a PurePos morfológiai elemzős változatát.

Szerb és horvát nyelvre Agić és munkatársai [17] készítettek szófaji címkéző és szótövesítő alkalmazást 2013-ban. A rendszert a HunPos és a CST szótövesítő [18] kombinációjából építették fel, és a SETimes.HR [17] korpuszon tanították. Az 1. táblázat eredményeiből látható, hogy PoS taggelés esetén a HuLaPos2 teljesítménye szignifikánsan meghaladja Agićék rendszerét, míg a szótövesítésben elért eredmény is közelít annak eredményességéhez. A különbség a javasló algo-

ritmus működéséből ered: a CST rendszerben a szótő-transzformációk nemcsak szuffixumok lehetnek, hanem a tetszőleges helyű változások is. Ezzel szemben a HuLaPos2 által használt guesser csak a szóvégi változást képes kezelni.

Georgi Georgiev és munkatársai [19] létrehoztak egy morfológiai lexikonnal és nyelvtani szabályokkal kiegészített irányított tanuláson alapuló szófaji egyértelműsítő rendszert bolgár nyelvre. Eszközüket a BulTreeBank korpuszon [20] tanították és tesztelték. A 2. táblázat eredményeiből látható, hogy a HuLaPos2 teljesítménye nagymértékben meghaladja a nyelvtani tudással nem rendelkező tisztán statisztikai módszereket használó rendszerek minőségét. Annak ellenére, hogy rendszerünk semmilyen nyelvspecifikus eszközzel nincs támogatva, jobban teljesít, mint a morfológiai lexikont használó eszköz, valamint pontossága megközelíti Georgiev által készített legjobb rendszerét (irányított tanulás + lexikon + szabályok).

2. táblázat. A HuLaPos2 rendszer minőségének összehasonlítása olyan rendszerekkel, amelyek csak szófaji egyértelműsítést csinálnak

Nyelv	Rendszer	Címkézés pontossága
bolgár	TnT	92,53%
	gépi tanulás	95,72%
	gépi tanulás + morf. lexikon	97,83%
	HuLaPos2	**97,86%**
	gépi tanulás + morf. lexikon + szabályok	97,98%
portugál	**HuLaPos2**	**93,20%**
	HMM-alapú PoS tagger	92,00%
angol	TnT	96,46%
	PBT (Mora and Sánchez [9])	96,97%
	HuLaPos2	**97,08%**
	Stanford tagger 2.0	97,32%
	SCCN [21]	97,50%

A HuLaPos2 rendszert teszteltük még morfológailag egyszerűbb nyelvek esetében is, mint a portugál és az angol. Mindkét esetben csak a PoS tagger eredményességét tudtuk összehasonlítani (2. táblázat), mivel az elérhető korpuszok nem tartalmazták a szavak lemmáit.

Portugál nyelvre a Maia és Xexéo [22] által 2011-ben készített HMM-alapú rendszert vettük összehasonlítási alapul. Ez az eszköz a Floresta Sintá(c)tica Treebank-en [23] lett tanítva, melyből az első 10% volt a teszthalmaz, a fennmaradó 90% pedig a tanító halmaz. Ugyanezekkel a beállításokkal a HuLaPos2 pontossága több mint 1%-kal felülmúlta a portugál címkéző eredményeit.

Ami az angol nyelvet illeti, a Penn Treebank [24] WSJ korpuszát használtuk az általánosan bevált elosztásban.[4]

[4] http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)

A 2. táblázat a HuLaPos2 és a másik négy rendszer által elért eredményeket mutatja. Megfigyelhető, hogy a HuLaPos2 meghaladja a TnT és a Mora és Sánchez-féle [9] rendszerek által elért értékeket. Az eredmények vizsgálatánál fontos még figyelembe venni, hogy algoritmusunk a tanítóanyagon kívül semmilyen más lexikai adatbázist, vagy előzetes tudást nem használ, így elmondható, hogy annak teljesítménye a maga nemében kiemelkedő.

6. Konklúzió

Írásunkban bemutattunk egy, a Moses keretrendszeren alapuló, nyelvfüggetlen teljes morfológiai egyértelműsítő rendszert. Ez az eszköz egyidejűleg végzi a szófaji egyértelműsítést és a szótövesítés feladatát egy trie-alapú suffix-guesser segítségével, amely hatékonyan kezeli a morfológiailag gazdag nyelvekre jellemző OOV szavak problémáját. A HuLaPos2 hat különböző nyelv legjobb rendszerével lett összehasonlítva. Szófaji egyértelműsítés tekintetében rendszerünk (az angol nyelv kivételével) jobb eredményt ér el a vizsgált taggerekhez képest. Mindemellett szótövesítés esetén is versenyképesnek bizonyult a nyelvfüggő vetélytársakkal szemben. Az angol nyelv esetén a HuLaPos2 meghaladja a közismert TnT rendszer eredményeit, valamint megközelíti az elérhető legjobb rendszer minőségét.

Köszönetnyilvánítás

Ez a projekt a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014. támogatásával készült.

Hivatkozások

1. Brants, T.: Tnt - a Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA (2000)
2. Halácsy, P., Kornai, A., Oravecz, C.: HunPos: An open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL, Stroudsburg, Association for Computational Linguistics (2007) 209–212
3. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the fifth conference on Applied natural language processing. ANLC '97, Stroudsburg, PA, USA, Association for Computational Linguistics (1997) 16–19
4. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13. EMNLP '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 63–70
5. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 275–283

6. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics **21** (1995) 543–565
7. Giménez, J., Màrquez, L.: SVMTool: A general POS tagger generator based on Support Vector Machines. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation. (2004) 43–46
8. Schmid, H.: Improvements In Part-of-Speech Tagging With an Application To German. In: In Proceedings of the ACL SIGDAT-Workshop. (1995) 47–50
9. Gascó I Mora, G., Sánchez Peiró, J.A.: Part-of-Speech tagging based on machine translation techniques. In: Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I. IbPRIA '07, Berlin, Heidelberg, Springer-Verlag (2007) 257–264
10. Laki, L.: Investigating the Possibilities of Using SMT for Text Annotation. In Simões, A., Queirós, R., da Cruz, D., eds.: 1st Symposium on Languages, Applications and Technologies. Volume 21 of OpenAccess Series in Informatics (OASIcs)., Dagstuhl, Germany, Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2012) 267–283
11. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria (2013) 539–545
12. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Association for Computational Linguistics (2007) 177–180
13. James, F.: Modified Kneser-Ney smoothing of n-gram models. Technical report (2000)
14. Laki, L.J., Orosz, Gy., Novák, A.: HuLaPos 2.0 – Decoding morphology. In: 12th Mexican International Conference on Artificial Intelligence, Mexico City, Mexico (2013)
15. Csendes, D., Csirik, J., Gyimóthy, T. In: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. Volume 3206 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004) 41–47
16. Novák, A.: What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
17. Agić, Ž., Ljubešić, N., Merkler, D.: Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, Association for Computational Linguistics (2013) 48–57
18. Jongejan, B., Dalianis, H.: Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, Association for Computational Linguistics (2009) 145–153
19. Georgiev, G., Zhikov, V., Simov, K.I., Osenova, P., Nakov, P.: Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In Daelemans, W., Lapata, M., Màrquez, L., eds.: EACL, The Association for Computer Linguistics (2012) 492–502
20. Chanev, A., Simov, K., Osenova, P., Marinov, S. In: The BulTreeBank: Parsing and conversion. Volume 309 of Current Issues in Linguistic Theory. John Benjamins, Amsterdam & Philadelphia (2007) 321–330

21. Søgaard, A.: Semisupervised condensed nearest neighbor for part-of-speech tagging. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 48–52
22. Maia, M.R.d.H., Xexéo, G.B.: Part-of-speech tagging of Portuguese using hidden Markov models with character language model emissions. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (2011) 159–163
23. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, thicker and easier. In: Proceedings of the 8th international conference on Computational Processing of the Portuguese Language. PROPOR '08, Berlin, Heidelberg, Springer-Verlag (2008) 216–219
24. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330

Mélyesetek a 4lang fogalmi szótárban

Makrai Márton

MTA SZTAKI Nyelvtechnológiai Kutatócsoport
e-mail: makrai@sztaki.mta.hu

Kivonat A predikátumok (főleg igék) argumentumhelyeit mélyesetekbe soroljuk, amelyek a szemantikai és szintaktikai tulajdonságok között teremtenek összeköttetést. A 4lang esetében a mélyesetek nyelvfüggetlen általánosításokat igyekszenek megragadni, melyek segítségével egy mondat jelentésreprezentációja elkészíthető.

A számítógépes szövegértés gyakran használt eszközei a mélyesetek, melyek egyszerre szemantikai és szintaktikai alapon osztályozzák a predikátumok (főleg igék és relációs főnevek, pl. *érdeke valakinek*) argumentumait, és így beazonosítható velük, hogy az egyes frázisok melyik szemantikai argumentum szerepét játsszák. A 4lang egy általános gépi szövegértés céljával készült rendszer lexikona. A cikkben a szótár rövid bemutatása és más erőforrásokkal való összehasonlítása után (1. szakasz) azt írjuk le, hogy hogyan működnek benne a mélyesetek (2. szakasz), és ezt hogyan valósítják meg (3. szakasz).

1. A 4lang fogalmi szótár

A 4langet [1]-ben mutattuk be először, majd a szavaknak egymás definiálásában való fontosságát jellemeztük a szótár segítségével [2], illetve azt vizsgáltuk, hogy a szavak jelentésének kompozícionalitása hogyan jelenik meg egy a 4langből készített vektoros nyelvmodellben (*continuous vector space model*) [3]. Mélyeseteket [4]-ben használtunk először, de még nem a most bemutatott esetkészletet. Itt a véglegesnek szánt esetrendszert mutatjuk be, melynek tesztelése még további kutatás tárgya lehet. Röviden összefoglaljuk a szótárnak azokat a tulajdonságait, amelyek a későbbiek szempontjából fontosak: többnyelvűségét és absztrakt jellegét. Magát a szótárat és néhány kapcsolódó linket a `http://hlt.sztaki.hu/resources/4lang/` címen talál az olvasó.

A 4lang tételei elvileg nyelvfüggetlenek. Az eredeti változatban négy nyelven szerepelnek a tételek (angolul, magyarul, latinul és lengyelül, innen az elnevezés), ami később kiegészült negyvenre [5]. A tételek más szempontból is absztraktabbak, mint a legtöbb hagyományos vagy gép által olvasható szótár tételei, ugyanis egy absztrakt fogalmi jelentést igyekeznek megragadni, így a szótár nem törődik a szófaji különbségekkel, és egy szóalaknak csak homonímia esetén feleltet meg több tételt. Utóbbit módszertanilag ahhoz kötjük, amikor egy szóalaknak (pl.

az angol *state*) egy másik nyelven két olyan szó felel meg, amelyek között nincs különösebb jelentéskapcsolat (szinkrón szinten ilyen a magyar *állam* és *állapot*).

A `4lang` tehát egy többnyelvű szótár absztrakt tételekkel. [1]-ben a Webster's Third [6], [2]-ben pedig a WordNet [7] egy-egy tételén szemléltettük ezt a különbséget. Most hadd említsük még meg a BabelNetet [8] és a VerbNetet [9]. A BabelNet a `4lang`-hez hasonlóan többnyelvű, de ismét csak poliszémebb felfogású, és nem csak szótár, hanem ontológia is. A VerbNetet az teszi érdekessé a jelen cikk szempontjából, hogy – bár csak az angolról szól – eseteket (tematikus szerepeket) is használ és formalizálja a jelentést. Utóbbit szemantikai predikátumokkal teszi, Moens és Steedman [10] eseményfelbontásához (*event decomposition*) hasonló módon. Tematikus szerepből 26 van, míg a `4lang`-ben, mint látni fogjuk, csak nyolc, és ebből is három lokatív. A `4lang` fő újdonsága itt az, hogy míg a VerbNet a tárgyas, tárgyatlan, passzív stb. használatoknál más-más vonzatkeretet (tematikusszerep-listát) ad meg, addig a `4lang` szerint egy fogalomnak csak egy reprezentációja van, és az ige által kiosztott összes mélyeset abban szerepel.

2. A mélyesetek szerepe a `4lang`ben

Amikor egy mondat jelentésreprezentációját akarjuk kiszámítani, fel kell térképezni a predikátum–argumentum viszonyokat. Elméleti nyelvészeti szempontból itt két támaszunk van: a szelekciós megszorítások, és a tág értelemben vett felszíni esetek (nyelvenként változó módon pl. a frázisok sorrendje, esetragok és/vagy adpozíciók). Kutatócsoportunk felfogásában a szelekciós megszorításoknak a *terjedő aktiváció (spreading activation)* felel meg a szótárban, a felszíni esetekkel kapcsolatos tudást pedig közvetve a mélyesetek kódolják. A mélyesetek szempontjából fontos, hogy a `4lang`minden nyelvhez egy nyelvspecifikus modullal fog kapcsolódni, ami megmondja, hogy az egyes mélyesetek az adott nyelvben mely felszíni esetekkel valósulnak meg. Ebben a cikkben a mélyesetekkel foglalkozunk, ezért az aktivációterjedést csak röviden és leegyszerűsítve vázoljuk fel.

Tekintsük az úgynevezett definíciós gráfot, melynek csúcsai a szótárban szereplő fogalmak, és két ilyen akkor van összekötve, ha valamelyik szerepel a másik definíciójában (lásd bővebben a már idézett [2]-ben), pl. a 'tej' össze van kötve a 'folyadék'-kal. Ha szeretnénk megtudni, hogy egy mondatban a *tej* szó az *iszik* szó melyik argumentumát tölti be, akkor megkeressük a két fogalom között a legrövidebb utat (élsorozatot) a gráfban. Szerencsés esetben ez áthalad a *folyadék* szón, és nagyjából megfelel a *tejet iszik* kifejezés reprezentációjának.

Térjünk most rá arra, hogy a felszíni esetekből hogyan lehet kiszámítani, hogy az egyes vonzatok melyik argumentum szerepét töltik be. Egy olyan kifejezés jelentésreprezentációját, amely egy régenst a vonzataival együtt tartalmaz (pl. igei frázis), a kompozícionalitás elve szerint a következőkből kell kiszámítanunk: a régens jelentésreprezentációja, a vonzatoké, valamint az az információ, hogy mindezek együtt milyen szerkezetet alkotnak. Az utóbbiról a `4lang` esetében úgy gondoskodunk, hogy a régens (tipikusan ige) jelentésreprezentációjában feltüntetjük, hogy az egyes vonzatok jelentésreprezentációjának hova kell kerül-

nie. Ehhez a többargumentumú predikátumok (pl. tárgyas igék) argumentumait meg kell tudnunk különböztetni. Ezt a vonzatok mélyesetére való hivatkozással tesszük. Módszerünk mögött az a közkeletű feltételezés húzódik, hogy az argumentumok szemantikai szerepe (pl. ágens) és szintaktikai tulajdonságai (a vonzat felszíni esete, mely mondatalternációkban vesz részt a szerkezet) között (nyelveken belül) rendszeres, és több esetben különböző nyelvekben is felbukkanó megfelelések vannak, ha nem is kivétel nélküliek.

Ahogy már írtuk, a mélyesetek nálunk csak arra szolgálnak, hogy a beazonosíthassuk, hogy melyik vonzat melyik. Ennek kapcsán nem árt talán hangsúlyozni, hogy az argumentumok mélyesetekbe való besorolása elsősorban nem szemantikai osztályozás. A számítógépes szemantikában sokszor érv lehetett két mélyeset közötti különbségtétel mellett az, ha a megfelelő argumentumok jelentése között van egy rendszeres különbség. Például Talmi a *hide/mislay, pour/spill, ...* igepárok tagjai közötti szándékosságbeli különbséget annak tulajdonítja, hogy az ágens pontosan milyen esetű. Nálunk ilyen különbségek nem indokolják új mélyeset bevezetését, hiszen a jelentés teljesen le van írva a lexikai tétel definíció részében.

A szemantikai alapú osztályozáshoz képest az, ahol az esetek száma nem haladhatja meg a legnagyobb argumentumszámot, amivel az igék között találkozunk. Mi erre sem törekszünk, hiszen szabályszerűségeket szeretnénk kihasználni a szemantikai szerep és a szintaktikai tulajdonságok között.

3. Az egyes viszonyok

3.1. Kevéssé tartalmas elemek

Hogyan ragadja meg a `4lang` az egyszerűbb függőségeket? Egyrészt bizonyos ragoknak, pl. a többes számnak fogalmi jelentésük van abban az értelemben, hogy annak a szerkezetnek a jelentésreprezentációjában, aminek a rag része, van egy olyan eleme, amiért a rag felel. Hasonlóak a produktív képzők és az adpozíciók is. Ezeket a viszonyokat (tő–rag, tő–képző, adpozíciós tárgy–adpozíció) már csak azért is egységesen kell kezelnünk, mert a `4lang` nyelvfüggetlen kíván lenni, és ugyanazt a szemantikai viszonyt különböző nyelvek különbözően fejezik ki, pl. azt a jelentést, ami a magyarban a birtokos személyrag fejez ki, az angolban birtokos névmás. Itt a funkcióelem reprezentációjának a tartalmasabbik elem reprezentációjában való helyét mindig a `REL` (*relációs, related*) kulcsszó képviseli, amit tágabb értelemben mélyesetnek is nevezhetünk.

3.2. Igei mélyesetek

Rátérve most már az igék argumentumaira, először is tisztáznunk kell, hogy mit értünk argumentum alatt. Csak a kötelező vonzatokat, vagy a szabad bővítményeket is? A felszíni argumentumokról beszélünk, vagy a az igének egy formális szemantikai fordításban megfelelő függvény argumentumairól? Ezekre a kérdésekre első közelítésben azt válaszoljuk, hogy azokat a felszíni argumentumokat jelenítjük meg mélyesettel egy ige definíciójában, amire a jelentés leírásához

szükség van. Egy másik szempont abból adódik, hogy a `4lang` absztrakt volta miatt nem teszünk különbséget ugyanazon igealak tárgyas (esetleg még több felszíni argumentummal rendelkező) és tárgyatlan használata között. A mélyeseteket úgy kell meghatároznunk, hogy az különböző használatokban ugyanaz a szereplő ugyanazt az esetet kapja. Ebből következik, hogy ha egy igének van tárgyas használata, akkor két szereplő mélyesetét is fel kell tüntetni. Végül árnyalja a képet, hogy amikor egy ige egy másik ige speciális eseteként definiálható, és az argumentumok is öröklődnek, akkor nem szükséges kiírni az eseteket, Például a *bite* igét `cut, INSTRUMENT tooth`-ként definiáltuk ('foggal szakít'), és a *harap* örökli a *szakít* argumentumait, ezért ezeket nem tüntettük föl.

Az igei mélyesetek megválasztásánál nem feladatunk, hogy különböző igei tövek szereplői között összhangot teremtsünk. Így például nem célunk, hogy a *János elad Péternek egy könyvet* és a *Péter vesz egy könyvet Jánostól* mondatok szereplői a két mondat esetében ugyanazokat a mélyeseteket kapják.[1] Végül nem írjuk bele az igető definícióba az úgynevezett külső szerepeket (*outer role*), vagyis azokat a lehetséges bővítményeket, amikkel egy olyan konstrukció tudja felruházni az igét, ami egész igeosztályokat (pl. mozgásigék) vagy akár minden igét érint, így a következő példákban a vastagon szedett frázisoknak megfelelő (nemlétező) argumentumpozíciót: *fest egy képet* ***valakinek****, alszik* ***egy órát****, átröpüli* ***az Atlanti-óceánt***[2]. A kauzációt (*úszik* → *úsztat*), képzésnek tekintjük, és így a képzőt tesszük érte felelőssé (annak ellenére, hogy az angolban zéróképzéssel történik).

AGT	383
PAT	311
REL	81
POSS	52
DAT	30
TO	17
FROM	11
AT	2

1. táblázat. Az egyes mélyesetek az őket használó szavak számával

A `4lang`ben 744 igét találunk. A mélyeseteket a 1. táblázat tartalmazza azzal a számmal együtt, hogy hány szónál fordulnak elő. A leggyakoribb mélyeset az ágens. A definíciók írásakor nagyjából problémamentesen el tudjuk dönteni, hogy egy többargumentumú igének melyik argumentuma az ágens (amit az `AGT` kulcsszó jelöl a szótárban). A `4lang`-ben második leggyakoribb mélyesetet, amit mi páciensnek neveztünk (`PAT`), sokszor csak a „szemantikailag jelöletlen" mélyeset-

[1] Előrebocsátjuk, hogy mindkét ige esetében a magyar, (vagy ami ugyanaz, az angol) alany lesz az ágens, és a tárgy a `PAT`.

[2] A külső szerepekről lásd bővebben [11] 9. fejezetét.

ként határozzák meg, de mivel a többi viszonylag egyértelműen beazonosítható, ez sem okoz problémát. A szintaxisban bevett unakkuzatív hipotézis szerint egyargumentumú igék argumentuma is lehet páciens (pl. *süllyed, fürdik*). Arról, hogy az intranzitív és tranzitív igék ágensét, illetve páciensét hogyan sorolják felszíni esetbe különféle nyelvek, jó összefoglalót ad Komlósy [12]. Áttekint számos ergatív (illetve aktív és alanyjelölő) nyelvet abból a szempontból, milyen esetet kap különböző egyargumentumú igék argumentuma. A 2. táblázatban mutatja be, hogy a különböző nyelvek hol húzzák meg a határt a két eset között egyfajta aktivitási skálán. Ezek az adatok azt sugallják, hogy egy nyelvfüggetlen esetrendszerben a bináris `AGT` vs. `PAT` felosztásnál finomabb különbségeket kell tennünk. Kérdés, hogy ez valóban javítana-e a rendszerünknek ezeken a nyelveken való teljeseítményén. Ezt egyelőre nem tudjuk megvizsgálni, így maradunk az egyszerűbb esetkészletnél.

	tárgyjelölő	ergatív 1.	ergatív 2.	aktív	lexikalizált aktív	alanyjelölő
Péter írja a levelet.	nom	ag	ag	ag	ag	ag
Péter ír.	nom	nom	ag	ag	ag	ag
Péter sétál.	nom	nom	nom	ag	ag/nom	ag
Péter beteg.	nom	nom	nom	nom	ag/nom	ag

tárgyjelölő	angol (eng), magyar (hun)
ergatív 1.	kabardi (kbd), avar (ava), adige (ady)
ergatív 2.	agul (agx), udi (udi)
aktív	bacbi (bbl)
lexikalizált aktív	grúz (kat), dakota (dak)
alanyjelölő	megrel (xmf), maidu (nmu)

2. táblázat. Egyargumentumú igék argumentuma különféle nyelvekben [7]. A nyelvek SIL kódját is feltüntettük.

Az ágenssel és a pácienssel lényegében a generatív szemantikai hagyományt követjük. Jobban eltérünk az előzményektől a datív (`DAT`) használatával. Az elnevezést a generatív szemantika legrégebbi szóhasználatából vesszük [13,14], és datívnak magyarítjuk, a datívusz szót fenntartva az ilyen nevű felszíni esetnek. Maga Fillmore később a datívot különbontotta experiensre, tárgyra (*Object*) és célra (*Goal*). Mi a datívot alapvetően csak legalább három felszíni argumentummal rendelkező igék esetén használjuk, más esetekben csak az ezekkel való hasonlóság alapján. A háromargumentumú igék egy része jelentésére nézve a *mond* speciális esete vagy legalábbis kommunikációról szól (*bevall, enged, parancsol, kijelent, megmagyaráz, kifejez, megtilt, hálás, köszön (valamit valakinek), bemutat, mond, mutat, esküszik*), egy másik csoport pedig az *ad* speciális esete vagy változata (*kölcsönad, átenged, bérbe ad, áldoz (istennek), ajánl, tartozik,*

fizet, elad, adományoz, ad, segít). Megjegyezzük, hogy a *valaminek tart* és a *valami(lyen)nek tűnik* igéknek az a vonzata, ami a magyarban datívuszos, nem datívban van, hanem a mienknél finomabb felosztás szerint [15] komplementum (Comp) lenne, amit mi a `PAT`-ba sorolunk.

A `4lang`ben három lokatív eset is van, a Fillmore-i Célnak (*Goal*), illetve Forrásnak (*Source*) megfelelő `TO`, illetve `FROM`, valamint a statikus `AT`. Meglehetősen messzire mentünk az absztrakcióban: ha egy vonzat sok nyelvben olyan felszíni esetet kap, ami mozgás céljának kifejezésére is használatos (a magyarban főleg határozóragok, az angolban pedig prepozíciók), akkor célnak tekintjük. Ide sorolódtak: *képes, hozzászokik, (hozzá)ad, összeadás, meghív, csatlakozik, tesz, hasonlít*. A másik két lokatív eset a forrás (*elfogad, kölcsönöz, vesz, levág, ered, eltávolít, bérel, kivon, elvesz*) és a statikus hely is (*valahol helyezkedik el, marad*).

A már említett nyelvspecifikus modulban lehetőség van megjelölni egyes igék egyes argumentumait felszíni esetekkel, amennyiben azok nem a mélyesetükből jósolható esetet kapják (*ferde eset, quirky case*). Viszont már az angol, magyar és német alapján világos, hogy maradnak igék, amelyeknél nem lehet általánosítani. Ekkor ugyanazt a `REL` kulcsszót használjuk, mint az egy felszíni argumentumú régenseknél (*prefer to something, jobban szeret valaminél*).

3.3. Relációs főnevek

Végül arról a viszonyról beszélünk, amit a *relációs főnevek* és a hozzájuk kapcsolódó szó (pl. az *érdek* szó esetén az érdekelt) között van. A jelenség, ami miatt az *érdek* főnevet relációsnak nevezem, kettős. Egyrészt az *érdek* szó előfordulásai között a birtokjelesek aránya légyegesen nagyobb, mint más főneveknél. Másrészt, és szemantikai szempontból ez az érdekes, bárhogy akarnánk is leírni az *érdek* szó jelentését, alighanem hivatkoznánk az érdekeltre. A két szó közötti grammatikai viszony a legtöbb esetben birtokos, de az esetek körülbelül egytizedében mást találunk. Az *alkalom* és a *szükség* szavak esetén az a szereplő, amit jobb híján célnak hívunk, magyarul szublatívuszba kerül (*-rA*), angolul (*occasion, need*) pedig *for* prepozíciót kap. A relációs főnevek reprezentációjában a két szó közötti grammatikai viszony szerint a `POSS`, illetve a `TO` kulcsszóval jelöljük azt a helyet, ahova a kapcsolódó szó (az érdekelt, illetve a cél) reprezentációja kerül. A `TO` ugyanaz az absztrakt cél, amivel már az igéknél is találkoztunk. A mélyesetek közvetítésével így a nyelvi viszony segít megtalálni a két dolog (az érdekelt és az érdek, illetve az alkalom és a cél) közötti szemantikai viszonyt. Nem foglalkozunk azokkal a relációs főnevekkel, amelyek produktívan vannak képezve valamely igéből, ugyanis ezeket a `4lang` nem különbözteti meg attól az igétől, amelyből képeztük őket.

4. Összegzés

Bemutattuk, hogyan működhetnek a mélyesetek egy olyan szabadon elérhető gépi szövegértő erőforrásban, amely a mélyeseteket közvetlenül az egyes nyelvek szavainál absztraktabb nyelvfüggetlen fogalmakhoz rendeli. A kutatás következő

lépése a mélyesetek működésének tesztelése egy gépi szövegértési feladatban. Az is későbbi kutatás témája, hogy egy tágabb szókincsbe tartozó igék szemantikai argumentumainak hogyan lehet mélyesetet tulajdonítani az őket definiáló szavaknak a szótárban levő emberi nyelvű definíciója segítségével. Végül távlati célként megemlítjük a generálást, ami úgy is felfogható, hogy adott egy r jelentésreprezentáció, és azok közül a (nem feltétlenül grammatikus) szósorozatok közül, amelyekhez a jelenlegi rendszer az r reprezentációt rendel, ki kell választani a leggrammatikusabbat, és az információs helyzetnek legjobban megfelelőt.

Köszönetnyilvánítás

Köszönöm témavezetőm, Kornai András munkáját, aki kitűzte a kutatási irányt, és fontos ötletekkel segített. Az esetkészlettel kapcsolatban fontos döntéseket a kutatócsoporton belül elsősorban Nemeskey Dáviddal hoztunk meg. Köszönöm Naszádi Katának a beszélgetéseket, amelyektől sokat tisztult a mélyesetekről való képem.

Hivatkozások

1. Kornai, A., Makrai, M.: A 4lang fogalmi szótár. In Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 62–70
2. Makrai, M.: Fogalmak fontossága a definíciós gráf vizsgálatával. In Váradi, T., ed.: VII. Alkalmazott Nyelvészeti Doktoranduszkonferencia, MTA Nyelvtudományi Intézet Budapest (2013)
3. Makrai, M., Nemeskey, D.M., Kornai, A.: Applicative structure in vector space models. In: Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, Sofia, Bulgaria, Association for Computational Linguistics (2013) 59–63
4. Nemeskey, D., Recski, G., Zséder, A.: Miből lesz a robot MÁV-pénztáros? In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2012)
5. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
6. Gove, P.B., ed.: Webster's Third New International Dictionary of the English Language, Unabridged. G. & C. Merriam (1961)
7. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM **38**(11) (1995) 39–41
8. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
9. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: A large-scale classification of english verbs. Language Resources and Evaluation **42**(1) (2008) 21–40
10. Moens, M., Steedman, M.: Temporal ontology and temporal reference. Computational Linguistics **14**(2) (1988) 15–28
11. Somers, H.L.: Valency and case in computational linguistics. Edinburgh University Press (1987)

12. Komlósy, A.: Deep structure cases reinterpreted. In Kiefer, F., ed.: Hungarian General Linguistics. John Benjamins. Amsterdam–Philadelphia (1982) 351–385.
13. Heringer, T.: Wertigkeiten und nullwertige verben im deutschen. Zeitschrift für Deutsche Spache **23** (1967) 13–34
14. Fillmore, C.: The case for case. In Bach, E., Harms, R., eds.: Universals in Linguistic Theory. Holt and Rinehart, New York (1968) 1–90
15. Chafe, W.L.: Meaning and the structure of language. Chicago: University Press (1970)

Statisztikai konstituenselemzés magyar nyelvre

Szántó Zsolt, Farkas Richárd

Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
szanto.zsolt@stud.u-szeged.hu, rfarkas@inf.u-szeged.hu

Kivonat Előadásunkban bemutatjuk, hogy a nyelvfüggetlen – valószínűségi környezetfüggetlen nyelvtanokat használó – Berkeleyparser [1] milyen eredményeket ér el a Szeged Treebanken, majd tárgyalunk két technikát, melyek jelentősen javítják az elemzések pontosságát morfológiailag gazdag nyelvekben.

Kulcsszavak: konstituenselemzés, morfológiai kódkészlet

1. Bevezetés

A szintaktikai elemzés szempontjából a világ nyelvei általában a morfológiai gazdagságuk szintjei szerint vannak csoportosítva (ami fordítottan arányos a nyelv konfigurációs szintjével). A skála egyik végében ott található az angol, egy erősen konfiguratív nyelv, míg a másik oldalon ott a magyar a maga gazdag morfológiájával és szabad szórendjével [2]. A szintaktikai elemzők általában az angol nyelvet figyelembe véve lettek kifejlesztve, ezzel szemben a világ nyelveinek jó része alapjaiban különbözik az angoltól. Különösképpen a morfológiailag gazdag nyelvek, melyek a legtöbb mondatszintű szintaktikai információit a morfológia (azaz a szavak) szintjén, és nem a szórenddel fejezik ki. Ezen különbségek miatt a morfológiailag gazdag nyelvek elemzése olyan technikákat igényel, melyek különböznek az angol nyelvre kifejlesztett módszerektől (vagy kiterjesztik azokat) [3]. Ebben a tanulmányban a konstituenselemzés tökéletesítésének érdekében két olyan technikát mutatunk be, amelyek speciálisan a morfológiailag gazdag nyelvek kihívásainak kezelésére hivatottak.

Az utóbbi két évtizedben jelentős mértékben fejlődtek a konstituenselemzők [4,5,1,6], ami elsősorban a Penn Treebank jelenlétének köszönhető [7]. Amíg angol nyelven is folyamatos fejlődés volt tapasztalható, a morfológiailag gazdag nyelvek treebankjei kevés figyelmet kaptak. Magyar nyelvre a Szeged Treebank [8], egy – nemzetközi viszonylatban is – nagyméretű, kézzel annotált konstituenskorpusz már közel 10 éve rendelkezésre áll. Annak ellenére, hogy ez kiváló alapanyagul szolgálhatna statisztikai[1] konstituenselemzők fejlesztéséhez, néhány korai kísérletet leszámítva, legjobb tudomásunk szerint ez idáig senki sem kísérelte ezt meg.

[1] az angol 'data-driven' kifejezés fordításaként használjuk a magyar 'statisztikai' szót

Ebben a tanulmányban a morfológiailag gazdag nyelvek két fő problémájára próbálunk meg választ adni. Ezen problémák az optimális preterminálisok (morfológiai kódok) halmazának megtalálása és a szóalakok nagy számának kezelése.

A sztenderd valószínűségi környezetfüggetlen nyelvtanokra épülő konstituenselemzők a preterminálisokat egy-egy struktúra nélküli címkének tekintik. Ezen címkék optimális halmazának meghatározása nagyon kritikus az elemzés hatékonyságára nézve. A két legkézenfekvőbb megoldás, hogy vagy csak a fő szófaji kódokat vagy a teljes morfológiai leírást használjuk címkének. Előbbi kódolással sok információt vesztünk, míg utóbbi esetén a preterminálisok magas száma miatt az elemzés lassú lehet, ill. a tanulás során az optimalizálási feladat kezelhetetlenné válik. Ezen problémák kezelésére kidolgoztunk egy új, teljesen automatikus módszert a morfológiai kódkészlet csökkentésére.

A másik probléma, hogy a toldalékolásnak köszönhetően a morfológiailag gazdag nyelvekben rengeteg eltérő szóalak található (ellentétben az angollal). Ennek következtében az ún. ismeretlen vagy ritkán látott szavak száma nagyon magas, ami negatív hatással van a konstituenselemzők hatékonyságára. Goldberg és Elhadad [9] gondolatait követve kiegészítjük a lexikai modellt külső lexikonok használatával. Megvizsgáljuk, egy teljesen felügyelt szófaji egyértelműsítő mennyire alkalmazható az általuk javasolt felügyelet nélkülivel szemben a külső lexikonok elkészítésére.

2. Korpusz, kiértékelési metrikák

A vizsgálatokhoz a Szeged Treebank [8] újságcikkekből álló alkorpuszát használtuk. A tanító halmazunkban összesen 8146 mondat található, míg ugyanez az érték a teszthalmazban 1051. Az egyes mondatokban átlagosan 21,76 token található. Összesen 680 morfológiai címkét tartalmaz a korpusz, ami 16 fő szófaji kód köré csoportosul. A teszthalmazon az ismeretlen szavak aránya 19,94%.

Kiértékeléskor a PARSEVAL [10] metrikát használtuk, illetve a hibátlanul leelemzett mondatok arányát vizsgáltuk.

3. Kiterjesztett lexikai modellek

Mielőtt bemutatnánk az ötleteinket és eredményeinket a preterminális halmazok optimalizálásra, szeretnénk ajánlani egy megoldást az ismeretlen szavak problémájára, mely kritikus fontosságú lehet a morfológiailag gazdag nyelvekben. Ennek fő oka ezen nyelvekben a toldalékolás következtében létrejövő rengeteg szóalak. Követvén Goldberg és Elhadad [9] ajánlását, kiterjesztettük a lexikai modellt a tokenek lehetséges morfológiai elemzéseinek gyakorisági információival.

Minden egyes t címkére és w szóra az alábbi képlet alapján becsültük a $P(t|w)$ valószínűséget:

$$P(t|w) = \begin{cases} P_{tb}(t|w), & \text{ha } c(w) \geq K \\ P_{ex}(t|w), & \text{ha } c(w) = 0 \\ \frac{c(w)P_{tb}(t|w)+P_{ex}(t|w)}{1+c(w)}, & \text{különben} \end{cases}$$

ahol a $c(w)$ a w tanító halmazon vett előfordulásainak a száma, a K egy előre definiált konstans, a $P_{tb}(t|w)$ a treebank alapján számolt valószínűség, míg a $P_{ex}(t|w)$ valószínűségeket egy külső lexikon alapján kalkuláljuk. A konstituenselemző számára szükséges $P(w|t)$ emissziós valószínűségeket megkaphatjuk a $P(t|w)$ valószínűségekből a Bayes szabály felhasználásával.

A kulcskérdés itt az, hogy hogyan is készítsük el a külső gyakorisági lexikont, amely $P_{ex}(t|w)$ becslésére szolgál. Goldberg és Elhadad [9] javaslata alapján baseline-nak egy olyan lexikont használtunk, melyben az adott szó lehetséges morfológiai elemzéseit egy morfológiai elemző segítségével határozzuk meg, és ezekre a valószínűségeket egyenletes eloszlással számítjuk.

Goldberg és Elhadad [9] jelentős javulásról számolt be héber nyelvre, amikor az egyenletes eloszlást használó baseline helyett a gyakoriságokat egy olyan nagyméretű korpuszon számolták le, amelyet felügyelet nélküli szófaji egyértelműsítő rendszer [11] használatával automatikusan annotáltak. Megmutatjuk, hogy felügyelt szófaji egyértelműsítéssel ugyanolyan mértékű javulás érhető el. Elsősorban az motiválta a felügyelt egyértelműsítő használatát, hogy – a felügyelet nélküli modellel szemben – nem igényel morfológiai elemzőt (amely meg tudná adni egy szóra a lehetséges morfológiai címkéket). Bár magyar nyelvre rendelkezésünkre áll morfológiai elemző, de ezen elemzők teljesen nyelvfüggők, ráadásul az sem garantált, hogy kompatibilisek az adott treebankkel, így közel sem biztos, hogy egy ezekre építő módszer általánosan használható lesz bármely morfológiailag gazdag nyelv esetén. Ezzel szemben bármikor felépíthetünk egy elfogadható felügyelt morfológiai egyértelműsítő rendszert az adott treebankünk tanító halmazán.

A címkézetlen szövegekben a morfológiai egyértelműsítés folyamatára a feltételes véletlen mezőkre (CRF) építő MarMot [12] szófaji egyértelműsítő rendszert alkalmaztuk. Ez a tisztán statisztikai elemző 97,6%-os pontosságot ért el a tesztthalmazunkon, amely versenyképes a nyelvfüggő szabályokat is alkalmazó magyar nyelvre használt szófaji egyértelműsítőkkel (például a magyarlanccal [13]).

1. táblázat. PARSEVAL eredmények és a hibátlanul elemzett mondatok aránya (EX) különböző külső lexikonok használata mellett.

	PARSEVAL	EX
BerkeleyParser	87.22	12.75
egyenletes eloszlás	87.31	14.78
teszt	88.29	15.22
teszt + MNSz	89.27	16.97

Az 1. táblázat megmutatja az eltérő $P_{ex}(t|w)$ becslések eredményeit a tesztthalmazon. Az első sorban az általunk abszolút baseline-ként használt 'BerkeleyParser' található, ami az elemző eredeti implementációja [1]. Az egyenletes eloszlással készített lexikonhoz a magyarlanc morfológiai elemzőjét használtuk.

Az utolsó két sor a szófaji egyértelműsítés felhasználásával kapott eredményeket mutatja be. Ehhez a MarMotot az újsághírek tanító halmazán tanítottuk, és ennek segítségével leelemeztettük a teszthalmazt, illetve – hogy tényleg nagyméretű korpusszal tudjunk dolgozni – 10 millió címkézetlen mondatot a Magyar Nemzeti Szövegtárból [14]. Az eredmények között külön beszámolunk a teszthalmazon ('teszt') és a teszthalmazon, illetve a nagyméretű korpuszon együttesen számolt ('teszt + MNSz') gyakoriságok mellett elért eredményéről.

Néhány előzetes kísérlet után beállítottuk a K értékét 7-re.

A 1. táblázatból látható, hogy az 'egyenletes eloszlás' mellett, habár a PARSEVAL értékben nem sokat javul, a tökéltesen elemzett mondatok aránya jelentősen javul. A 'teszt' konstrukció tekintélyes növekedést mutatott az 'egyenletes eloszlással' szemben is, ami összhangban van a Goldberg és Elhadad által megállapítottakkal. Emellett láthatjuk azt is, hogy a nagyméretű címkézetlen korpusz használata szintén jelentősen javulást hozott az eredményekben. A későbbi eredmények vizsgálatához innentől kezdve a Magyar Nemzeti Szövegtárra és a teszthalmazra építő külső lexikont tartalmazó megvalósítást fogjuk használni.

4. Morfológiai kódok automatikus összevonása

A preterminális címkék halmazának optimális megadása kritikus lehet bármely valószínűségi környezetfüggetlen nyelvtant használó konstituenselemző számára. Morfológiai jellemzők törlésével csökkenthetjük a feladat bonyolultságát, de el is veszíthetünk a szintaxis számára hasznos információkat. Ebben a fejezetben leírunk egy általunk kidolgozott eljárást a preterminálisok optimális halmazának automatikus megadására, és a hatékonyságát empirikus eredmények alapján vizsgáljuk különböző baseline-okkal összehasonlítva.

4.1. Eljárás morfológiai jellemzők értékeinek összevonására

A múltban már jelentek meg publikációk a morfológiai kódok számának automatikus csökkentésével kapcsolatban. Ezek egyikében Dehdari [15] bemutatott egy rendszert, melyben az egyes morfológiai jellemzőket egységként kezelte, és ezen egységek iteratívan kerültek törlésre, majd az így kapott új kódkészletet úgy értékelte ki, hogy a tanítástól kezdve újrafuttatta a konstituenselemzőt. Ezzel kapcsolatban két probléma is felmerül. Az első, hogy véleményünk szerint a morfológiai jellemzőket nem szabad egységként kezelni, hiszen egy adott jellemző eltérő értékei viselkedhetnek különbözően. Vegyük például a fokot a melléknevekben, itt az alap- és felsőfok azonosan viselkedik (összevonható), amíg az előbbi két érték megkülönböztetése a középfoktól hasznos lehet a szintaktikai elemző számára, mert a középfokú mellékneveknek általában rendelkeznek egy vonzattal (például: *Kati szebb, mint Zsófi*), míg az alap- és felsőfok nem. A második, hogy az előbbi cikkben az egyes morfológiai jellemzők kerültek törlésre függetlenül attól, hogy milyen szófajhoz tartoznak, azaz ha az eset (Cas) jellemző törlődött, akkor törlődött a főnevek, illetve a melléknevek jellemzői közül

is, pedig előfordulhat, hogy az egyes jellemzők egy adott szófaj esetén hasznosak, de más szófaj esetén törölhetők.

Az alábbi megfigyelésekre alapozva terveztünk egy új módszert, ami a fő szófaji kódokból kiindulva iteratívan összevonja az egyes morfológiai jellemzők értékeit, miközben az eltérő szófajokhoz tartozó (azonos) jellemzőket külön kezeli. A folyamat eredményeként kapunk egy csoportosítást az egyes morfológiai jellemzők lehetséges értékei felett. A mi megközelítésünknek egy speciális esete lesz az, amikor egy morfológiai jellemző kitörlődik. Ez akkor fordulhat elő, ha az adott morfológiai jellemző minden értéke egy csoporttá vonódik össze, ekkor a kérdéses jellemzőnek nem lesz többé megkülönböztető szerepe. Ennek következtében a mi munkánkra tekinthetünk úgy, mint az előbbi módszer egy általánosítására.

Ezen általános megközelítés jelentősen megnöveli a lehetséges preterminális halmazok számát, melyek egyenkénti kiértékelése megvalósíthatatlan lenne egy külső elemző folyamatos újratanításával (a BerkeleyParserrel egy átlagos méretű korpuszon a tanítás és elemzés több mint 1 órát vesz igénybe). Elképzelésünk szerint nem szükséges az elemző újratanítása minden egyes preterminális halmazra. Globális célunk, hogy a konstituenselemzés-beli hasznosságuk alapján válogassunk az egyes halmazok között. Ez megegyezik a BerkeleyParser rejtett állapotokat összevonó eljárásának motivációjával. A BerkeleyParser miután véletlenszerűen szétbontotta a nemterminális alállapotokat, újratanítja a nyelvtant, majd minden egyes szétbontásra kiszámítja, hogy mekkora veszteséggel jár az egyes szétbontott alállapotok összevonása. Ha ez az információveszteség kicsi, a szétbontással keletkezett alállapotok nem hordoztak elég hasznos információt, ezért összevonhatjuk őket. A mi feladatunk ugyanez, azaz meg kell találnunk a megfelelő összevonásokat a morfológiai jellemzők értékeire. Ennek következtében a preterminális szinten – a BerkeleyParser által létrehozott alállapotok helyett – a morfológiai jellemzők értékeire meghívjuk az előbb említett összevonó eljárást. Ennek következtében a BerkeleyParser bináris elágazású véletlenül szétbontott hierarchiája helyett, a mi alállapot-keresési terünk egy háromszintes hierarchia lesz, ahol az első szinten a fő szófaji kódok, a másodikon a morfológiai jellemzők és a harmadikon az egyes jellemzők értékei találhatók. Mivel ez a hierarchia nem bináris elágazású, ezért módosítottuk a BerkeleyParser idevonatkozó implementációját.

A gyakorlatban első lépésként tanítjuk a BerkeleyParsert a sztenderd módon a teljes kódkészlet használatával, majd a preterminális szimbólumok alállapotait újra egyesítjük. Ezután az összes fő szófaji kód-morfológiai jellemző párt különkülön, egymástól függetlenül vizsgáljuk. Minden egyes jellemző esetén az adott jellemző értékeit mint alállapotokat fogjuk használni, melyek valószínűségeit egyenletes eloszlással adjuk meg. A nyelvtanban direkt módon újra tudjuk számolni a lexikai valószínűségeket (preterminális $\rightarrow$ terminális átmenetek), annak köszönhetően, hogy ismerjük az új alállapotaink előfordulásait az egyes konstituensfákban. Ezekután kiszámítjuk jellemzőnként az összes alállapotpárra a valószínűségben történt veszteségét. Ezen információk felhasználásával minden jellemzőre létrehozunk egy teljes gráfot, melyben a csúcsok a preterminális

alállapotai (jellemző értékei) és az élek súlyai a két alállapot összevonásával kapott veszteségek. Az így kapott gráfokból kitöröljük a legnagyobb súllyal rendelkező éleket (a kitörlendő élek arányát a *th* metaparaméter segítségével szabályozhatjuk). Végül az egyes gráfokban megkeressük az összefüggő komponenseket, és ezen komponensek értékeit összevonjuk, az így kapott új értékek lesznek az adott morfológiai jellemző új értékei.

4.2. Baseline preterminális halmazok létrehozása

A javasolt módszert négy módszerrel állítjuk szembe. A két legegyszerűbb irány preterminális halmaz készítésére a fő szófaji kódok és a teljes morfológiai leírás használata. Ezen felül magyar nyelvre rendelkezésünkre áll egy köztes méretű kódhalmaz is, melyet a magyarlanc fejlesztésekor nyelvészeti szempontokat figyelembe véve kézzel hoztak létre [13]. Ez a manuálisan létrehozott kódhalmaz eltérő szófaji kódok esetén eltérő morfológiai jellemzőket tartalmaz, és az összevonások benne a morfológiai értékek szintjén történtek, ami alapján nem lehet meglepő, hogy az előző szakaszban bemutatott automatikus összevonó eljáráshoz ezen korábbi kézi megvalósítás is erős inspirációként szolgált.

Az utolsó baseline-unk a Dehdari [15] által javasolt kísérlet magyar nyelvre való megismétlése. Ezért a teljes morfológiaijellemző-halmazból kiindulva mindig töröltünk egy-egy jellemzőt, és az így kapott új halmazokkal újratanítottuk a konstituenselemzőnket. Azt tapasztaltuk, hogy a leghatározottabb visszaesést a PARSEVAL statisztikában a 'Cas' jellemző törlése okozta, míg a legenyhébbet a 'Type' törlése mellett kaptuk. Mivel a névszók esetragjai (Cas) hordozzák a mondat szintaktikai felépítése szempontjából legfontosabb információt, azaz hogy az adott névszó pontosan milyen nyelvtani szerepet tölt be az adott mondatban (pl. tárgy, részeshatározó stb.), nem meglepő, hogy ennek törlése esetén a parser teljesítménye jelentősen visszaesik. Ezzel szemben a Type jellemző pusztán a nyílt szóosztályok néhány fajtájában fordul elő (pl. a dátumot, időt jelölő számsorokat különíti el egymástól), ami egy szemantikai jellegű megkülönböztetés, és az adott egység szintaktikai viselkedésére nincs különösebb hatással.

4.3. Eredmények különböző preterminális halmazokkal

A 2. táblázat összesítve tartalmazza a baseline módszerekkel és a saját automatikus összevonó megoldásunk által megkapott preterminális halmazokkal mért eredményeket. Az összevonó algoritmussal két különböző címkehalmazt is megadtunk, melyek eltérő küszöbérték (*th*) mellett lettek összevonva.

A fő szófaji kódok és a teljes morfológiai leírás közötti különbség meglepően magas, ebből következik, hogy a preterminálisok által hordozott morfológiai információk rendkívül hasznosak a konstituenselemző számára, és hogy a BerkeleyParser képes sok száz elemű preterminális halmazok kezelésére. Magyarra azt találtuk, hogy az egyes jellemzők teljes eltávolításától az eredmények nem javulnak. Ez a felfedezés szögesen ellentmond Dehdari [15] arab nyelvre tett megfigyeléseivel, ahol a 'Case' eltávolításától a PARSEVAL eredmény 1%-kal lett

2. táblázat. PARSEVAL eredmények és a hibátlanul elemzett mondatok aránya (EX) eltérő preterminális halmazok mellett.

	#pt	PARSEVAL	EX
fő szófaji kód	16	83.47	7.52
manuális	72	86.43	13.04
teljes	680	89.27	16.97
teljes - Cas	479	84.76	9.53
teljes - Type	635	89.15	16.97
összevont ($th = 0.5$)	378	89.28	**17.73**
összevont ($th = 0.1$)	642	**89.40**	16.49

jobb. Megfigyeltük, hogy a baseline eredmények is teljesen eltérnek a két nyelv között, míg magyarra a teljes morfológiai leírás sokkal eredményesebbnek bizonyult a fő szófaji kódoknál, addig ugyanez a két érték arabra Dehdari eredményei alapján közel azonos volt.

A táblázat szintén tartalmazza az általunk tervezett eljárás két különböző eredményét. A th=0.1 esetben csak pár morfológiai jellemző érték került összevonásra, és ez enyhe javulást eredményezett a teljes kódhalmazt tartalmazó konfigurációval szemben. A másik esetben, ahol a th értéke 0.5, közel azonos eredményt kaptunk a teljes morfológiai leírással, miközben feleannyi preterminálist használtunk (ráadásul a hibátlanul elemzett mondat aránya releváns javulást mutatott). Következésképpen, habár statisztikailag nem lett jobb az eredmény, mint a legjobb baseline esetében, de az elemzés futási ideje majdnem a felére csökkent.

Összességében az összevonó megoldásunk a teljes morfológiai leírásnál jobb preterminális halmazokat talált meg, melyek az új címkék számától függően javítottak az eredményeken vagy gyorsították az elemzést.

5. Konklúzió

Ebben a tanulmányban vizsgáltuk a konstituenselemzők hatékonyságát magyar nyelvre, ezen felül két olyan technikát mutattunk be, amelyek az elemzés javítására szolgálnak morfológiailag gazdag nyelveken.

A fő eredményünk a preterminális összevonó eljárás, ami az előző munkáknál egy általánosabb és gyorsabb megoldást ad köszönhetően annak, hogy nincs szükségünk a konstituenselemző lehetséges preterminális halmazonkénti újratanítására. Az összevonó eljárásnak köszönhetően javítani tudtunk az elemzés pontosságán és sebességén is.

Kísérleteztünk külső korpuszok felhasználásával is a lexikai modellben. Megmutattuk, hogy felügyelt szófaji egyértelműsítés használatával jelentős javulást lehet elérni a rendszer pontosságában.

Köszönetnyilvánítás

Szántó Zsolt kutatásait a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

Farkas Richárd kutatásai az Európai Unió és Magyarország támogatásával, az Európai Szociális Alap társfinanszírozásával a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosító számú „Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program" című kiemelt projekt keretei között valósultak meg.

Hivatkozások

1. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. (2006) 433–440
2. Fraser, A., Schmid, H., Farkas, R., Wang, R., Schütze, H.: Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language. Computational Linguistics **39**(1) (2013) 57–85
3. Tsarfaty, R., Seddah, D., Kübler, S., Nivre, J.: Parsing morphologically rich languages: Introduction to the special issue. Computational Linguistics **39**(1) (2013) 15–22
4. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. (2000) 132–139
5. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05 (2005) 173–180
6. Huang, L.: Forest reranking: Discriminative parsing with non-local features. In: Proceedings of ACL-08: HLT. (2008) 586–594
7. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330
8. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: TSD. (2005) 123–131
9. Goldberg, Y., Elhadad, M.: Word segmentation, unknown-word resolution, and morphological agreement in a hebrew parsing system. Computational Linguistics **39**(1) (2013) 121–160
10. Abney, S., Flickenger, S., Gdaniec, C., Grishman, C., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., Strzalkowski, T.: Procedure for quantitatively comparing the syntactic coverage of English grammars. In Black, E., ed.: Proceedings of the workshop on Speech and Natural Language. (1991) 306–311
11. Goldberg, Y., Adler, M., Elhadad, M.: EM can find pretty good HMM POS-taggers (when given a good start). In: Proceedings of ACL-08: HLT. (2008) 746–754
12. Mueller, T., Schmid, H., Schütze, H.: Efficient higher-order CRFs for morphological tagging. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 322–332

13. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP. (2013)
14. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation. (2002) 385–389
15. Dehdari, J., Tounsi, L., van Genabith, J.: Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic. In: Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, Dublin, Ireland, Association for Computational Linguistics (2011) 12–21

Többszintű szintaktikai reprezentáció kialakítása a Szeged FC Treebankben

Simkó Katalin Ilona[1], Vincze Veronika[2], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
kata.simko@gmail.com
rfarkas@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Napjainkban a két leggyakrabban használt szintaktikai reprezentációs elmélet a konstituens és a függőségi nyelvtan. A Szeged Treebank mondatai mindkét leírással manuális annotáltak. E cikkben beszámolunk egy olyan automatikusan átalakított, többszintű reprezentáció kialakításának munkálatairól, amely e két elemzés előnyös tulajdonságait egyesíti a mondatok szintaktikai leírásában.

1. Bevezetés

A létező szintaktikai elméletek közül jelenleg a két leginkább elterjedt a konstituens és a függőségi szintaxis. A Szeged Treebank mondatai is ezen reprezentációs elméleteknek megfelelően rendelkeznek manuális konstituens [1] és függőségi [2] elemzésekkel. Mindkét reprezentációnak megvannak az előnyei és a hátrányai is. A kétféle elemzés előnyeinek kihasználása céljából készül jelenleg automatikus átalakítással a Szeged Treebank leírására egy, a konstituens és függőségi fák, valamint a szavak morfológiai elemzéseit felhasználó, összetett szintaktikai reprezentáció. A reprezentáció kialakításakor hangsúlyozottan törekszünk arra, hogy a magyar nyelv szintaktikai sajátságait a lehető legnagyobb mértékben szem előtt tartsuk, ugyanakkor kiemelt szempontként kezeljük azt is, hogy a létrejövő treebank alkalmas legyen magyar nyelvű statisztikai szintaktikai elemzők betanítására is.

Ebben a munkában részletesen ismertetjük a többszintű szintaktikai reprezentáció kialakítása során követett irányelveket. Példákon keresztül megmutatjuk, hogyan kezelünk egyes nyelvi jelenségeket, valamint kitérünk arra is, hogy elemzésünk miben különbözik a Szeged Treebank eddigi változataiban követett függőségi, illetve konstituens alapú megközelítésektől, illetve szót ejtünk arról is, hogy elemzésünk hogyan viszonyul a szintén több nyelvi elemzési szinttel operáló LFG nyelvelméleti kerethez [3].

2. Konstituens és függőségi nyelvtanok

Bár a konstituens és a függőségi nyelvtanoknak is megvannak a hátrányai, mégis ezek a legelterjedtebben használt szintaktikai reprezentációk.

A konstituens reprezentáció a mondatokat összetevőkre bontja, amik összefüggő, jelentéssel bíró alkotóelemei a mondatnak. Tagmondatokra, azokon belül pedig igékre és bővítményeikre osztja a mondatokat. A szigorú konstituens elemzési elméletben az összetevők nyelvtani szerepére csak a szórendből következtethetünk, ami kötött szórendű nyelveknél, mint az angol jól működhet, de a magyar esetében kevésbé működőképes megoldás. A számítógépes nyelvészetben léteznek megoldások, amelyek az argumentumok felcímkézésével jelzik a nyelvtani szerepet, de ezek a konstituens nyelvtan szigorúan vett elméleti nyelvészeti hátterébe nem illenek bele. Nehezen elemezhetőek a nem folytonos konstituensek is, azaz azok az egybe tartozó elemek, amelyek nem egymás mellett jelennek meg a mondatban, mint például egyes mondatokban a genitív esetű birtokos és a birtoka.

Függőségi elemzésben a mondat szavai közvetlenül egymáshoz kapcsolódnak absztrakt csomópontok nélkül. Ezzel jól reprezentálhatóak a nyelvtani szerepek a mondatban és a nem folytonos összetevők kezelése is egyszerű feladat, elveszítjük viszont az összetartozó szavak egységként való kezelésének lehetőségét. Mindemellett a tagmondatok és mellérendelések kezelése például kevésbé intuitív, mint a konstituens elemzésben.

Mivel mindkét reprezentáció tartalmaz fontos információkat a magyar és a hasonlóan gazdag morfológiájú nyelvek szintaxisára vonatkozóan, nem eldöntött, hogy melyik a jobb leírás az ilyen nyelvek esetében. Hasonlóan, léteznek mind konstituens, mind függőségi elemzők a magyar nyelvre, melyek a Szeged Treebank különböző változatain lettek betanítva [4], azonban az automatikus elemzések kiértékelése során használatos mutatók sem teszik le egyértelműen a voksot egyik reprezentáció mellett sem. Ezen okokból döntöttünk egy olyan szintaktikai reprezentáció létrehozása mellett, amely egyesíti a két elmélet által kódolt információkat.

A Szeged Treebank mondatai kézzel annotált konstituens és függőségi elemzéssel is el vannak látva. A kétféle reprezentáció részben megegyező, részben az adott reprezentációnak megfelelő információkat kódol a mondat szintaktikai szerkezetével kapcsolatban. Ezeket az információkat egyesítjük egy új, többszintű szintaktikai leírásban.

3. Többszintű szintaktikai reprezentáció

A Szeged Treebank többszintű szintaktikai reprezentációja a lexikai funkcionális grammatika [3] elmélethez hasonló szerkezetű és a már létező, kézzel annotált konstituens és függőségi elemzések és morfológiai kódok felhasználásával jön létre. Az LFG-hez hasonlóan a különféle nyelvtani jellemzőket különböző szinteken jelenítjük meg.

A LFG reprezentációk több különböző struktúrát rendelnek a mondatokhoz. Ezek különböző szintaktikai szerkezeteken kívül szemantikai, fonológiai és egyéb nyelvi szintekhez kapcsolódó információkat is hozzákapcsolnak a mondat kifejezéseihez. A struktúrák egy többszintű reprezentáció alkotórészeit képezik ebben a keretben, egy-egy kifejezéshez a leírás több különböző szintjéről másmás információk társulnak és ezek együtt, egymással összekapcsolva alkotják az LFG elméletbeli reprezentációját az adott mondatnak.

Az LFG struktúrái közül a szintaktikai szempontból legalapvetőbb c- és f-struktúrák létrehozása mellett döntöttünk. A c-struktúra a mondat felszíni szerkezetét tükrözi, azt összetevőkre bontja. Az f-struktúrában a mondat argumentumszerkezete, illetve morfológiai információk jelennek meg attribútum-érték párokként. A két szerkezet szavai és nagyobb összetevői egymással összeindexelve, közösen alkotják ezt a többszintű modellt.

A magyar nyelv bizonyos jelenségeinek ebben a modellben való elemzéséről már nagyon sok cikk született [5,6], de a magyart általánosan leíró LFG nyelvtan legjobb tudomásunk szerint nem létezik. Jelen átalakítás alapelveinek lefektetésekor egy átfogó jellegű szabályrendszert igyekeztünk létrehozni, és a kisebb számban előforduló speciális nyelvi jelenségek kezelésére átvesszük a Szeged Treebank előző verzióiban kifejlesztett megoldásokat.

4. Átalakítás

4.1. C-struktúra

A c-struktúra átalakítása a Szeged Treebank konstituens elemzéséből indul ki. Ez az átalakítás viszonylag kevés módosítással jár. Megtartjuk a kézzel annotált frázisokat és hozzájuk adunk egy-egy indexet, ami összekapcsolja őket az f-struktúra megfelelő részeivel.

Így a konstituensnyelvtan előnye, az összetevős struktúra megmarad ebben az új modellben is, az ebben nehezen reprezentálható nyelvtani szerepek pedig más szinten vannak kezelve.

4.2. F-struktúra

Címkék. Az f-struktúra a mondat argumentumszerkezetét tükrözi. Ezen a szinten találhatóak a kifejezésekhez tartozó nyelvtani szerepek, és a nem folytonos összetevők elemzése is megoldható. Leginkább a függőségi nyelvtanban kódolt információval feleltethető meg, ezért a Szeged Dependencia Treebank és a mondatok szavaihoz rendelt morfológiai kódok átalakításával hozzuk létre.

Ezen a szinten a szintaktikai információ attribútum-érték párokból álló szerkezetben jelenik meg. Minden kifejezés f-struktúrájában megtalálhatóak a hozzátartozó releváns morfológiai adatok és a kifejezés különböző vonzatainak f-struktúrái. A függőségi nyelvtanban található relációk címkéit itt attribútumok címkéiként jelennek meg, az ezekhez kapcsolódó érték a kapcsolódó kifejezés f-struktúrája.

A mondat PRED jegye alatt megtaláljuk a fő elemet és a vonzatait zárójelben. A mondatok fő eleme a függőségi nyelvtan ROOT eleme, vonzatai a függőségi nyelvtanban hozzá csatlakozó szavak. A PRED jegy után a releváns morfológiai jegyek találhatóak, amelyeket a szavak morfológiai kódjából nyerünk.

Ezután a predikátum argumentumai következnek a nyelvtani szerepüknek megfelelő címkével. A függőségi nyelvtan SUBJ (alany) és OBJ (tárgy) relációi azonos nevű címkék lesznek az f-struktúrában. A kötelező vonzatok, a függőségi nyelvtanban DAT (részes eset) és OBL (egyéb eset) relációban állók egy közös, OBL címkét kapnak, míg a különböző határozói szerepű vonzatok (MODE, LOCY, FROM, TO, TLOCY, TFROM, TTO függőségi reláció) ADJ (adjunktum) címke alá kerülnek. Az INF, PA és AUX relációkkal rendelkező főnévi igenevek, melléknévi igenevek és segédigék szintén megtartják a függőségi relációjuk nevét az f-struktúra-beli címkéjükben.

A vonzatok f-struktúrája hasonló felépítésű: a PRED jegy az adott kifejezést jelöli, utána a vonzatait, módosítóit találjuk. Ezután a szófajának megfelelő morfológiai jegyek értékei következnek. A vonzatokat OBL vagy DAT függőségi relációval módosító, kötelező bővítmények itt is OBL címke alá kerülnek. Az ATT és MODE viszonyúak ADJ címkét kapnak. A névszókat módosító birtokosok POSS címkével kerülnek a birtok f-struktúrájába. A határozott és határozatlan névelők DEF=+ és DEF=- jegyekként jelennek meg a szerkezetben.

A névszói predikátumok függőségi PRED relációját az LFG elméletnek megfelelően [7,8] PREDLINK címkével jelöltük az f-struktúrákban. Ennek mintájára a többszavas névelemek kezelésére a függőségi NE viszonyt NELINK-ké alakítottuk, az összetett számnévi kifejezések NUM relációját pedig NUMLINK-ké.

Összetett mondatok. Az összetett mondatok kezelésében szintén az LFG-ben használt megoldást választottuk. Alárendelő szerkezetek és vonatkozói mellékmondatok esetén a főmondat PRED elemének egy vonzata az alárendelt mondat fő eleme, a beágyazott mondat f-struktúrája COMP címkével jelenik meg a főmondat f-struktúrájában. Mellérendelés esetén a mellérendelt kifejezések f-struktúrái egymás mellett jelennek meg. A kifejezéseket összekapcsoló esetleges kötőszavak alárendelés esetén az alárendelt mondat f-struktúrájában, mellérendelés esetén a mellérendelt tagok f-struktúrái alatt, CONJ-FORM címke alatt találhatóak.

Kötelező jegyek. Az f-struktúrában az egyes kifejezések alatt megtalálható kötelező morfológiai jegyeket az adott kifejezés morfológiai kódjából nyerjük ki. Az, hogy egy szónál milyen jegyeknek kell kötelezően megjelenni, a szó szófajától függ.

Az MSD kódban tárolt információk közül a szintaktikailag relevánsakat jelenítjük meg. Az ige altípusa, száma, személye, az igemód, igeidő és határozottság az ige f-struktúrájában jelenik meg. A névszói vonzatok esetében a szám és az eset jelenik meg kötelezően. Melléknevek esetén ezeken felül a fokozás, névmásoknál a személy.

Hely- és időhatározók. A Szeged Treebankben található három-három hely- és időhatározó típus megkülönböztetését az átalakított többszintű reprezentációba nem vettük át, mivel úgy gondoljuk, hogy ezen megkülönböztetés már túlmutat a szintaxis szintjén. Az irányhármasságot is kifejező hely- és időhatározói címkéket minden esetben ADJ jegyként kezeltük a mondatok f-struktúrájában.

A későbbiekben ezt az információt egy újabb struktúrába tervezzük felvenni, amelyben megtennénk ezt a szinte már szemantikai megkülönböztetést a hely- és időhatározók típusai között.

5. Virtuális csomópontok

A magyar LFG reprezentációjával kapcsolatban ugyanúgy felmerül a virtuális csomópontok problémája, mint a függőségi elemzésben. Mivel mindkét elmélet kerüli a fonológiailag jelen nem levő kifejezések megjelenítését a szintaktikai struktúrákban, a magyarban megjelenő kétféle virtuális összetevő kezelése nehézségeket okozhat.

A magyarban előforduló egyik ilyen meg nem jelenő összetevő a *van* ige harmadik személyű, kijelentő mód, jelen idejű alakja. A *Józsi katona* mondat esetén például nem jelenik meg az ige, ami más személy, mód vagy igeidő esetén már igen, például *Józsi katona volt*.

A másik típus az ellipszis, az a több nyelvre is jellemző jelenség, amikor egy már elhangzott szót vagy kifejezést nem mondunk ki újra, illetve a több tagmondatban ismétlődő kifejezéseket csak a tagmondatok egyikében szerepeltetjük. A ki nem mondott kifejezés lehet a tagmondat fő igéje, vagy annak bármely argumentuma, illetve az argumentum kisebb része. A *Józsi katona volt, Béla pedig pék* mondat esetén például a második tagmondatból a *volt* ige elliptálva van.

A virtuális csomópontok mindkét típusánál hasonló megoldás mellett döntöttünk. A virtuális kifejezések a mondathoz tartozó c-struktúrában nem jelennek meg, mivel az szigorúan a mondat felszíni szerkezetét rendezi frázisokba. Ezek a kifejezések csak az f-struktúrában jelennek meg, ami a szigorú LFG elméletben szintén kerüli a ki nem mondott kifejezések reprezentálását, viszont az ott megjelenített viszonyok leírásához fontos, hogy kitöltsük ezeket a csomópontokat is.

Az f-struktúrában a PRED jegyben jelöljük, hogy virtuálisról van szó: VAN vagy ELL értéket kap. A további jegyeket csak a VAN kapja meg, azok közül is csak azokat, amelyek biztosak: az igemód, igeidő és személy.

6. Eltérések az LFG-től

A Szeged Treebank átalakításakor főként az LFG elméletben [3] használt megoldásokat követtük, így a reprezentáció nagyon hasonló a lexikai funkcionális grammatika c- és f-struktúráihoz. Néhány ponton viszont eltértünk a szigorú LFG elmélettől. A következőkben ismertetünk néhányat ezen eltérések közül.

6.1. C-struktúra

Az LFG reprezentációk c-struktúrái a generatív nyelvtanokban használt bináris, X-vonás elméletnek megfelelő fákból állnak [9].

Az általunk átalakított c-struktúrák a Szeged Treebank konstituens fáihoz hasonlóan nem követik a szigorú chomskyánus nyelvtant, hanem a fő elem szófajának megfelelő frázisokra bontják a mondatokat.

6.2. Topik és fókusz pozíciók

Az LFG elemzésben a mondatok f-struktúrájában jelölve van a topik és a fókusz pozíció is, főleg a magyarhoz hasonló diskurzuskonfigurációs nyelvek szintaktikai leírása esetén.

A Szeged Treebank átalakítása során nem használtuk az f-struktúrában a topik és fókusz pozíciókat, mivel az erre vonatkozó információ sem a meglévő konstituens, sem a meglévő függőségi treebankben nincs kódolva, és így automatikus konvertálásuk nem megoldható. A topik és fókusz jelölése egy későbbi lépésben belekerülhet az f-struktúrákba kézi annotációval.

6.3. Fonológiailag üres névmási kategóriák

Bár az LFG kerüli az üres kategóriák felvételét az elemzésbe, pro elemek mégis megjelennek ki nem mondott névmások helyén az f-struktúrában. A magyarban gyakran ki nem tett személyes névmási alany és tárgy helyére például egy pro kerül az LFG elemzés f-struktúrájába.

Mivel a Szeged Treebank egyik verziója sem jelöli a fonológiailag üres névmásokat, az átalakítás során az ehhez hasonló esetekben nem vettük fel a pro PRED jegyű elemet, az ehhez tartozó jegyeket egy szinttel feljebb jelenítjük meg. Például egy elhagyott alany esetén annak száma és személye a magyarban megjelenik az igén, így ezeket a jegyeket ott reprezentáljuk ahelyett, hogy egy pro PRED jegyű alanyt vennénk fel az f-struktúrába ezekkel a jegyekkel.

7. A Szeged FC Treebank kialakítása

A fentiekben ismertetett elveket a gyakorlatba átültetve kialakítjuk a Szeged Treebank egy újabb verzióját, a Szeged FC Treebanket. Ezt elsődlegesen automatikus konverzió segítségével állítjuk elő a meglévő konstituens- és függőségi reprezentációk alapján, minimálisra csökkentve az utólagos kézi javításokat. A létrejövő új treebank kitűnő lehetőséget teremt arra, hogy létrehozzunk egy olyan statisztikai szintaktikai elemzőt, amely kifejezetten a magyar nyelv szintaktikai sajátságaira van optimalizálva, ugyanakkor egyesíti magában a konstituens és függőségi elemzők nyújtotta előnyöket is.

A Szeged FC Treebank kialakítása a Szeged Treebank konstituens és függőségi elemzéseinek automatikus konvertálásával történt a már leírt szabályok mentén. Az alábbiakban bemutatjuk egy példán keresztül az átalakítás különböző lépéseit.

A c-struktúrát a konstituens fákból egyszerűen a nyelvtani szerepjelölések eltávolításával nyertük, l. 1. és 2. ábrák.

Az f-struktúra és a függőségi nyelvtan között már nagyobb különbség látható, vö. 3. és 4. ábrák. A példamondatban az alá- és mellérendelő szerkezeteken kívül a birtokos szerkezetek kezelése is látható a két különböző elméleti keretben.

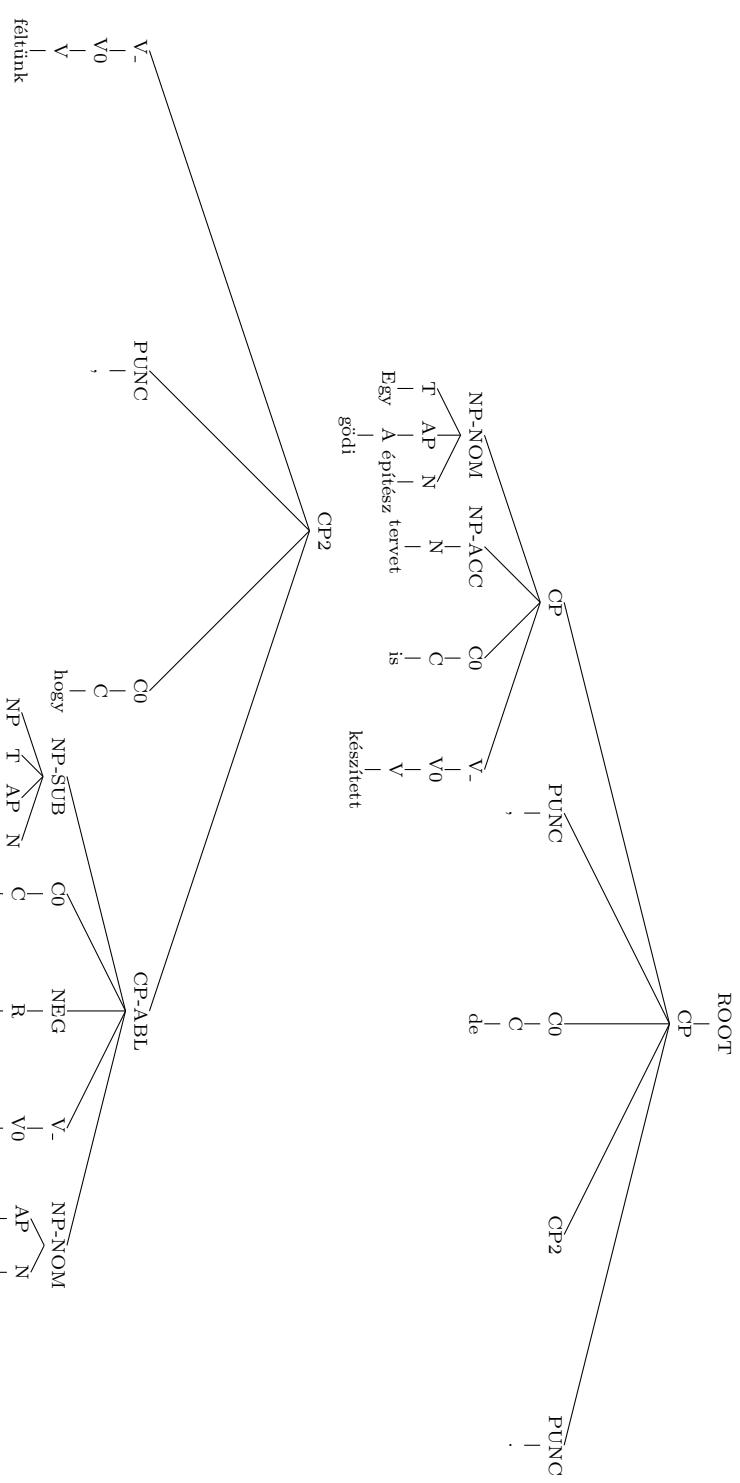
A Szeged FC Treebank reprezentációi a Szeged Korpusz mondataihoz a fent láthatóakhoz hasonló c- és f-struktúrákat rendelnek. Ez a két leírás együtt alkotja az új treebank elemzését.

8. Összegzés

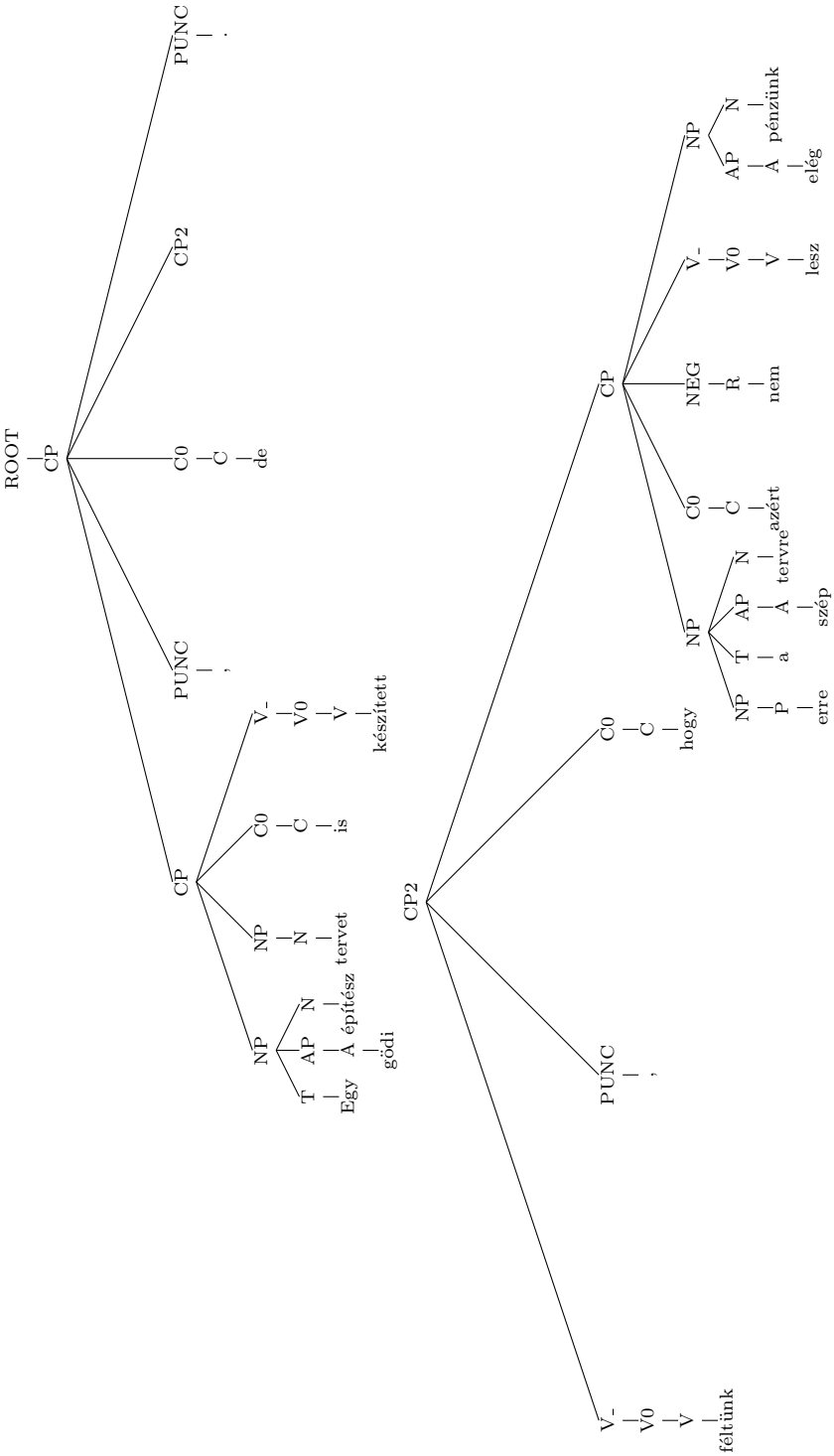
Ebben a munkában bemutattuk a készülő Szeged FC Treebank elméleti alapját képező többszintű szintaktikai reprezentációt, mely egyesíti magában a konstituens és függőségi reprezentációk előnyeit, ugyanakkor kifejezetten a magyar nyelv szintaktikai sajátságaira van szabva. Az LFG elméletéhez hasonlóan, e reprezentáció is c és f-struktúrában jeleníti meg a releváns szintaktikai információkat, azonban attól néhány fontos vonásban eltér. Az újonnan létrejövő treebank reményeink szerint egy új, a magyar nyelvet minden eddiginél hatékonyabban feldolgozni képes statisztikai szintaktikai elemző létrehozásának alapjául szolgálhat.

Köszönetnyilvánítás

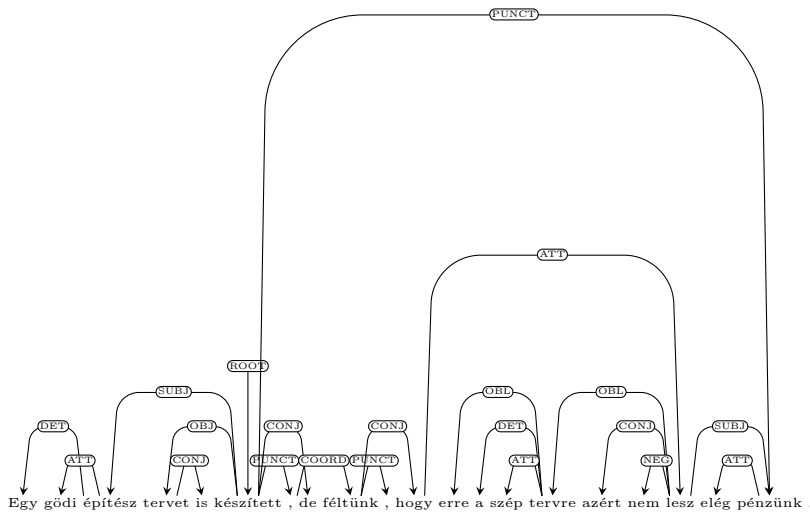
A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.



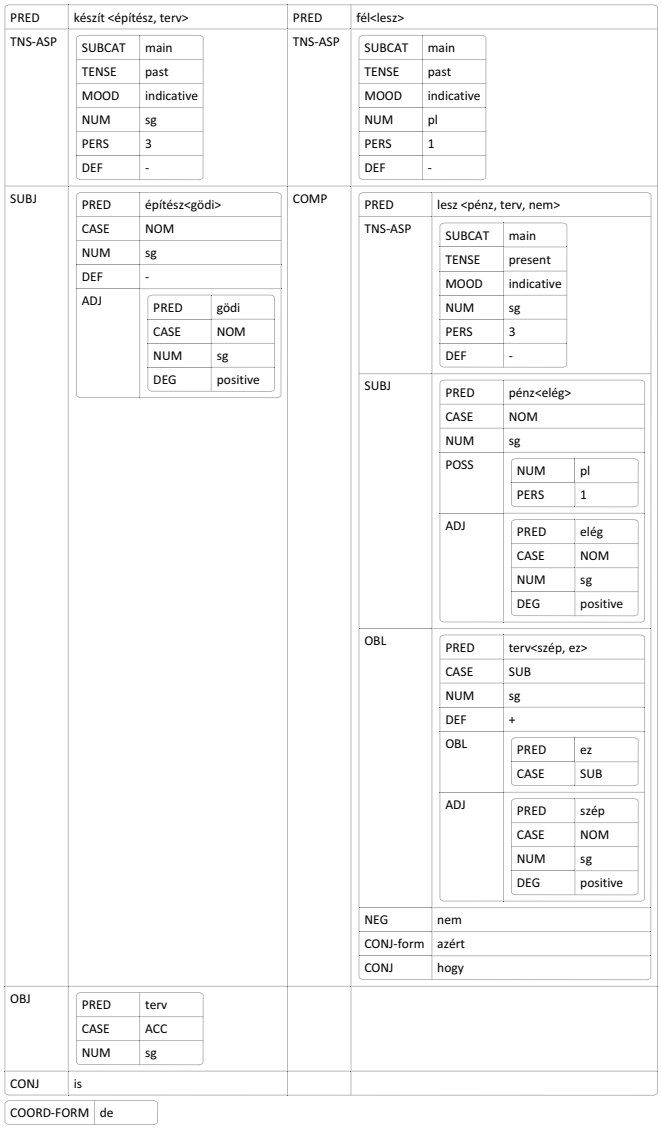
1. ábra. Konstituensnyelvtan szerinti reprezentáció.



2. ábra. C-struktúra.



3. ábra. Függőségi fa.



4. ábra. F-struktúra.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged TreeBank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005. Lecture Notes in Computer Science, Berlin / Heidelberg, Springer (2005) 123–132
2. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
3. Bresnan, J.: Linear order, syntactic rank, and empty categories: On weak crossover. In Dalrymple, M., Kaplan, R.M., Maxwell, J.T., Zaenen, A., eds.: Formal Issues in Lexical-Functional Grammar. CSLI Publications, Stanford, CA (1995) 241–274
4. Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A.: Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA, Association for Computational Linguistics (2013) 146–182
5. Laczkó, T.: Grammatical Functions, LMT, and Control in the Hungarian DP Revisited. In Butt, M., King, T.H., eds.: The Proceedings of the LFG '04 Conference, University of Canterbury (2004)
6. Rákosi, Gy., Laczkó, T.: Inflecting Spatial Particles and Shadows of the Past in Hungarian. In Butt, M., King, T.H., eds.: The Proceedings of the LFG 2011 Conference, Hong Kong (2011) 440–460
7. Butt, M., Niño, M., Segond, F.: A Grammar Writer's Cookbook. CSLI Publications, Stanford, CA (1999)
8. Attia, M.: A Unified Analysis of Copula Constructions in LFG. In Butt, M., King, T.H., eds.: The Proceedings of the LFG '08 Conference, University of Sydney, Australia (2008) 89–108
9. Chomsky, N.: Lectures on Government and Binding. Dordrecht, Foris (1981)

Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé

Prószéky Gábor[1,2,3], Indig Balázs[2], Miháltz Márton[1], Sass Bálint[1]

[1] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[3] MorphoLogic
e-mail: {proszeky.gabor, indig.balazs, mihaltz.marton, sass.balint}@itk.ppke.hu

Kivonat Cikkünkben egy, az eddigi megközelítésektől jelentősen eltérő nyelvelemző rendszert ismertetünk, mely a következő alapelvek szem előtt tartásával készül. (1) A *pszicholingvisztikai* indíttatás azt jelenti, hogy amennyire csak lehetséges, az emberi nyelvfeldolgozás mintájára alakítjuk ki a modellt. (2) *Performanciaalapú* rendszerként minden olyan nyelvi megnyilatkozást megpróbálunk feldolgozni, ami (leírt szövegekben) előfordul, nem helyezünk hangsúlyt az elméletileg létező, de a gyakorlatban meglehetősen ritka jelenségek kezelésére. Ugyanakkor bármilyen – rosszul formált, agrammatikus – szöveget igyekszünk nyelvi megnyilvánulásnak tekinteni és értelmezni. (3) Szigorúan *balról jobbra*, szavanként dolgozzuk fel a szöveget. A még el nem hangzott elemeket teljes mértékben ismeretlennek tekintjük, rájuk nem hivatkozunk. (4) Az elemző architektúrája eredendően *párhuzamos*. A hagyományos megközelítéssel szemben, ahol az elemzések egy láncot alkotó modulsor végén alakulnak ki, itt az éppen elemzendő szót folyamatosan, párhuzamosan jelen lévő szálak (morfológiai elemző, korpuszgyakorisági szál stb.) egyszerre vizsgálják és együttesen, egymással kommunikálva, egymás hibáit javítva határozzák meg az elemzést. (5) Nem a mondatot, hanem az akár több mondatból álló *megnyilvánulást* tekintjük reprezentálandó alapegységnek, lehetővé téve a mondaton belüli és mondatok közötti anaforikus viszonyok egységes kezelését. (6) Ennek megfelelően, illetve a különböző jelenségek egyidejű kezelése miatt a *reprezentáció* nem feltétlenül fa, hanem egy akár különböző típusú éleket tartalmazó összefüggő gráf. Az elvi megalapozást követően az elemző pilot megvalósítását is bemutatjuk. A pilot program az alapelveket szemlélteti és emellett néhány, az ismertetett elveknek megfelelő elemzési lépést is elvégez.

1. Bevezetés

A nyelvészet utolsó évtizedeiben egyeduralkodónak mondható generatív modellek informatikai szempontból nem igazán nyújtanak hatékony megoldást a valóságban előforduló, azaz a nem feltétlenül tökéletesen szerkesztett szövegek elemzésére. Ennek az is az oka, hogy a Chomsky [1] által bevezetett és az ezt követő

generatív technikákban a transzformációk nem invertálhatók, de ez nem lehet a fő ok, hiszen léteznek Chomskyétól eltérő, transzformációmentes generatív modellek is. Ám mindegyik esetében igaz, hogy ezekben a modellekben a „hatékony elemezhetőség" nem a generatív közelítésben preferált kompetencia, hanem a performancia érdeklődési körébe tartozik. A *performanciaalapúság* számunkra azt jelenti, hogy minden nyelvi megnyilatkozás feldolgozandó, ami „előfordul"; viszont ami elvben ugyan lehetne, de valójában nem fordul elő, az valamilyen értelemben kevésbé lényeges. Az emberi nyelvfeldolgozás a nyelvi megnyilatkozással egy időben – ha tetszik: balról jobbra – halad, és igyekszik minden olyan információt felhasználni, mely a megnyilatkozás értelmezéséhez szükséges, még akkor is, ha az – a hagyományos grammatikai értelemben – nem feltétlen tökéletesen szerkesztett. Nincs tehát mód a megnyilatkozások még el nem hangzott, vagy le nem írt részére hivatkozni, azaz legfeljebb feltételezni, valószínűsíteni lehet bizonyos még meg nem jelent összetevőket a már elhangzottak, leírtak alapján, egészen addig, míg a megnyilatkozás be nem fejeződik. Ez nem jelenti azt, hogy nem léteznek olyan megnyilvánulások, amelyek a legvalószínűbbnek tűnő elemzési megoldást „kijátszva", olykor visszalépéses működésre kényszerítik az emberi elemzőt is, ám ezeket úgy tűnik, hogy a hétköznapi kommunikációban a grice-i maximák [2] betartásából következően a kommunikációban kerüljük, és inkább csak viccek, vagy szándékos félrevezetés alkalmával fordulnak elő. Ennek a bizonyítására nagyméretű szövegkorpuszokat kezdtünk építeni[4].

2. Elméleti háttér, összevetés más rendszerekkel

Az elemző architektúrájának kialakításához először megvizsgáltuk a legfontosabb performanciaalapú elemzőket [4], továbbá azokat az elméleti közelítéseket is, melyek a hatékony elemezhetőséget a kompetencia körébe sorolják [5], és azt láttuk, hogy a ma ismert számítógépes mondatelemzők szinte kizárólag egyirányú feldolgozást végeznek, azaz nincs oda-vissza kapcsolat a különböző nyelvi szintek között. Ez a hibák felhalmozódásához vezet, amire általában egy egyszerű gyakoriságon alapuló szűrő a gyakorlati megoldás. A megvalósítandó analitikus grammatikai (a továbbiakban: AnaGramma) elemző viszont párhuzamos szálakon többféle nyelvi elemzést indít, melyekkel egyidejűleg jelennek meg más, a feldolgozandó szöveghez kapcsolható jelentést és világismeretet kezelő szálak. Elemző algoritmusunk tehát egyfajta konszenzust keres a különböző „tudások" között [6]. Amint tehát a humán információfeldolgozásban, a mi elemzőnkben is egyidejűleg és szorosan működnek együtt a nyelvi elemzést és az értelmezést végző modulok (amik a valóságban egy-egy agyi területnek felelnek meg [7]).

Mivel a tervezett reprezentáció legközelebb a függőségi leírásokhoz áll, megvizsgáltuk a hagyományos, kompetenciaalapú világ különböző, létező, hatékony

[4] Szeretnénk a kialakítandó elemző „súlypontját" is a megfelelő helyre tenni, ezért a nagy korpuszok építésére és feldolgozására irányuló kutatásunk egy másik célja a modern grammatikaelméletek által sokat vizsgált, sokszor igen bonyolult – de a hétköznapi életben meglehetősen ritka – nyelvi szerkezetek előfordulási gyakoriságainak vizsgálata. [3]

függőségi elemzőit is, amilyen például a MaltParser [8], a Stanford Parser [9], vagy a véges állapotú függőségi elemző [10]. Ezek valóban a nyelvi egységek egymás közötti viszonyainak leírását célozzák meg, de olyan erősen kötődnek az egymás után következő mondatok szeparált feldolgozásához, hogy nem találtuk őket közvetlenül felhasználhatónak. A magyar nyelvre egyébként történtek korábban függőségi megközelítések, mind szabályalapúak, mint például a holland DLT rendszerhez készített nyelvtan [11], mind adatorientáltak, mint a Szeged Treebank függőségifa-formátumú változata [12]. Ami viszont a magyar nyelvi jelenségek leírását illeti, az eddig készített legátfogóbb magyar mondatelemző, a *MetaMorpho* fordítórendszer magyar nyelvi elemzőjének szabályrendszere is rendelkezésünkre áll [13], bár az nem a függőségi leíráson alapul. Az összes fenti elemző közös tulajdonsága, hogy ezek egyike sem kezeli megfelelően a többértelműségek feloldását, és meglehetősen rossz a hibatűrésük.

Mint Prószéky [14] utal rá, a nyelvi szerkezetek elemzés közbeni kiválasztása közben hozott döntéseink felül tudják bírálni a lexikont. A korábban kialakított nyelvi ismereteket összegző szótárakat és az eddig leírt szintaktikai szerkezeteket adatbázisként használó szabályalapú elemzők és az egyes szerkezetek korábbi gyakoriságára építő valószínűségi elemzők [15] kizárólag csak a „múltbéli" ismeretekre, múltbéli statisztikákra alapozva tudják meghozni döntésüket. Az AnaGramma-elemzésben egy olyan megoldást szándékozunk megvalósítani, melyben az aktuális bemenet esetleges szokatlan felépítését sem a korábbi statisztika támogatásának a hiánya, sem a mechanikusan alkalmazott szabályok sokszor félrevezető elemzési kimenete nem „zavarja meg". Kiinduló hipotézisünk az, hogy a nyelvhasználó fejében két rendszer él: egy a tanult szerkezetekre építő és egy aktuális döntéseket hozó, mely az elhangzó nyelvi elemek valós idejű feldolgozását akkor is képes megvalósítani, ha a „megtanult" szerkezetek egymásnak ellentmondó (például egymáshoz nem illeszkedő jegyszerkezeteket tartalmazó) nyelvtani információkat hordoznak.

A felsorolt eszközök egyike sem kezeli helyesen a többértelmű szerkezeteket és rossz a hibatűrésük, így az újraírószabály-alapú rendszerekben egyetlen nem ismert szó, vagy egy szokatlan, a rendszer számára ismeretlen fordulat az egész elemzés kudarcát okozhatja. A jelenleg kialakítás alatt álló és az elemzésre, mint elsődleges feladatra összpontosító AnaGramma ezzel szemben

1. egyidejűleg több szálon, időben monoton halad (a valódi emberi feldolgozást jobban közelítve), gyakorlatilag visszalépés nélkül (de ennek lehetőségét nem zárja ki)[5];
2. nem tárol „fölöslegesen hosszú ideig" később nem használandó elemzési ágakat (de ez a „hosszú idő" persze szerkezetenként nagyon különböző lehet);
3. mindezekkel együtt, illetve mindezek ellenére: az emberi információfeldolgozáshoz hasonlóan (ha azzal nem is összemérhető mértékben) gyors; és
4. a „hiba" fogalmát nem ismeri, vagyis csak az aktuálisan adott toleranciaszint (ami egy külső paraméter) szerint kezelhető elemei vannak (ezáltal dolgozhatóak föl a helyesírási hibák, az agrammatikus szerkezetek, a szokásos emberi hibák, esetleg beszélt nyelvi átiratok, vagy a nem-anyanyelvűek szövegei is).

[5] ezért egy inkrementálisan balról jobbra haladó elemzőt készítettünk kiindulásként

Elemzőnkben tehát szakítunk a hagyományos „pipeline" architektúrával. A legfőbb ok, hogy a hagyományos architektúrákban az alacsonyabb szinteken képződött hibák javítás nélkül kerülnek át magasabb szintekre és felerősödnek, ezzel rontva a későbbi modulok kimenetének minőségét. Az AnaGrammában több feldolgozási szint (pl. morfológia, igei szerkezetek felismerése, korpuszgyakoriságok, ontológiák és világismeretek) párhuzamosan működnek külön-külön *erőforrásszálként* kiegészítve vagy éppen felülbírálva egymást, minden egyes elemzési lépésben. Az alapelveinkből az is következik, hogy az elemzés folyamán nem használhatunk olyan tradicionális értelemben vett POS-taggert, ami globális információ felhasználásával dönt a mondat minden eleméről. Ehelyett egy olyan n-gram modellt használunk, ami ugyan rendel valószínűségeket az aktuális szóhoz kapcsolható címkékhez, ám csupán az elhangzott, illetve leírt, az aktuális pozíciótól tehát balra álló, azaz a *megelőző* szavak alapján.

Megvizsgáltuk azt is, hogy mely nyelvi elemek indítanak el szövegértelmezés közben valamilyen literális vagy kategoriális predikciót. Néhány ilyen „üzenet" részletes elemzése alapján arra jutottunk, hogy a lehetséges alternatív ágak egyikén-másikán néhány lépés után nem folytatódik az elemzés. Megjegyezzük, hogy bár ez a jelenség a hagyományos táblázatos elemzők [16] világából ismert, azok nem tesznek különbséget a szerkezetek közt aszerint, hogy ezek közül melyik mennyire tipikus, vagy épp mennyire ritka. A tervezett elemző az emberi nyelvfeldolgozás hatékonyságából kiindulva igyekszik elkerülni a kombinatorikus robbanást is, ezért használja az előismeretek összegzéseként kialakított statisztikát: a gyakori szerkezetek sokszor elemzés nélkül, kész belső szerkezettel jelennek meg a feldolgozásban. Informatikai szakszóval ezt gyorsítótárazásnak (angolul cache-elésnek) mondanánk, ám a jelenség a pszicholingvisztikában is jól ismert, és az emberi nyelvértelmezés esetében ezt *egészleges feldolgozásnak* nevezik.

Az eddig megvalósított kompetenciaalapú modellek a nem nyelvi információfeldolgozó alrendszerekkel „természetüknél fogva" semmilyen együttműködést nem feltételeznek. Kijelenthetjük viszont, hogy a performancia nem választható el más kognitív folyamatoknak a nyelvre gyakorolt hatásától, ezért az első elemzési lépéstől kezdve az AnaGramma-módszer a nyelvi, és a modell kidolgozottságától függően bizonyos nyelven kívüli modulok (világismeret, hangulat stb.) párhuzamos kezelésére épít. Ráadásul, a szokásos megoldásoktól eltérően, nem egyes mondatokat, hanem teljes „megnyilvánulásokat" (egy gondolategységet átfogó, általában bekezdésnyi szövegeket) dolgozunk fel, hiszen egy-egy konjunktív elem jelenléte vagy hiánya nem okozhatja az azonos tartalom felszíni különbségek miatti radikálisan különböző feldolgozását, pusztán a mondathatárok különbözősége miatt.

Az egyes mondatok reprezentációi nem a szokásos fastruktúrákban képzelendők el, hiszen a részszerkezetek teljes összekapcsolása nem feltétlen egyetlen mondaton belül valósul meg, továbbá a referenciális elemek is ugyanezen reprezentációban megjelenő, de a hagyományos generatív felfogástól eltérő éleket vezetnek be a leginkább a függőségi leírásra hasonlító AnaGramma-reprezentációkba. A mondatok egyes részeinek referenciális alapon való összekötése (vonatkozó név-

mások, visszautalások kezelése stb.) egy sajátos összefüggő gráfot eredményez[6]. Kimenetként nem pusztán szintaktikai, hanem szemantikai jellegű információkat is szeretnénk megkapni: az elemző célja beazonosítani az összes szereplőt és eseményt, meghatározva a szükséges koreferencia-viszonyokat is. A rendszer végül is létrehoz egy olyan, a mondatot, illetve a bekezdést reprezentáló összefüggő irányított gráfot, amelynek segítségével válaszolni tud majd az olyan kérdésekre, hogy például ki, mit csinált, hol és mikor. Egy ezen az elven a gyakorlatban is működő pilot megoldás jelenleg a következő magyar nyelvi jelenségeket képes kezelni: elváló igekötő, birtokos szerkezet, tagadás, felsorolás (tekintetbe véve, hogy felsorolás tagjai csak valamilyen szempontból egységes elemek lehetnek), értelmező, illetve a vessző írásjel funkciójának meghatározása (azaz, hogy felsorolásra, közbevetésre, vagy értelmezőre utal-e).

3. Az architektúráról

Az elemző balról jobbra halad végig a nyelvi elemeken, amik a mi jelenlegi megvalósításunkban a szavak. Feldolgozza tehát a soron következő szót, tekintetbe véve az összes futó szál által szolgáltatott információt, majd (a) lezár, (b) elindít vagy (c) változatlanul hagy szükséges szálakat. Ha több szabály illeszkedik egy elemre (pl. egyszerre valaminek a birtoka is, és valamilyen esetben is áll), akkor az összes illeszkedő szabályhoz tartozó *strukturális szálaknak* el kell indulniuk. Ezek a lépések együtt határozzák meg, hogy az adott elemnél mi történjen: valamilyen típusú szál induljon, záruljon le, vagy él keletkezzen két elem között a reprezentációban. A szabályokat egyébként a prototípusban még kézzel „gyártottuk", de a későbbiekben elsősorban a statisztikai feldolgozások kimenetén megjelenő minták segítségével hozzuk őket létre, illetve – mint már említettük – felhasználunk rendelkezésre álló nyelvtani adatbázisokat is (ilyen például a MetaMorpho szintaktikai mintáinak egy része).

Alapvetően kétféle, nyelvi elemek által indítható száltípus látszik szükségesnek. A *felkínálás* jellegű szál információt ad az adott elemről (pl. alanyesetű), míg az *igény* jellegű szál keres egy adott tulajdonságú elemet vagy szálat. Például a birtok igényel egy alanyesetű vagy datívuszos alakot, a névutó egy alanyesetű (vagy megfelelő raggal ellátott) alakot, a névelő az esetragos NP-fejet, a tárgyas ige a tárgyat, amire (mindenképpen) szüksége van. Azt állítjuk, hogy a különböző nyelvi aspektusokat figyelő szálak együttműködésének mellékhatása a morfológiai egyértelműsítés és a kombinatorikus robbanások megelőzése, mely utóbbi jelenség a szabályalapú rendszereknél gyakran felmerül a hosszabb mondatok feldolgozása folyamán, akár még morfológiailag egyértelműsített tokenek esetén is.

A tervezett kimenet a feldolgozott szövegből épített szintaktikai-szemantikus relációk hálózata, ami alapján egy lekérdező rendszer meg fog tudni válaszolni

[6] Projektünkben a fentiek szellemében megindult a fent említett (automatikus) korpuszépítés, a főnévi csoportok és mondatvázak mintázatainak (reguláris) szövegkorpuszokban való vizsgálata [3], az új elemző architektúrájának kialakítása, a reprezentációépítés, sőt, az igevonzatok automatikus szemantikus kategorizálása is [17].

olyan kérdéseket, melyek csak implicite vannak benne az eredeti mondatban. Jelenleg a mondatban felismert és azonosított függőségi relációkból épült egységek fája készül el, kiegészülve azokkal a szemantikai jellegű információkkal, amelyeket az elemzés során a szövegből nyertünk.

Reprezentációnkban előnyben részesítjük a nem fa formájú, hanem általában DAG-formájú függőségi gráfokat. Mivel a *koreferenciák* nem feltétlenül azonos mondatban jelennek meg, így az általuk bevezetett élek „színe" más, így nem tudják elrontani a szerkezetet, azaz miattuk nem kaphatunk irányított, körmentes gráfot eredményül.

Az *ellipszis* jelenségek kezelése miatt megengedjük a szálaknak, hogy túlléphessenek a mondathatáron. Úgy véljük, hogy az elemzésnek nem szabad megállnia a mondatok végén, mert az egymagukban álló mondatokkal szemben a hosszabb megnyilatkozások az emberi kommunikáció természetes egységei. Az egymást követő mondatok témája sokszor azonos, ezért a természetes emberi kommunikáció során lehetséges – és többnyire meg is történik – az egyes elemek kihagyása (az ellipszis jelenség), ami a legtöbb hagyományos elemzőnél komoly problémákat okoz. A rendszerünk által feldolgozandónak szánt nyelvi egységek néhány mondatból álló összefüggő szövegek, ám a sok mondatból álló, nagyobb művek feldolgozását egyelőre nem szándékozzuk megcélozni.

Nagyon fontos számunkra a szövegben előforduló események szereplőinek azonosítása, és a *koreferenciaviszonyok* meghatározása, más szóval annak meghatározása, hogy mely szereplők azonosak a világban („ki kicsoda?"). Más szóval, szeretnénk helyesen kezelni, hogy mely szereplő „új" a szöveg egy adott pontján való megjelenésekor, és mely nyelvtani elem utal egy korábban már megjelent szereplőre, illetve van-e, és ha igen, milyen kapcsolata a korábbiakkal. Lényegében szereplőnek tekinthető az összes névszó, az igeragokból kikövetkeztethető alanyi, tárgyi szereplők, sőt, Davidson nyomán a *neo-davidsoniánus eseményszemantika* [18] elveinek megfelelően maguk az események (azaz az igék) is, mivel vissza tudunk utalni rájuk. Az azonosítás érdekében a korábbi mondatokat és minden korábbi szereplőt folyamatosan nyilvántartunk.

További szálak hasznosítják a *lexikai egységek* és *lexiko-szintaktikus szerkezetek* gazdag leírását, amit a *MetaMorpho* elemző adatbázisait felhasználva építettünk fel. Például egy „felkínálás" típusú szál alapvető szintakto-szemantikus jellemzőikkel (élő, ember, absztrakt stb.) annotálja az egyes egységeket a rendelkezésre álló, mintegy 118 000 szót és többszavas kifejezést tartalmazó adatbázisból. Egy „igény" szál pedig a *MetaMorpho* 35 000 darabos nyílt konstrukciós szabályhalmazából kapcsolatokat javasol az igék, főnevek és melléknevek a lehetséges argumentumai között (például: *eszik valamit, ellenségesség valamivel szemben, érdeklődés valamivel kapcsolatban*). Ezeken túl kísérletezünk még az igei szerkezetben megjelenő argumentumok predikciójával, mely a korpuszbeli adatok (együttes előfordulás, gyakoriság), az ontológiai információk [19] és a lexikon (a *MetaMorpho* elemző igei szerkezetek adatbázisa) információira támaszkodik. Építünk még a *Mazsola* projektből [20] származó ige-főnév együttes előfordulások adatbázisára, és azon is dolgozunk, hogy össze tudjuk kapcsolni

őket a *Magyar WordNettel*, hogy általánosított szemantikai osztályokat találjunk az igei szerkezetek szemantikus szelekciós megszorításai között [17].

4. A rendszer működésének alapelveiről

Új elvű nyelvi elemzőnk kialakításának első lépéseként azonosítottuk a feldolgozás során használni kívánt formális utasítástípusokat. Meg kellett találnunk azokat az elemeket, amik meghatározzák, hogy milyen fajta elemek jöhetnek utánuk. Például a névelőt követő főnévi csoport végén valahol egy főnévnek, egészen pontosan valamilyen főnévi szerepű elemnek kell állnia. Előbb-utóbb megjelennek a szövegben olyan elemek, melyek „kielégítenek" egy korábbi „igényt". Például az igei argumentumok kitöltenek egy helyet a már korábban látott ige vonzatkeretében. Ha az ige maga valamely argumentuma után jön, akkor a korábban megjelent argumentumok – ha megfelelő jegyeik kompatibilisek – automatikusan kitöltik a szerkezetet a megfelelő módon.

Vannak azonban a fentiektől eltérő, más típusú műveletek is: a konjunkciós szerkezetek például csak akkor azonosíthatóak, ha egy konjunktív elem ténylegesen feltűnik. Ez lehet „és", „vagy" vagy épp egy erre szolgáló vessző, mert ezek vezetik be a konjunktív szerkezet következő tagját. Ha a rendszer felismer egy ilyen elemet (de csak akkor!), módosítania kell az utolsóként feldolgozott elem reprezentációját a felismert szerkezetnek megfelelően, hiszen az előző elem volt ennek a konjunktív szerkezetnek az első tagja, amit az előző lépésben, annak feldolgozásakor még nem tudhattunk róla. A konjunkciót egyébként egyetlen egységként kezeljük, anélkül, hogy állást foglalnánk arról, hogy van-e az ilyen szerkezeteknek feje. Jelenleg épp az ilyen, exocentrikus szerkezetekre vonatkozó műveletek balról jobbra történő feldolgozásának formalizálásán dolgozunk.

Pilot implementációnkban megpróbálunk kezelni néhány olyan gyakori, alapvető jelenséget, amiket nem feltétlenül egyszerű kezelni más keretrendszerekben. Ilyenek például

- az elváló igekötő és az igető, illetve a birtokos szerkezetek részeinek összekapcsolása,
- a felsorolások/koordinációk (amik azonos típusú elemekből állnak) komplex egységként való felismerése,
- a vessző szerepének felismerése aszerint, hogy mit vált ki: mellékmondatot, felsorolást, zárójeles kifejezést/közbevetést vagy értelmezőt,
- a tagadás hatókörének felismerése.

Elemzőnk elkészült prototípusát *újsághírek összefoglalóin* teszteltük, melyeket a `www.inforadio.hu` RSS csatornájáról töltöttünk le. A két-három mondat hosszúságú hírek általában egyetlen politikai vagy gazdasági eseményt írnak le. Nyelvi komplexitásuk közel áll ahhoz, amit modellezni szeretnénk, ezért megfelelő bemenetül szolgálnak az elemző számára. Először a bemeneti szöveget előfeldolgozásként lemmatizáltuk (ezzel mintegy modelláltuk a flektáló nyelvek toldalékolt alakjainak „szótári lookup" jellegű kezelését). A morfológiai többértelműségek ezen a szinten természetesen meg kell, hogy maradjanak, mert – mint

korábban említettük – nem használhatjuk a jól ismert egyértelműsítő eljárásokat, mivel azok általánosságban megsértik a monoton balról jobra haladó elemzést. Néhány rövidhír részletes elemzése alapján arra jutottunk, hogy egyfajta dinamikus (azaz az aktuális szó kategóriájától függő) előrenéző stratégiát érdemes használnunk, ugyanis a legtöbb alternatív elemzési ág gyorsan befejeződik, mert a különféle szálak nem engedik folytatódni őket néhány lépés után. Megjegyezzük, hogy ez a jelenség jól ismert a hagyományos táblázatos elemzőknél, de azok nem képesek különbséget tenni a különböző struktúrák között azok tapasztalati gyakorisága alapján. Mi ezeknél a döntéseknél állandóan tekintetbe veszünk egy olyan korpuszgyakorisági szálat, mely a háttérben fut és egy nagy korpusz adataira támaszkodik. Ez tájékoztat arról, hogy a meglévő elemzés mennyire felel meg a szokásos mintázatoknak, illetve döntési helyzetben segít választani több lehetséges alternatíva közül. Mindig csak a (balról jobbra) soron következő szót értékeljük ki, figyelve, hogy milyen gyakorisági viszonyban áll az eddigiekkel. Például az *esik* alak után alanyként a *szó* meglehetősen jól elfogadható, mert kb. 15%-ot képvisel az *esik* mellett. Ha viszont ezen a lépésen is túl vagyunk, akkor már nagyon várjuk a *-rÓl* ragos alakot, mivel az 90%-os valószínűségű az *esik szó* kifejezés esetében.

Fontos, hogy a rendszer kategóriáinak kialakításánál csak a szükséges általánosításokat tegyük meg. Például adott esetben létrehozhatunk olyan – a hagyományos nyelvtani kategóriáktól eltérő – szófajt, amely adott esetben egyetlen kivételes szót (pl. *is*) tartalmaz, vagy dönthetünk úgy, hogy az alany- és a birtokosesetet a többi esettől teljesen elkülönítve, új néven kezeljük.

5. Összefoglalás

Kutatásunk egy pszicholingvisztikai motivációjú, performaciaalapú, párhuzamos feldolgozást végző nyelvi elemzőt céloz meg. Megpróbáltuk összegyűjteni az ehhez a működéshez szükséges ismereteket az irodalomból, de azt találtuk, hogy az emberi nyelvfeldolgozás általunk vizsgált aspektusait egyetlen ma működő elemző sem elégíti ki megfelelően, így lefektettük egy új elképzelés, az AnaGramma alapjait. A kidolgozott elvek működtetéséhez első lépésként egy minimális képességű, de a működés alapjait mégis bemutatni képes pilot programot is készítettünk, melynek forráskódja megtalálható az alább internetes oldalon:
`https://github.com/ppke-nlpg`.

Köszönetnyilvánítás

Köszönjük a TÁMOP-4.2.1.B – 11/2/KMR-2011–0002 és a TÁMOP: 4.2.2/B – 10/1–2010–0014 projektek részleges támogatását.

Hivatkozások

1. Chomsky, N.: Syntactic structures. The Hague:Mouton (1957)
2. Grice, H.P., Harman, G.: Logic and conversation. Encino:Dickenson (1975)
3. Endrédy, I., Novák, A.: Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve. A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai (2013) 297–301
4. Bunt, H., Merlo, P., Nivre, J.: Trends in Parsing Technology. Dordrecht:Springer (2010)
5. Pritchett, B.L.: Grammatical competence and parsing performance. University of Chicago Press (1992)
6. Pléh, Cs.: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1999)
7. Csépe, V.: Az olvasó agy. Akadémiai Kiadó, Budapest (2006)
8. Nivre, J.: Inductive dependency parsing. Springer (2006)
9. De Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC. Volume 6. (2006) 449–454
10. Oflazer, K.: Dependency parsing with an extended finite-state approach. Computational Linguistics **29**(4) (2003) 515–544
11. Prószéky, G., Koutny, I., Wacha, B.: A dependency syntax of Hungarian. Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation) (1989) 151–181
12. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: LREC. (2010) 1855–1862
13. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: A robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation (2004) 138–142
14. Prószéky, G.: Számítógépes morfológia. In Kiefer, F., Bánréti, Z., eds.: Morfológia (Strukturális magyar nyelvtan III). Volume 3. Akadémiai Kiadó, Budapest (2000) 151–1064
15. Brants, T., Crocker, M.: Probabilistic parsing and psychological plausibility. In: Proceedings of the 18th conference on Computational linguistics-Volume 1, Saarbrücken:Association for Computational Linguistics (2000) 111–117
16. Révész, G.: Bevezetés a formális nyelvek elméletébe. Akadémiai Kiadó, Budapest (1979)
17. Miháltz, M., Sass, B., Indig, B.: What do we drink? Automatically extending Hungarian WordNet with selectional preference relations. In: Joint Symposium on Semantic Processing. (2013) 105–109
18. Terence, P.: Events in the semantics of English: A study in subatomic semantics (1990)
19. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the Fourth Global WordNet Conference (GWC-2008), Szeged, University of Szeged (2008) 311–321
20. Sass, B.: The Verb Argument Browser. 11th International Conference on Text, Speech and Dialog (TSD) (2008) 187–192

III. Szemantika, ontológia

A ℜeALKB tudástár metamodell-vezérelt megvalósítása

Kilián Imre[1], Alberti Gábor[2]

[1]PTE TTK Informatika Tanszék
[2]PTE BTK Általános Nyelvészeti Tanszék
7624 Pécs, Ifjúság útja 6.
kilian@gamma.ttk.pte.hu[1], alberti.gabor@pte.hu[2]

Kivonat: Az elmúlt évben a ℜeALIS természetes nyelvi elemző és értelmező rendszer[1] [1] tudáskezelő rendszerével kapcsolatos elméleti meggondolásokról számoltunk be [2]. Az elméleti meggondolások mellett egy sor deszkamodell-szerű tesztprogram futtatása engedte meg a derűlátó előrejelzéseket. A deszkamodellek integrációja megkezdődött: a jelen írás ennek előrehaladásáról számol be. A választott megoldás két szempontból is érdekes. Egyrészt a szoftver felületei révén programozható, és a Szemantikus Web projektum OWL ontológialeíró nyelvével [3] felülről kompatibilis, vagyis kész OWL ontológiák betölthetők. Másrészt a tudástár hátterében annak különválasztott metamodellje áll, és a programozható felületen keresztül a tudáselemek metamodell-vezérelt módon hozhatók létre és kérdezhetők le. A következtetések szemszögéből nem cél a teljesség. Egyes korai következtetések betöltési időben belefordíthatók a tudásbázis Prolog tárgymodelljébe, más következtetések későiek, vagyis ha a Prolog saját következtetési mechanizmusa nem lenne elegendő, akkor metaintepreterrel megvalósíthatók.

Logikai programozás és metaszintek

Alapfeltételezés, hogy logikai eszközök kezelését csakis logikai programozási nyelven: a gyakorlatban a Prolog valamelyik dialektusában érdemes megvalósítani. A Prolog következtetési képességei azonban elégtelenek – azok kiterjesztésére mindenképpen szükség van.

Egy logikai következtető rendszer a konkrét alkalmazói adatok kezelése mellett azok modelljét is adatként kezeli, folytonosan módosítja, fejleszti. Ezért a tudástárban legalábbis a modell modelljét, a metamodellt kell beprogramozni. A metamodell már elszakad az alkalmazói világtól, és a modell logikai szerkezetét, a modellelemek összefüggéseit írja le.

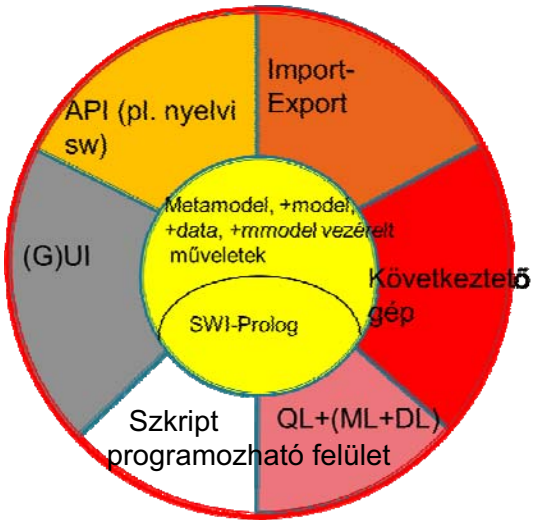
Azért, hogy a logika szerkezet maga is kellően rugalmasan fejlődhessen, célszerű magát a metamodellt sem rögzítetten beprogramozni, hanem legalábbis elválasztva,

[1] A szerzőket e cikk alapjait jelentő kutatásaikban és a konferencia-részvételben a TÁMOP 4.2.2.C-11/1/KONV-2012-0005 (Jól-lét az információs társadalomban) kutatási projektum támogatta.

adatszerűen leírni, és rajta általános algoritmusokat kidolgozni. Erre a megközelítésre szintén a Prolog nyelv a legjobb választás.

A Neumann elvű számítógépek sikere, de a körülvevő élő világ is alátámaszthatja: a metaszintjeiket átmetsző rendszerek különleges fejlődési képesség lehetőségét zárják magukba. Megvizsgáljuk ezért azt, hogy a tudástár esetében a metaszinteket hol lehetséges és célszerű átvágni.

A tudástár felépítése és felületei



1. ábra: A szoftver felületei

A szoftver magja a Prolog nyelven megvalósított tudástár, amely az ANSI Prolog szabványhoz közelálló dialektusban, az SWI-Prolog rendszeren készült. A rendszer a külvilággal az egyes felületein keresztül érintkezik. A tervezett (és részben megvalósított) felületek a következők:

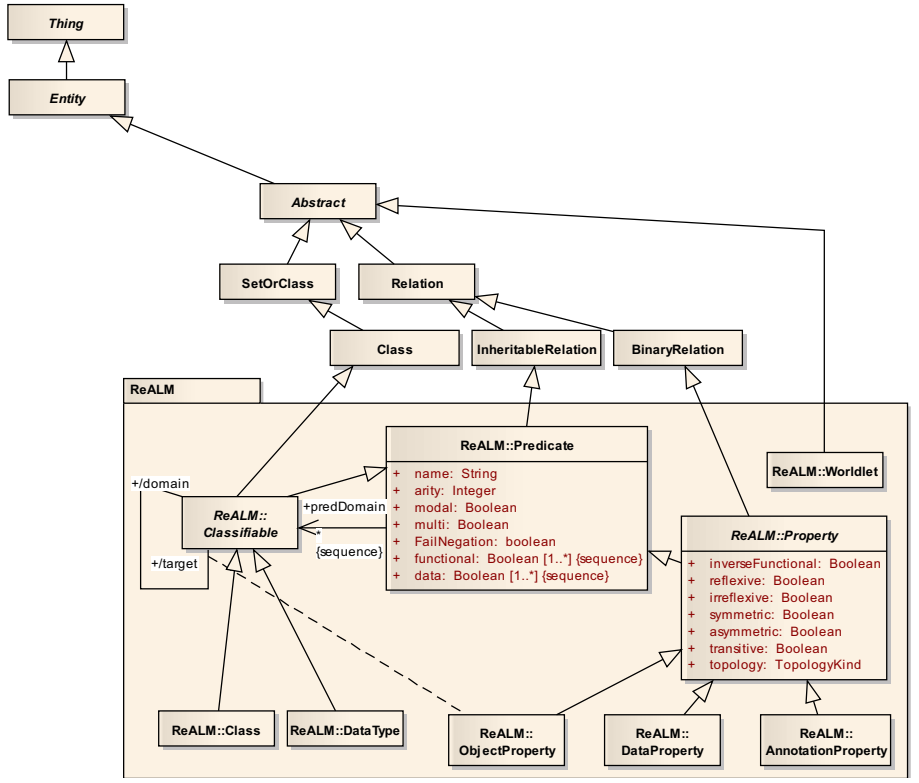
- a tudástárnak rögzített programozható felülete (API) van. Ehhez férhetnek hozzá a nyelvi feldolgozó szoftverek, pl. a ℜeALIS elemző, de a kezelői felület szintén ide kapcsolódik. A felület Prolog nyelvű, amit a megvalósítás adta módon lehet hagyományos programnyelvből meghívni. Jelenleg a Java kapcsolódás van használatban.
- a tudástárhoz egy Java Swing GUI felületet csatoltunk. Ez lehetőséget ad a tudástár adatszerkezeteinek (világmodell, adatréteg, modellréteg) a grafikus böngészésére és módosítására, valamint tesztelési célra egy közvetlen Prolog ablakot is biztosít. A Swing felület monolit felépítménnyel egyrészt könnyen programozható, másrészt a Java alaptechnológia lehetővé teszi a Swing felület lecserélését pl. Java Beans, JSP vagy más rokon technológiára akkor, ha ügyfél-kiszolgáló megoldás szükséges.
- a már meglevő tudástárak anyagának újrafelhasználása érdekében a rendszer OWL ontológiák beolvasására és mentésére is képes lesz. Ezekből egyelőre a beolvasás van az SWI-Prolog alatt használatos Thea csomag [6] segítségével, de offline módon megvalósítva. A Thea közvetlen Prolog formátumra fordít, amit jelenleg a Prolog `consult/1` műveletével tudunk beolvasni.
- a tárolt adatok lekérdezésére egy lekérdező nyelvi felület megvalósítása szükséges. Evégett az Object Query Language (OQL) nyelvet [7], mint az SQL objek-

tum orientált kiterjesztését, valamint az OWL ontológiák lekérdezésére létrehozott SparQL nyelvet [3] célszerű megvalósítani. Jelenleg csupán a Prolog saját eszközeit használhatjuk.

- az egyes részműveletek egymás utáni megvalósítására és gyors, dinamikus programozására valamilyen szkript programíró környezet használható. Jelenleg ez a lehetőség is csupán a Prolog saját eszközeit jelenti.
- Végül, de nem utolsósorban összetettebb következtetések elvégzésére következtető csomag csatolása is szükséges. Itt számba jöhetnek Interneten elérhető következtető csomagok, esetleg Prolog nyelven megvalósított csomagok, és a Prolog sajátmaga is, olyan feladatokra, amelyekre a szegényes képességei elegendőek.

Metamodell

Az import/export műveletek miatt igyekeztünk valamiféle szabványos ontológialeíró nyelvhez illeszkedő megoldást választani. Ezért a Szemantikus Web projektum OWL ontológialeíró nyelvét tiszteletben tartó, de azt bővítő metamodellt határoztunk meg:



2. ábra: A metamodell illesztése a SUMO155 ontológiához

- Megőriztük a 'Class' (osztály) fogalmat és a kétoldalú relációkat magukba foglaló 'Property' (tulajdonság) fogalmat, valamint a tulajdonságok felbontását annotációkra, amelyek String típusúak és megjegyzés-jellegű értékűek, valamint adattulajdonságokra, amelyek skaláris értékűek, és általános objektumtulajdonságokra.
- A tulajdonságok az OWL-ban rögzített metatulajdonságokat kapják.

A bővítés a következő újdonságok bevezetését jelentette:

- Bár kétoldalúval tetszőleges reláció is leírható, és az OWL ontológiai tervminták között is fellelhető hasonló célú minta, az osztályok és a tulajdonságok általánosításaként felvettük a tetszőleges argumentumszámú predikátum fogalmát.
- Tetszőleges argumentumszámú relációkra viszont a `domain` metatulajdonság értéke vektoros: minden argumentumsorszámhoz megadja az argumentum típusát. Minthogy a relációk esetén kitüntetett érték-argumentum (`range`) nincs, így nem is bonthatók annotációs, adat- és objektumrelációkra. Az argumentumokhoz viszont alaptípusok (`String`, `Integer`, stb.) is rendelhetők.
- Ugyanígy, általános relációk esetében a kétoldalúakra vonatkozó egyes metatulajdonságok rögzítése is értelmetlen. Kivétel a függvényszerű működés, amit a `functional` metatulajdonsággal, de annak vektoros értékével adhatunk meg, és azt fejezi ki: a többi argumentum értékének rögzítése esetén az adott argumentum értéke egyértelmű.

Felvettünk egy sor új metatulajdonságot, amelyekkel osztályok és tulajdonságok is jellemezhetők:

- `modal`: modálisan értelmezendő relációk kifejezésére.
- `negation`: a reláció felett explicit negációt használunk, mert nem elegendő a Prolog rendszerekből ismert kudarcalapú negáció (Negation As Failure) alkalmazása
- `multi`: a reláció nem tiszta kétértékű logikában értelmezhető

A metamodellt a metaszint-átvágás végett célszerű az alkalmazói modellbe, illetőleg a csúcsontológiába beilleszteni. Ezt a műveletet fogalomkonszolidációnak is nevezhetjük: a csúcsontológia fogalom- ill. tulajdonság-taxonómiájából levezetjük a metamodell fogalmait. Ha csúcsontológiának a SUMO155 szabadon elérhető ontológiát választjuk [3], akkor az illesztés az alábbi kapcsolatok létrehozását jelenti.

- a világocskakapcsolatok leírására a `subWorldlet` relációt használjuk.

`subWorldlet` $\subseteq$ **`abstractProperty`**

- relációszerkezetek leírására a `subPredicate` relációt használjuk.

`subPredicate` $\subseteq$ `collectionRelation`

- ennek részrelációi a relációleszármazást leíró `subProperty`, valamint az osztályleszármazást leíró `subClass` tulajdonságok.

`subProperty` ⊆ `subPredicate`, ill. `subClass` ⊆ `subPredicate`

- A `ReALM:Classifiable` egy közvetlen példány nélküli, ún. absztrakt osztály, amely a SUMO155 osztályfogalmából van levezetve, és magában foglalja a ℜeALM osztály-, valamint skaláris alap-adattípus fogalmát.

`ReALM:Classifiable` ⊆ `Class`,
`ReALM:Class` ⊆ `ReALM:Classifiable`,
`ReALM:Datatype` ⊆ `ReALM:Classifiable`

- A `ReALM:Predicate` fogalom a SUMO `InheritableRelation` fogalmának kiterjesztése.

`ReALM:Predicate` ⊆ `InheritableRelation`

- A `ReALM:Property` fogalom sajátmaga Predicate fogalmából, ill. a SUMO `BinaryRelation` fogalmából van levezetve.

`ReALM:Property` ⊆ `ReALM:Predicate`
`ReALM:Property` ⊆ `BinaryRelation`

- Végül pedig: a `ReALM:Worldlet` fogalom közvetlenül a SUMO `Abstract` fogalmának kiterjesztése.

`ReALM:Worldlet`⊆ `Abstract`

- Adatszinten egyetlenegy objektumpéldány, a gyökérvilágocska megadása szükséges.

```
Worldlet(root).
```

Prolog futási modell

A metamodelljével rögzített logikai nyelv alapvetően egy Prolog kóddá van leképezve, amelynek a formátumát a megfelelő futási modell rögzíti. Ez a magasabb rendű vagy nem klasszikus logikai szerkezeteket elsőrendű logikába képezi le. Ennek megfelelően a következő átalakítások történnek:

- A közismert logikai alapműveletek Prologban közvetlenül is ábrázolhatók.
- Az egyes logikai metaszintek számára külön Prolog modulokat használunk (`model`: a modelszint, `ontology` az adatszint, és `metamodel` a metamodell számára).
- A modalitást egy külön Prolog argumentumban ábrázoljuk [4, 5]. Az itt ábrázolt modális címke szintaxisa magában foglalja a multimodális logikai szerkezet összes vonását. Vagyis a modális címke szintaxisa lehetőséget ad temporálisan, episztemikusan és deontikusan is modális állítások kifejezésére.

- A diszkrét vagy többértékű logikai értékeket egy külön Prolog argumentumban ábrázoljuk. A megfelelő Prolog hívás hamis jellege jelzi a teljesen lehetetlen eseményt, minden egyéb esetén a Prolog argumentum értéke jelzi a lehetségesség, ill. a bizonyosság mértékét.

A rögzített Prolog futási modell előnye az is, hogy a használt ontológiát is végső soron a Prolog `consult/1` műveletével töltjük be. Az ontológiára épülő esettanulmányok és egyéb példák szintén ugyanígy, Prolog formátumban készíthetők el és tölthetők be.

Korai következtetések

Korai következtetésnek azokat a következtetéseket nevezzük, amelyek valamiféle általános következtetési szabály (pl. öröklődés) közvetlen alkalmazásával keletkeznek. A korai következtetések tekinthetők az interpretálás helyett a tudásbázisba közvetlenül belefordított következtetésnek is. A korai következtetések általában a választott magasabb rendű logikai rendszer axiómáiból állnak elő. A ℜealKB rendszerben a következő korai következtetéseket valósítjuk meg:

- A modális világ szuperindividuális régiójában a világocskák felett öröklődés érvényes. Ez minden egyes predikátumhoz (vagyis többparaméteres relációhoz, tulajdonsághoz és osztályhoz) hozzávesz egy új szabályt, miszerint minden olyan dolog igaz, ami az ősvilágocskában igaz.
- A gyökérvilágocska felett található a mód nélküli világ, ahonnan gyökérvilágocska minden állítását örökli.
- Ha egy objektum egy osztály példánya, akkor példánya az ősosztályénak is. Az öröklődés ilyen megfogalmazása igaz modálisan és mód nélkül is.
- Ha egy példánypár vagy példány n-es példánya egy tulajdonságnak vagy egy predikátumnak is, akkor példánya az őstulajdonságnak, ill. az őspredikátumnak is.

Metamodell-vezéreltség

A modellvezéreltség azt jelenti: olyan műveleteket valósítunk meg, amelyek bár az alkalmazói példányokon dolgoznak, de paraméterként megkapják azokat az alkalmazói modellelemeket is, amelyeknek a példányai. Vagyis a megvalósított műveletek nemcsak az adattartalomtól függetlenek (mint minden tisztességes szoftver esetén), de a konkrét modelltől is. Így, ha a modell változtatása szükséges, akkor a futó kódok nem muszáj újraírni, a módosítást elegendő csupán a modellben megtenni.

Metamodell-vezérelt lehet egy modelltárház szoftver akkor, ha vagy egy általános modelltárházat tervezünk, amely egyidejűleg esetleg több metamodellű rendszert is kezelni kíván, vagy a metamodell maga sem rögzített, ezért a fejlesztés során metamodell-módosításokat is tekintetbe kell vennünk.

A ℜeALKB tudástár ilyen értelemben metamodell vezérelt, és a megfelelő metamodell-, ill. modellelemmel paraméterezett CRUD (Create, Read, Update, Delete) műveleteket valósít meg.

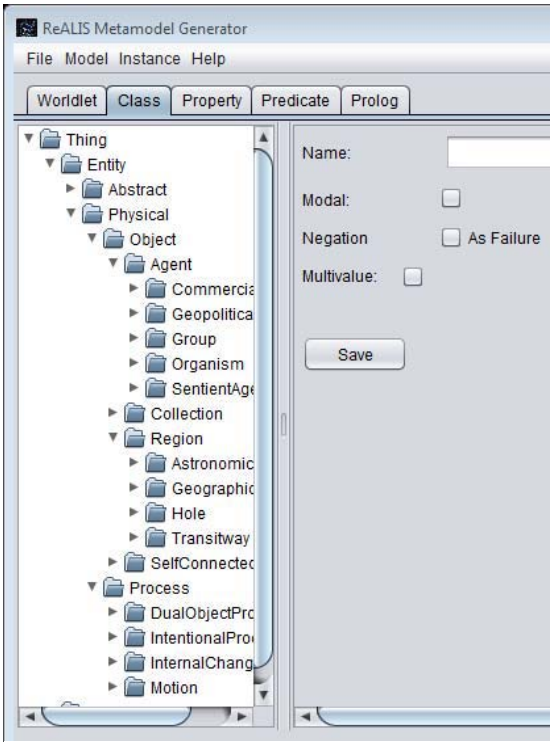
A Java Swing kezelői felület használata

eALKB rendszer lehetőségeit használja ki így a ℜeALM nyelv modellelemeinek, és a modellelemeknek megfelelő példányelemek kezelésére alkalmas. Ez a modell- és példányelemek létrehozását, módosítását, törlését valamint ellenőrzését jelenti.

A felület tartalmaz egy közvetlen Prolog végrehajtást lehetővé tevő ablakot is. Külön érdekesség a kétoldalú homogén relációkra vonatkozó általános böngésző alkalmazása, amely a reláció topológiájától függően alkalmaz vezérlőelemeket (Tree, List stb.)

Az általános böngésző legfontosabb alkalmazásai a világocska-szerkezet és az osztályszerkezet feletti böngészők, de ugyanígy alkalmazható pl. földrajzi objektumok között a részterület feletti viszonyra, vagy akár híres személyek családfájára.

Eredmények, továbbfejlesztés



3. ábra: A Java Swing kezelői felület

A vázolt rendszer fejlesztés alatt áll, létezik, működik, bemutatható. Amint egy viszonylag stabil és kerek változat elkészül, nyilvánosan elérhetővé kívánjuk tenni, és felajánljuk a tudományos közösségnek használat és továbbfejlesztés céljából.

A jelenleg legfontosabb célunk egy működő és stabil változat létrehozása és közzététele. Ha ez sikerült, akkor kerülhet sor a továbbiakra…

- hiányzó modulok megvalósítása és a rendszerbe illesztése
- ügyfél-kiszolgáló felépítmény megvalósítása
- egyes tételbizonyítók és megoldók rendszerbe integrálása
- konkrét lekérdezési nyelvek megvalósítása
- a jelenlegi bővített, de alap-

jaiban kétértékű logikai modell tágítása fuzzy irányba

Itt szeretnék köszönetet mondani a ℜeALIS projektbéli munkatársaimnak, Alberti Gábornak, Kleiber Juditnak és Károly Mártonnak a nyelvészeti információk önzetlen átadásáért, és a jól célzott, és egyben megfelelően adagolt, a cikk végső példányára is kiható megjegyzéseikért.

Hivatkozások

1. Alberti G.: ℜeALIS. Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
2. Kilián I.: A ReALIS tudástároló és következtető alrendszere In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia 2013. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2013) 225–235
3. Niles, I., Pease, A.: Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology. In: Proceedings of Measuring Intelligence and Performance of Intelligent Systems Conference (2001)
4. Grosof, B. N., Horrocks, I., Volz, R., Decker, S.: Description Logic Programs: Combining Logic Programs with Description Logic. In: Proceedings of the Twelfth International World Wide Web Conference, ACM (2003) 48–57
5. Ohlbach, H.J.: A Resolution Calculus for Modal Logic. FB Informatik, University of Kaiserslautern, Germany (1988) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.5003, letöltve: 25-Jun-12.)
6. Vassiliadis, V., Wielemaker, J., Mungall, C.: Processing OWL2 ontologies using Thea: An application of logic programming. In: Proceedings of OWL: Experiences and Directions (OWLED), CEUR Workshop Proceedings, Vol-529. (2009) (http://www.webont.org/owled/2009, letöltve: 25-Jun-12)
7. Cattell, R. G. G., Barry, D., et al.: The Object Data Standard: ODMG 3.0 Morgan Kaufmann publishers San Francisco, USA (1999)

Bizonytalanságot jelölő kifejezések azonosítása magyar nyelvű szövegekben

Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged Árpád tér 2.
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink egyik intenzíven vizsgált területe a számítógépes nyelvészeti kutatásokban. Ebben a cikkben bemutatjuk magyar nyelvű annotált korpuszunkat, melyben kézzel bejelöltük a nyelvi bizonytalanság különféle fajtáit jelző nyelvi elemeket. A korpusz arra is lehetőséget kínál, hogy beszámoljunk az első, magyar nyelvű bizonytalanságazonosító gépi tanuló rendszer eredményeiről.
Kulcsszavak: információkinyerés, szemantika, korpusz

1. Bevezetés

A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink számítógépes nyelvészeti kutatásának egyik fontos problémaköre [1]. A feladat fontossága abban rejlik, hogy a különféle számítógépes nyelvészeti alkalmazásokban lényegi szerep jut a tényszerű és a bizonytalan, illetve tagadott információ megkülönböztetésének, hiszen például információkinyerés és szemantikus keresés esetében a felhasználónak többnyire tényszerű információra van szüksége, így alkalmazástól függően a rendszer vagy kiszűri a bizonytalan / tagadott szövegrészeket, vagy pedig a tényektől elkülönítve adja őket vissza a felhasználónak. A problémára eddig elsődlegesen angol nyelvű szövegeken nyújtottak megoldásokat [1,2]. Ebben a cikkben bemutatjuk kézzel annotált, magyar nyelvű bizonytalansági korpuszunkat, és beszámolunk az első eredményekről a nyelvi bizonytalanságot jelölő elemek automatikus felismeréséről magyar nyelvű szövegekben.

2. A bizonytalanság típusai

A nyelvi bizonytalanságot hagyományosan a mondat szemantikájához szokták kötni, azonban vannak olyan bizonytalanságot jelző nyelvi elemek is, melyek ezzel szemben a mondat (közlés) kontextusában – diskurzusbeli tényezőknek

köszönhetően – válnak többértelművé. Például a *Lehet, hogy esik az eső* mondat alapján nem tudjuk eldönteni, hogy esik-e az eső (szemantikai bizonytalanság), viszont a *Számos kutató szerint az MSZNY a legjobb magyar konferencia* mondatból az nem derül ki, hogy pontosan kinek (illetve hány kutatónak) a véleményéről esik szó, így a közlés forrása marad bizonytalan (diskurzusszintű bizonytalanság). Ebben a cikkben követjük a [2], illetve [3] cikkekben felvázolt osztályozást a bizonytalanság különböző fajtáira nézve, illetve a magyar nyelvre alkalmazzuk azt, annotációs elveinket a fentiek alapján kialakítva.

A szemantikai bizonytalanságnak több osztálya is létezik. Egy propozíció episztemikusan bizonytalannak számít, ha a világtudásunk alapján nem tudjuk eldönteni ebben a pillanatban, hogy igaz-e vagy hamis. Ugyanez igaz a hipotetikus bizonytalanságra is, ide sorolhatók a feltételes mondatok, illetve a vizsgálati bizonytalanság – utóbbi különösen tudományos cikkekben gyakori, hiszen a kutatási kérdést gyakran a vizsgálati bizonytalanság nyelvi eszközeivel fogalmazzák meg a szerzők. A modalitás nem episztemikus típusai (például doxasztikus bizonytalanság, mely a hiedelmekkel függ össze, illetve a dinamikus modalitás különböző fajtái, melyek többek között a szükségszerűséghez kapcsolódnak) szintén ebbe a nagyobb csoportba sorolhatók.

A diskurzusszintű bizonytalanságnak három osztályát különböztethetjük meg [3]. Először, a *weasel* kifejezésekhez nem tudunk egyértelműen forrást rendelni (azaz nem tudjuk, kihez köthető az adott információ), más esetben pedig hiányzik a közlésből egy fontos és releváns információrészlet, amely azonban az adott helyzetben szükséges lenne. Másodszor, a *hedge* szavak homályossá teszik bizonyos mennyiségek vagy minőségek pontos jelentését. Harmadszor, a *peacock* kifejezések bizonyítatlan (vagy bizonyíthatatlan) értékeléseket, minősítéseket vagy túlzásokat fejeznek ki.

A bizonytalanságot jelző kulcsszavakra itt mutatunk néhány példát:

Episztemikus: **Lehet**, hogy esik.

Dinamikus: Mennem **kell**.

Doxasztikus: Azt **hiszi**, hogy a Föld lapos.

Vizsgálat: A felvétel manipuláltságáról **vizsgálatot folytattak**.

Feltételes: **Ha** esik, itthon maradunk.

Weasel: **Egyesek** szerint inkább megszállást kellene mondani.

Hedge: A belga lakosság **kb.** 10%-a él Brüsszelben.

Peacock: Apafi négy évet **keserves** tatár fogságban töltött.

Az angolra alkalmazott osztályozást változtatások nélkül vettük át a magyarra, azonban a magyar nyelv sajátságainak megfelelően az annotációs elveket némileg átalakítottuk. Például az episztemikus bizonytalanságot a magyarban igen gyakran a *-hat/-het* képző fejezi ki, míg az angolban ez segédigék (pl. *can*, *may*) használatával történik. Ezekben az esetekben az angol korpuszban a segédigét jelöltük meg mint bizonytalanságot jelző elemet, a magyarban azonban a teljes szóalakot, mivel a képző külön címkézésére nem volt lehetőségünk morfémákra bontott nyelvi adatbázisok híján.

A [2] és [3] munkákhoz hasonlóan e cikkben is a diskurzusszintű bizonytalanság mindhárom fajtájával, illetve a szemantikus bizonytalanság négy fajtájával (episztemikus, vizsgálati, feltételes és doxasztikus) foglalkozunk.

3. Kapcsolódó irodalom

A bizonytalanságot jelző nyelvi elemek vizsgálata napjaink számítógépes nyelvészeti kutatásainak egyik népszerű témája. Ezt jelzi többek között a CoNLL-2010 verseny megrendezése, melynek témája a nyelvi bizonytalanság azonosítása volt biológiai cikkekben és Wikipedia-szócikkekben, angol nyelven [1], illetve a Computational Linguistics folyóirat tematikus különszáma (Vol. 38, No. 2), melyet a bizonytalanság és tagadás automatikus azonosításának szenteltek. Az eddigi vizsgálatok túlnyomórészt az angol nyelv köré csoportosulnak, és elsődlegesen újsághíreket, biológiai publikációkat vagy orvosi dokumentumokat, illetve Wikipedia-szócikkeket elemeznek (vö. [2, 4, 5]).

A felügyelt gépi tanulási eljárások megkövetelik egy annotált korpusz létét. Noha számos, bizonytalanságra épített korpusz elérhető a világban (a teljesség igénye nélkül megemlítve néhányat: BioScope [6], Genia [4], FactBank [5], a CoNLL-2010 verseny korpuszai [1]), ezek azonban angol nyelvűek. A magyar nyelvű kutatások egyik fontos előkészületi lépésének bizonyult tehát egy kézzel annotált, magyar nyelvű adatbázis elkészítése, melyben nyelvész szakértők bejelölték a bizonytalanságot jelző nyelvi elemeket.

A bizonytalanságot azonosító rendszerek eleinte szakértői szabályok alapján működtek (pl. [7,8]), az utóbbi időben azonban gépi tanulásra épülnek, többnyire felügyelt tanulási módszereket hasznosítva (pl. [9,10] és a CoNLL-2010 versenyen részt vevő rendszerek [1]). A legutóbbi tendenciákkal összhangban e cikkben bemutatunk egy felügyelt tanulásra épülő modellt, mely gazdag jellemzőtérrel rendelkezik: lexikai, morfológiai, szintaktikai és szemantikai jegyekre egyaránt épít, továbbá kontextuális jellemzőket is figyelembe vesz.

4. A korpusz

A hUnCertainty korpusz magyar nyelvű Wikipédia-szócikkekből áll, összesen 1081 bekezdést, 9722 mondatot és 180 000 tokent tartalmaz. A szövegek kiválogatása során összegyűjtöttük a legtipikusabb angol nyelvű bizonytalan kulcsszavak magyar megfelelőit, majd az olyan bekezdések kerültek bele a korpuszba, amelyek legalább egyet tartalmaztak e kulcsszavak közül. Mindemellett olyan bekezdések is a korpusz részét képezik, amelyek nem tartalmazták ezen kulcsszavak egyikét sem, így törekedve a korpuszbeli adatok kiegyensúlyozottságára.

A korpuszban kézzel jelöltük meg a bizonytalanságért felelős nyelvi elemek (kulcsszavak) több fajtáját. A korpuszban előforduló kulcsszavak arányát az 1. táblázat mutatja.

Mint látható, a korpuszban a diskurzusszintű bizonytalanság kulcsszavai dominálnak. Ez összhangban van a korábban angol nyelvű Wikipedia-szócikkeken

elért eredményekkel [3], így valószínűleg a kulcsszavak ilyen eloszlása a Wikipédia-szövegek sajátja nyelvtől függetlenül.

1. táblázat. Bizonytalanságot jelző kulcsszavak.

Kulcsszó típusa	#	%	Eltérő kulcsszavak száma
Hedge	2100	35,12	439
Weasel	2150	35,95	598
Peacock	788	13,18	400
Diskurzusszintű összesen	5038	84,25	1437
Episztemikus	441	7,37	184
Doxasztikus	316	5,28	67
Feltételes	154	2,58	46
Vizsgálat	31	0,52	22
Szemantikus összesen	942	15,75	319
Összesen	5980	100	1756

Ha a mondatok szintjén vizsgáljuk a bizonytalanságot, azt találjuk, hogy a korpuszban 3710 (39,22%) bizonytalan mondat szerepel (azaz legalább egy kulcsszót tartalmaznak). Ezek közül 3344 mondat tartalmaz diskurzusszintű bizonytalanságot jelölő kulcsszót (35,35%), és 746 pedig szemantikus bizonytalanságra utaló kulcsszót (7,89%).

A 2. táblázat foglalja össze a leggyakoribb magyar episztemikus és doxasztikus kulcsszavakat. Az első tíz kulcsszó adja az összes előfordulás 42 és 79%-át ezen kulcsszavak esetében. Mivel a feltételes és a vizsgálati kulcsszavak nem mutatnak nagy változatosságot a korpuszban, csak a legalább háromszor előforduló elemeket soroljuk fel itt: a *vizsgál* és *tanulmányoz* szavak adják a vizsgálati kulcsszavak 29%-át, illetve a *ha*, *akkor* és *amennyiben* szavak a feltételes kulcsszavak 68%-át.

2. táblázat. A leggyakoribb episztemikus és doxasztikus kulcsszavak.

Episztemikus	#	%	Doxasztikus	#	%
valószínűleg	79	17,87	szerint	151	47,63
talán	28	6,33	tart	25	7,89
feltehetőleg	15	3,39	tekint	19	5,99
állítólag	14	3,17	állít	18	5,68
feltehető	11	2,49	vél	10	3,15
lehet	10	2,26	tulajdonít	7	2,21
lehetséges	10	2,26	gondol	6	1,89
feltételez	7	1,58	tesz	5	1,58
tekinthető	7	1,58	hisz	4	1,26
lehetőség	6	1,36	vall	4	1,26

A 3. táblázatban találhatók meg a leggyakoribb, diskurzusszintű bizonytalanságot jelölő kulcsszavak. A tíz leggyakoribb kulcsszó az esetek 40, 31 és 26%-át fedi le a weasel, hedge és peacock előfordulásoknak.

3. táblázat. A leggyakoribb diskurzusszintű kulcsszavak.

Weasel	#	%	Hedge	#	%	Peacock	#	%
számos	150	8,60	általában	127	6,18	fontos	50	6,36
egyes	134	7,68	gyakran	119	5,79	jelentős	39	4,96
egyik	118	6,76	később	99	4,82	ismert	25	3,18
más	100	5,73	nagyon	50	2,43	híres	23	2,93
néhány	66	3,78	főleg	47	2,29	nagy	17	2,16
különböző	34	1,95	nagy	46	2,24	kiemelkedő	15	1,91
egyéb	29	1,66	igen	43	2,09	komoly	11	1,40
sok	27	1,55	néhány	40	1,95	erős	10	1,27
bizonyos	22	1,26	főként	37	1,80	kiváló	9	1,15
többek között	19	1,09	mintegy	36	1,75	egyszerű	9	1,15

Néhány kulcsszó több bizonytalansági osztályt is jelölhet, ugyanakkor a kulcsszavak nem minden előfordulása jelöl ténylegesen bizonytalanságot az adott kontextusban. Az első esetre példa a *nagy* szó, amely hedge és peacock kulcsszó is lehet attól függően, hogy fizikai vagy minőségi nagyságra utal-e. A második esetet illusztrálja az *igen* szó: határozószóként előfordulhat hedge-ként, mondatszóként azonban nem jelöl bizonytalanságot.

Minthogy a hUnCertainty korpusz annotációs elvei angol korpuszok építése során használt elveken alapulnak [2,3], az angol és magyar korpuszokból származó adatok összevethetők egymással. Például a szemantikai és diskurzusszintű bizonytalanság kulcsszavai hasonló arányban fordulnak elő mindkét nyelvű Wikipédia-szövegekben. A kulcsszavak szintjén pedig megfigyelhetjük, hogy azonos jelentésű szavak szerepelnek a leggyakoribb kulcsszavak között, például *valószínű*, *lehetséges*, *hisz*. E tények arra utalnak, hogy a [2] és [3] munkákban bemutatott osztályozás több nyelvre is alkalmazható.

5. A bizonytalanság automatikus azonosítása

Annak érdekében, hogy automatikus úton azonosítsuk a bizonytalanságot jelölő kulcsszavakat, kifejlesztettünk egy gépi tanuláson alapuló módszert, melyet a következőkben ismertetünk részletesen. Méréseinkhez a hUnCertainty korpuszt vettük alapul, melyet a magyarlanc elemzőt [11] felhasználva morfológiailag és szintaktikailag elemeztünk.

5.1. Gépi tanulási módszerek

Korábbi angol nyelvű kísérleteink alapján a szekvenciajelölés bizonyult a legeredményesebbnek a bizonytalanság automatikus azonosításában [2], így a magyar nyelvű anyagon végzett méréseinket is feltételes véletlen mezőkön (CRF)

[12] alapuló módszerrel kiviteleztük. Kísérleteink kiindulópontjaként egy magyar nyelvre implementált, MALLET alapú névelem-felismerő rendszer [13] szolgált, a felhasznált jellemzőket természetesen a bizonytalanságazonosítási feladat sajátságaira szabva, melyeket az alábbiakban ismertetünk:

- **Felszíni jellemzők:** a szó írásmódjával kapcsolatos jellemzők (tartalmaz-e írásjelet, számot, kis/nagybetűket, szóhossz, mássalhangzó bi- és trigramok...)
- **Lexikai jellemzők:** a hasonló elvek alapján annotált, rendelkezésre álló angol nyelvű korpuszokból [2] minden bizonytalansági típushoz kigyűjtöttük a leggyakoribb kulcsszavakat, és ezeket magyarítva listákba rendeztük őket. A listákat bináris jellemzőként használtuk fel: ha az adott szó lemmája előfordult valamelyik listában, akkor igaz értéket kapott az adott jellemzőre nézve.
- **Morfológiai jellemzők:** minden szó esetében felvettük annak fő szófaját, illetve lemmáját a jellemzők közé. Igék esetében továbbá megvizsgáltuk, hogy ható igéről van-e szó, feltételes módú-e az ige, illetve T/1. vagy T/3. alakban fordul-e elő. Főnevek esetében felvettük jellemzőként, hogy egyes vagy többes számban állnak-e. Külön jelöltük a névmások esetében azt is, ha határozatlan névmásról volt szó, illetve mellékneveknél a fokot is felvettük a jellemzők közé.
- **Szintaktikai jellemzők:** minden szóhoz felvettük annak szintaktikai címkéjét, továbbá főnevek esetében megvizsgáltuk, hogy rendelkezik-e névelővel, illetve igék esetében felvettük, hogy van-e alanya.
- **Szemantikai/pragmatikai jellemzők:** egy általunk összeállított, beszédaktusokat tartalmazó lista alapján megvizsgáltuk, hogy az adott szó beszédaktust jelölő ige-e. Mindemellett a kulcsszavakhoz hasonlóan, angol nyelvű, pozitív és negatív jelentéstartalmú szavakat tartalmazó listákat [14] is magyarítottunk, és megnéztük, hogy a szó lemmája szerepel-e az adott listában.

Az adott szó környezeti jellemzőjeként felvettük a tőle egy vagy két szó távolságra levő szavak szófaji kódját és szintaktikai címkéjét is.

A fentiekben leírt jellemzőkészlet alapján tízszeres keresztvalidációt használva hajtottuk végre méréseinket a hUnCertainty korpuszon. Mivel csak a tokenek körülbelül 3%-a funkcionál kulcsszóként a korpuszban, így szükségesnek láttuk a tanító adatbázis szűrését: a kulsszót nem tartalmazó mondatoknak csak a fele került bele a tanító halmazba. Továbbá mivel a vizsgálati bizonytalanság kulcszavai összesen 31 előfordulást mutattak, ezt az ritka osztályt nem vettük figyelembe a rendszerünk létrehozásánál, így a kiértékelésben sem szerepel.

5.2. Baseline mérések

Baseline mérésként egyszerű szótárillesztést használtunk. A lexikai jellemzők között említett listákat jelöltük rá a korpuszra: amennyiben a szó lemmája megegyezett az adott lista egyik elemével, a bizonytalanság adott típusának címkéztük fel.

6. Eredmények

A 4. táblázat mutatja a baseline, valamint a gépi tanuló kísérletek eredményeit. A kiértékelés során a pontosság, fedés és F-mérték metrikákat alkalmaztuk.

4. táblázat. Eredmények.

	Szótárillesztés			**Gépi tanuló**			**Különbség**
Típus	Pontosság	Fedés	F-mérték	Pontosság	Fedés	F-mérték	
Weasel	26,03	38,50	31,06	59,26	34,74	43,80	+12,74
Hedge	55,86	29,92	38,97	64,59	50,02	56,38	+17,41
Peacock	23,29	30,63	26,46	37,85	13,80	20,22	-6,38
Episztemikus	49,57	37,34	42,59	63,95	36,03	46,09	+3,5
Doxasztikus	25,24	65,20	36,40	54,31	33,54	41,47	+5,07
Feltételes	29,66	67,74	41,26	47,12	31,61	37,84	-3,42

A táblázatból jól látszik, hogy a gépi tanuló megközelítés eredményei két osztály kivételével minden esetben meghaladták a baseline szótárillesztés által elért eredményeket. Ez elsődlegesen a pontosság javulásának köszönhető, mely kivétel nélkül minden osztályra nézve jóval magasabb lett a szekvenciajelölő megközelítés esetén. Ezzel szemben a fedési értékek nagyobb változatosságot mutatnak: míg a hedge osztály esetében ez is nőtt, a weasel és episztemikus kulcsszavaknál nem változott jelentős mértékben, addig a peacock, doxasztikus és feltételes kulcsszavaknál drasztikus visszaesést figyelhetünk meg. Vélhetően a gyenge fedésre vezethető vissza az is, hogy a peacock és feltételes kulcsszavaknál a szótárjelölő megközelítés magasabb F-mértéket ért el, mint a gépi tanuló algoritmus.

7. Az eredmények megvitatása

Elért eredményeink azt igazolják, hogy a magyar nyelvben is lehetséges a bizonytalanságot jelölő kifejezések automatikus azonosítása szekvenciajelölő megközelítéssel. A szótárillesztés során a legjobb eredményeket az episztemikus, feltételes és hedge kulcsszavakon értük el, míg a szekvenciajelöléssel a hedge, episztemikus és weasel osztályokon születtek a legjobb eredmények. Mindezek alapján a hedge és episztemikus osztályok tűnnek a legkönnyebben felismerhetőknek. Az eredmények arra is utalnak, hogy azon (szemantikai) osztályok esetében, ahol kicsi volt a különbség a szótárillesztés és gépi tanulás eredményei között, az adott bizonytalanságtípus nyelvi jelölésmódja elsődlegesen lexikális (és kevésbé többértelmű) eszközökkel valósul meg. Ugyanakkor a diskurzusszintű bizonytalanság kulcsszavainak felismerésében nagyobb szerepet játszik a gépi tanulás, ami annak köszönhető, hogy esetükben igen fontos szerepe van a kontextusnak (diskurzusnak), így egy szekvenciajelölő algoritmus sikeresebben tudja megoldani a feladatot.

Amennyiben eredményeinket összevetjük a korábban angol nyelvű Wikipedia-szócikkeken elért, szemantikai bizonytalanságot azonosító rendszer által elértekkel [2], azt láthatjuk, hogy angol nyelven könnyebbnek tűnik a feladat: 0,6 és 0,8 közötti F-mértékekről számol be a cikk. Azonban nem szabad figyelmen kívül hagynunk két fontos tényezőt. Egyrészt a két nyelv közti tipológiai különbségeknek köszönhetően az angolban inkább lexikálisan meghatározott a bizonytalanság jelölése, a magyarban pedig inkább morfológiai eszközök valósítják meg ezt: például a ható igéket a magyarban a *-hat/-het* képző jelöli, az angolban pedig a *may*, *might* stb. segédigék. Így a szóalak, illetve lemma jellemzőként való szerepeltetése angolban már viszonylag jó eredményekhez vezethet, magyarban azonban ezek a jellemzők önmagukban (morfológiai jellemzők felvétele nélkül) kevésbé hatékonyak. Másrészt az adatbázis nagysága jelentősen különbözik a két esetben: míg körülbelül 20000 annotált angol mondat állt rendelkezésre, addig a magyarban ez a szám nem érte el a 10000-et. Az annotált adatok mennyiségének fontosságát igazolják az angol nyelvű mérések is: azokban az esetekben, amikor csupán néhány ezer annotált mondat állt rendelkezésre, az elért F-mértékek – doméntől és kulcsszótípustól függően – 0,1-0,8 között mozogtak.

A peacock és a feltételes kulcsszavak esetében a szekvenciajelölő módszer rosszabbul teljesített a szótárjelölő megközelítésnél: mindkét esetben a pontosság nőtt ugyan, de a fedés jelentős visszaesést mutatott. Ez alapján szükségesnek ígérkezik a rendszer felülvizsgálata, továbbá új, speciálisan ezekre az osztályokra kifejlesztett jellemzők definiálása.

A gépi tanuló rendszer kimenetét részletesen is megvizsgáltuk hibaelemzés céljából. Azt találtuk, hogy elsődlegesen a többértelmű kulcsszavak egyértelműsítése jelent problémát. Például a *számos* vagy *sok* szavak lehetnek szövegkörnyezettől függően weasel és hedge kulcsszavak is, vagy a *nagy* lehet peacock és hedge is. Az ehhez hasonló eseteket a rendszer időnként rossz osztályba sorolta. Gyakori hibaforrásnak számítottak azok a kulcsszavak is, amelyek gyakran használatosak nem kulcsszó jelentésben is, mint például a *tart* ige, amely lehet doxasztikus kulcsszó (*vki vmilyennek tart vkit/vmit*), azonban más jelentésben nem kulcsszó (pl. *vki vhol tart vmit*, *vki vhol tart vmiben* stb.). Egy sajátos hibának bizonyult az episztemikus osztálynál a tagadást tartalmazó kulcsszavak fel nem ismerése: a *nem zárható ki*, *nem tudni* stb. alakokat a rendszer nem jelölte meg kulcsszóként.

8. Összegzés

Ebben a cikkben bemutattuk a hUnCertainty korpuszt, amely az első kézzel annotált, magyar nyelvű bizonytalansági korpusz. A korpusz lehetőséget adott arra, hogy beszámoljunk az első eredményekről a nyelvi bizonytalanságot jelölő elemek automatikus felismeréséről magyar nyelvű szövegekben. A szekvenciajelölésen alapuló, gazdag jellemzőtérrel dolgozó megközelítésünk által elért eredményeink bizonyítják, hogy magyar nyelvre is alkalmazható a bizonytalanság nyelvi modellje, illetve a bizonytalanságot jelölő kulcsszavak automatikus azonosítása is megoldható.

A jövőben módszereinket szeretnénk továbbfejleszteni, elsősorban a jobb fedés elérésének irányába, mindemellett más jellegű szövegekben is szeretnénk annotálni, illetve automatikusan azonosítani a bizonytalanságot jelölő kifejezéseket.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task, Uppsala, Sweden, Association for Computational Linguistics (2010) 1–12
2. Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics **38** (2012) 335–367
3. Vincze, V.: Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 383–391
4. Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics **9**(Suppl 10) (2008)
5. Saurí, R., Pustejovsky, J.: FactBank: a corpus annotated with event factuality. Language Resources and Evaluation **43** (2009) 227–268
6. Vincze, V., Szarvas, Gy., Farkas, R., Móra, Gy., Csirik, J.: The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. BMC Bioinformatics **9**(Suppl 11) (2008) S9
7. Light, M., Qiu, X.Y., Srinivasan, P.: The language of bioscience: Facts, speculations, and statements in between. In: Proc. of the HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases. (2004) 17–24
8. Chapman, W.W., Chu, D., Dowling, J.N.: Context: An algorithm for identifying contextual features from clinical text. In: Proceedings of the ACL Workshop on BioNLP 2007. (2007) 81–88
9. Medlock, B., Briscoe, T.: Weakly Supervised Learning for Hedge Classification in Scientific Literature. In: Proceedings of the ACL, Prague, Czech Republic (2007) 992–999
10. Özgür, A., Radev, D.R.: Detecting speculations and their scopes in scientific text. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Association for Computational Linguistics (2009) 1398–1407
11. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013) 763–771

12. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML-01, 18th Int. Conf. on Machine Learning, Morgan Kaufmann (2001) 282–289
13. Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: Proceedings of the 9th international conference on Discovery Science. DS'06, Berlin, Heidelberg, Springer-Verlag (2006) 267–278
14. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)

Mit iszunk? A Magyar WordNet automatikus kiterjesztése szelekciós preferenciákat ábrázoló szófajközi relációkkal

Miháltz Márton, Sass Bálint

MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport, 1444 Budapest, Pf. 278.
{mihaltz.marton, sass.balint}@itk.ppke.hu

Kivonat: A cikkben bemutatott, folyamatban lévő munkálatok célja a Magyar WordNet automatikus kiegészítése új, különböző argumentumpozíciók szelekciós preferenciáit ábrázoló ige-főnév relációkkal. Bemutatunk egy algoritmust, amely korpuszgyakorisági adatok és a WordNet hierarchikus szerkezete alapján megkísérli azonosítani a vonzatpozíciók szemantikai típusait legjobban reprezentáló HuWN hipernima-algráfokat. Az eljárás segítségével minden, a korpuszban megtalálható, esetraggal vagy névutóval jelölt igei argumentumpozíciót igyekszünk lefedni. Nem célunk egyértelmű, kizárólagos kategóriák kijelölése, ehelyett súlyozott listák segítségével igyekszünk felsorolni a megfigyelt példákból általánosítható leggyakoribb típusokat. Az eredmények reményeink szerint a Magyar WordNet felhasználóin felül az általunk fejlesztett szintaktikai elemző számára is hasznos erőforrásként fognak szolgálni. A cikkben bemutatunk néhány előzetes eredményt és szót ejtünk néhány felmerülő kérdésről.

1 Bevezetés

1985-ös első kiadása óta a Princeton Wordnet (PWN) [5] mára általánosan elterjedt lexikális szemantikai erőforrássá vált a nyelvtechnológiai kutatásokban és alkalmazásokban. Szabad hozzáférhetősége, tekintélyes lefedettsége és folyamatos fejlődése mind hozzájárultak sikereihez.

Története során több lehetséges irány megfogalmazódott a PWN további javíthatósága szempontjából. Az NLP-felhasználók szemszögéből a PWN egyik hiányossága, hogy a szófajokon belül meglévő gazdag relációrendszerhez képest jóval kevesebb szófajok közötti (különböző szófajú synseteket összekapcsoló) relációt tartalmaz. A főnevek, igék, melléknevek és határozószók alhálózatai között jelenleg csak morfológiai (derivációs) kapcsolatok vannak, pl. *research* (ige) — *researcher* (fn), *engage* (ige) — *engagement* (fn) stb.

Jelen kutatás célja, hogy automatikus módszereket találjunk arra, hogy a Magyar WordNetet (HuWN) [9] bizonyos, az igéket és főneveket összekötő relációkkal egészítsük ki korpuszadatok alapján. E relációk az igék és az ige mellett megjelenő adott esetragú/névutójú bővítmények között hoznak létre kapcsolatot úgy, hogy megadják a szóban forgó vonzat szemantikai típusának általánosítását legjobban reprezentáló főnévi WordNet csomópontot, pl. *{eszik}-{étel}, {ír}-{írásmű}* stb. Ez az információ

többek között felhasználható jelenleg folyó, pszicholingvisztikai relevanciájú nyelvi elemző fejlesztését célzó projektünkben is (ld. [10] és Prószéky et al jelen kötetben).

A cikk további felépítése a következő: a következő részben röviden érintjük a magyar igei argumentumszerkezet szintaxisának és szemantikájának néhány releváns kérdését, majd ismertetjük kutatásaink céljait. A 3. részben bemutatjuk a vonatkozó irodalmat, a 4. részben az általunk javasolt algoritmust, majd az 5. részben néhány előzetes eredményt. Végül ismertetjük a további lehetséges fejleszési irányokat.

2 Háttér

A magyarban az igei argumentumokat (komplemenseket) szintaktikailag az esetragok, illetve a névutók adják meg. Ezek a relációk függvényei az egyes igék vonzatkereteinek: különböző igei vonzatkeretek különböző morfoszintaktikai pozícióihoz különböző névszói fogalmak tartozhatnak (pl. *figyel valami*RE*, elkezdődik valami*Ø*, odaéget valami*T*, érdeklődik valami* UTÁN stb.)

Másfelől ez a kötődés széles spektrumot mutathat: az egyik véglet az olyan idiomatikus, nem-kompozícionális ige–igei módosító kapcsolatoké, mint pl. *hangot ad (valaminek), issza a szavát, napvilágra hoz, tenyerén hordoz* stb. A másik végletet az olyan vonzatok képviselik, amelyeknek megfeleltethetők – egy vagy több – olyan szemantikai osztállyal, amelyek produktívan képesek az adott pozícióban elfogadható kifejezések szemantikai kategóriáját megjósolni (szelekciós preferenciák): *eszik valamit {étel, ennivaló}, ír valamit {írás, írásmű}, kiönt valami {víz, víztömeg}* stb.

Gyakran egy adott ige adott vonzatpozíciójához több szemantikai kategória is tartozik, pl. *iszik valamit {folyadék: víz, sör, bor, tej, ...}* | *{becsült mennyiség: pohár, csepp, korty, ...}*. Ezek a kapcsolatok a vonatkozó kategóriákba tartozó elemek gyakoriságainak függvényében eltérő mértékű asszociációt fejezhetnek ki az ige és a fogalomosztály között.

Az alábbiakban bemutatott módszerekkel megkíséreljük a különböző argumentumpozíciókra jellemző szemantikai kategóriákat korpuszadatok alapján automatikusan megtalálni, és ezeket a Magyar WordNetben új ige-főnév relációkkal ábrázolni. Az új relációtípus minden példányához két tulajdonságot szeretnénk társítani: egyrészt a vonzatpozíciót leíró morfoszintaktikai megkötéseket (esetrag vagy névutó), másrészt a a korpuszban mért adatok kiszámított kapcsolati erősségét, melynek célja az azonos pozícióban megadható szemantikai osztályok egymáshoz képesti szerepének számszerűsítése. Például az *{iszik}-[case=ACC, p=0,8]-{folyadék}, {iszik}-[case=ACC, p=0,2]-{becsült mennyiség}* két olyan kapcsolatot jelöl, amely az *iszik* ige két, tárgyesetű vonzatpozíciójában megfigyelt szemantikai kategóriát ad meg. A *{folyadék}* és a *{becsült mennyiség}* synsetek itt önmagukon kívül összes indirekt hiponimáikat is reprezentálják, így megadnak egy-egy fogalomosztályt.

3 Kapcsolódó munka

A szelekciós preferenciák feltérképezése kulcsfontosságú az írott nyelv szemantikai feldolgozása szempontjából. A vizsgálatok célja annak megállapítása, hogy milyen szójelentések gyakoriak és/vagy megengedettek bizonyos szavak adott szintaktikai környezetében. Resnik [12, 13] munkáját követve több tanulmány is a WordNetre támaszkodott a szelekciós preferenciák megállapításában ([2, 3, 22]).

Míg az utóbbi időkben ismertetett megközelítések a Latent Dirichlet Allocation (LDA) módszerekre koncentáltak ([15, 6, 14]), az általunk bemutatott kísérlet [13]-hoz áll közelebb. A magyar nyelv esetében elsőként kíséreljük meg az igék szelekciós tulajdonságainak automatikus feltérképezését. Munkánk nem csupán az ige-tárgy (direct object) viszony szelekciós megkötésének klasszikus problémájával foglalkozik, hanem figyelembe vesszük az összes lehetséges szintaktikai argumentumtípust is (20 fölött szám a magyarban), [1] javaslatával összhangban.

Szemben azokkal a megközelítésekkel, melyek célja csupán adott argumentumszerepben előforduló szavak halmazának azonosítása (pl. [4, 17, 14]), a [13] által felvázolt és [6] által is követett iránynak megfelelően kutatásunk célja szemantikus osztályok (típusok) címkéinek hozzárendelése az argumentumpozíciókhoz. Ezt a rendelkezésünkre álló legnagyobb kiterjedésű magyar nyelvű nyelvi ontológia, a Magyar Wordnet fogalmai csomópontjainak és taxonómiai relációinak felhasználásával szándékozunk megvalósítani.

4 Módszerek

A feladat megoldására általunk alkalmazott eljárás bemenete egy szóhalmaz (egy adott ige mellett adott bővítménypozíciójában előforduló főnevek gyakorisági listája), kimenete pedig e bővítményeket reprezentáló (általánosító) HuWN synsetek súlyozott, rendezett listája. Mindegyik kimenő synset a belőle kiinduló, hiponima-relációval alkotott algráfot reprezentálja. A kimenő synseteknek az alábbi feltételeket kell minél teljesebb mértékben kielégíteniük:

Lefedettség: a synset, illetve hiponima-leszármazottai tartalmazzanak minél többet a korpuszbeli szavak közül.

Sűrűség: a synsetből kiinduló algráf minél kevesebbet olyan szót tartalmazzon, ami nincs benne az input szólistában.

Használható általánosítások: a synset és a belőle kiinduló hiponima-algráf fejezze ki az argumentpozícióba tartozó korpuszszavak jelentéseinek általánosítását, de ne legyen túl általános. Például, kevés haszna van, ha minden igei argumentumhoz az *{entitás}* fogalmat társítjuk, mivel keveset mond az egyes argumentumok szemantikai preferenciáinak sajátosságairól.

Automatikus jelentés-egyértelműsítés: ha egy igei argumentumként szereplő szónak a WN-ben több jelentése van (több synsetbe is tartozik), elvárjuk, hogy az algoritmus csak a releváns jelentés(ek) általánosításához tartozó kapcsolato(ka)t generálja. Például, az *iszik* tárgyaként előforduló *kávé* főnév két jelentésének hiponimái

közül ne a *{termés, gyümölcs}*, hanem az *{ital, italféle}* felé konvergáljon az általánosítás.

A fenti feltételek alapján javasolt algoritmusunk vázlatosan az alábbi lépésekből áll:

1. Először megkeressük az összes lehetséges synsetet, amik az input szavakat tartalmazzák (azok összes lehetséges jelentéseit), majd ezekből generáljuk a lehető leghosszabb, hipernima-reláció szerinti útvonalakat a WN gyökércsomópontjaiig. Minden, ezeken az útvonalakban bárhol szereplő csomópont (synset) a továbbiakban szemantikaiosztály-**jelölt** lesz.
2. Ezt követi a jelöltek **szűrése**: elvetjük azokat a jelölteket, amelyek csak egyetlen egy korpuszszót reprezentálnak és a korpuszszót tartalmazó synset (direkt vagy indirekt) hipernimái. Ezzel a lépéssel kiszűrjük az általánosítást nem hordozó jelölteket.
3. A következő lépésben pontozzuk a fennmaradó jelölteteket két tényező figyelembevételével: hány bemeneti szót **fednek le** és milyen **sűrű** a jelölt által megadott részgráf a bemeneti szavakra nézve (a részgráf által lefedett bemeneti szavakat tartalmazó synsetek számának és a részgráf csomópontjai számának hányadosa). Az alábbi képlettel határozzuk meg *c* synset-jelölt pontszámát (ahol *subgr(c)* a *c*-ből kiinduló hiponima-részgráf, I_c a *subgr(c)* által lefedett bemeneti korpuszszavak halmaza):

$$Score(c) = |I_c| \cdot \frac{|\{s \in subgr(c) : w \in s, w \in I_c\}|}{|subgr(c)|}$$

4. A pontozás alapján rangsorolt jelöltek közül az **N legjobbat** adjuk vissza. Ezen a ponton történhet a bemeneti szavak jelentés-egyértelműsítése: ha az N legjobb synset között van legalább kettő, ami ugyanannak a bemeneti szónak eltérő jelentéseit fedi le, akkor a (leg)magasabb ponttal bíró jelöltet tartjuk meg, a többit elvetjük. Ezt addig ismételjük, amíg nem marad több többértelműség.

A HuWN-be ezután felvehetjük az új relációkat, amelyekben az igei synseteket összekötjük a nyertes főnévi synsetekkel. A kérdéses vonzatra vonatkozó morfoszintaktikai információn felül megadjuk a kapcsolat erősségét is, melyet a lefedett szavak korpuszgyakoriságai alapján adhatunk meg (ld. 6. rész).

Az algoritmus futtatásához felhasználtuk a *Mazsola* igei bővítménytár [16] adatbázisait. A Mazsola a 187 millió szavas Magyar Nemzeti Szövegtár [20] alapján készült, 20,24 millió tagmondatban azonosították a finit igéket és az igei bővítményekként funkcionáló főnévi csoportokat, majd ezeket csoportosították szintaktikai jellemzők (esetrag, névutó) szerint.

Annak eldöntésére, hogy milyen igéknek milyen vonzatai vannak, felhasználtuk a *MetaMorpho* magyar-angol fordítóprogram szintaktikai elemzőjében [11] használt igei vonzatkeret-leíró adatbázis anyagát is. Az adatbázis több mint 18 ezer igéhez 33 ezer vonzatkeret-leírást tartalmaz, melyek megadják az adott jelentésben szereplő lehetséges vonzatpozíciókat és az azokra érvényes, attribútumokkal kifejezett lexikai, morfológiai és szintaktikai megkötéseket. A Magyar WordNet fejlesztése során az igei synsetekhez hozzárendelték ebből az adatbázisból a megfelelő vonzatkeret-

leírásokat is [9]. Ez az információ felhasználható az új ige-főnév relációk létrehozásakor az igei synsetek egyértelmű kijelölésében.

A fentiek segítségével 25 500 különböző igei vonzatkeret 32 000 lehetséges argumentumpozíciójához készítettünk szógyakorisági listákat, melyeken futtatni tudtuk szelekciós preferenciákat általánosító algoritmusunkat.

5 Eredmények

Mivel jelenleg még dolgozunk egy olyan kiértékelési módszertanon, melynek segítségével az algoritmus eredményét humán annotátorok ítéleteivel tudnánk összevetni, eredményeink szemléltetésére bemutatunk néhány kiragadott példát.

Az 1. táblázatban felsoroltunk 6 kiválasztott igei vonzatpozíciót és az algoritmusunk segítségével hozzájuk rendelt, legnagyobb ponttal rendelkező szemantikai osztályt (HuWN synseteket).

1. táblázat: Automatikusan azonosított szemantikai osztályok az igevonzatokhoz

Ige és vonzatpozíció	Szemantikai kategória
iszik ACC	{folyadék}
kigombol ACC	{ruha}
olvas ACC	{könyv}
ül SUP	{ülőbútor}
vádol INS	{bűncselekmény}
megold ACC	{nehézség}

A 2. táblázatban bemutatjuk az *iszik* ige tárgyesetű vonzatpozíciójához tartozó 5 legmagasabb pontot elérő szemantikai kategóriát, valamint ezek pontszámát, a lefedett korpuszszavak számát (c) és a kategória kiszámított sűrűségét (d).

2. táblázat: Az *iszik* tárgyesetű vonzatpozíciójához rendelt 5 legmagasabb pontot elérő synset

Pont	Szemantikai kategória	Lefedettség	Sűrűség
9,1	{folyadék}	26	0,35
8,796	{ital, italféle, italféleség}	25	0,351
4,888	{szeszes ital, szesz, ital, alkohol}	16	0,305
4,375	{rövidital, tömény ital, tömény szesz, tömény}	7	0,625
3,759	{táplálék, tápanyag}	28	0,134

A HuWN hierarchiáját megvizsgálva észrevehetjük, hogy a *{folyadék}* csomópont hipernimája az *{ital, italféle, italféleség}* fogalomnak, amely viszont hipernimája a *{szeszes ital, szesz, ital, alkohol}* synsetnek. Felmerül a kérdés, hogy ezek közül melyikhez (melyekhez) szeretnénk az *{iszik}* igei fogalmat (accusativusi minősítésű kapcsolattal) hozzárendelni? Ha a legáltalánosabb és legtöbb pontot szerzett fogalmat preferáljuk, akkor a *{folyadék}* synsetre esik a választásunk. Egy másik nézőpontból viszont az *{ital, italféle, italféleség}* relevánsabb lehet, hiszen nem minden folyadék alkalmas emberi fogyasztásra. Bizonyos alkalmazásokban viszont fontos információ

lehet a korpuszadatok tanúsága szerint a *{szeszes ital, szesz, ital, alkohol}* fogalommal megjelenő erős kapcsolat is. Azáltal, hogy meghagyjuk az *N* legmagasabb pontot elérő szemantikai kategóriát minden argumentumpozícióban, valamint ábrázoljuk ezek relatív asszociációs erősségét is, szándékaink szerint a létrehozott erőforrás jövőbeli felhasználói számára biztosítjuk a lehetőséget arra, hogy céljaik és igényeik szerint maguk hozhassák meg ezeket a döntéseket.

6 További munka

Jelenleg módszereink továbbfejlesztésén dolgozunk. Amint elérhetővé válik egy kiértékelési metodológia, lehetséges lesz a jelölteket pontozó formula további finomhangolása, valamint kísérletezhetünk a kapcsolati erősségek beállításának optimális módjával is. További, felhasználható információk a bemeneti szavak korpuszbeli gyakoriságai, a jelölt synsetek mélysége a HuWN hálózatában és az átlagos távolságok a jelölt algráfokban.

Amint láttuk, a fent vázolt megközelítésben ige-vonzat párokhoz rendeltük hozzá az abban a pozícióban előforduló főnevek listáját, és az alapján határoztuk meg a szemantikai preferenciákat leíró legvalószínűbb HuWN synseteket. Az igék bővítményei azonban kölcsönhatásban vannak egymással: gyakran előfordul, hogy az egyik bővítmény megkötése (adott szóval való kitöltése) esetén egy másik bővítményben egy speciális (csak az első bővítményben lévő szóra jellemző) szelekciós preferenciával találkozunk. Ilyen például az 'ad -t' esetén a 'hírt ad' -rÓl bővítménye, vagy a 'húz -t' esetén a 'hasznot húz' -bÓl bővítménye. Ahogy azt [19] is hangsúlyozza, fontosnak tartjuk, hogy továbblépjünk a több bővítményt egyszerre kezelni tudó modellek felé, melyek képesek felismerni a 'hírt ad', 'hasznot húz' stb. összetett egységeket és ezek argumentumainak szelekciós preferenciáit.

Mechura [8] szerint a WordNetben található egységek nem teljesen felelnek meg a szelekciós preferenciák által megkívánt egységeknek, és felteszi a kérdést: hogyan kellene egy ontológiának kinézni ahhoz, hogy a szelekciós preferenciákban szerepet játszó szemantikai típusokat pontosan tudja ábrázolni? Az algoritmusunk segítségével előállított kategóriák vizsgálata elvezethet a válaszhoz.

7 Összefoglalás

A tanulmányban bemutattunk egy módszert a Magyar WordNet automatikus kiegészítésére új, szelekciós preferenciákat ábrázoló relációkkal, ami hasznos lehet a szövegfeldolgozó alkalmazások számára. Eredményeink érdekesek lehetnek a pszicholingvisztikai kutatások szempontjából is, mivel betekintést nyújthatnak a mentális lexikon szófajközi viszonyaiba.

Köszönetnyilvánítás

Köszönjük a TÁMOP-4.2.1.B – 11/2/KMR-2011–0002 és a TÁMOP: 4.2.2/B – 10/1–2010–0014 projektek részleges támogatását.

Hivatkozások

1. Brockmann, C., Lapata, M.: Evaluating and combining approaches to selectional preference acquisition. In: Proceedings of EACL (2003) 27–34
2. Calvo, H., Gelbukh, A., Kilgarriff, A.: Distributional Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. In: Proceedings of CI-CLing (2005) 177–188
3. Clark, S., Weir, D.: Class-Based Probability Estimation Using a Semantic Hierarchy. In: Computational Linguistics 28:2 (2002) 187–206
4. Erk, K.: A simple, similarity-based model for selectional preferences. In: Proceedings of ACL (2007) 216–223
5. Fellbaum, C. (szerk.): WordNet: An Electronic Lexical Database. MIT Press: Cambridge (1998)
6. Guo, W., Diab, M.: Improving Lexical Semantics for Sentential Semantics: Modeling Selectional Preference and Similar Words in a Latent Variable Model. In: Proceedings of NAACL-HLT (2013) 739–745
7. Kuti, J., Varasdi, K., Gyarmati, Á., Vajda, P.: Language Independent and Language Dependent Innovations in the Hungarian WordNet. In: Proc. of The Fourth Global WordNet Conference, Szeged, Hungary (2008) 254–268
8. Mechura, M.B.: What WordNet does not know about selectional preferences. In: Dykstra, A., Schoonheim, T. (szerk.): Proceedings of the 14th Euralex International Congress, Ljouwert/Leeuwarden: Fryske Akademy (2010) 431–436
9. 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. (szerk.) Proceedings of The Fourth Global WordNet Conference. Szeged: University of Szeged (2008) 311–321
10. Prószéky, G.: Kutatások egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozás irányában. In: Ladányi, M., Vladár, Zs. (szerk.) A XI. MANYE-konferencia előadásai (megjelenés alatt)
11. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator. In: Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: Foundation for International Studies (2004) 138–142
12. Resnik, P.: Selectional constraints: an information-theoretic model and its computational realization. Cognition 61 (1996) 127–159
13. Resnik, P.: WordNet and Class-Based Probabilities. In: Fellbaum (1998a)
14. Rink, B., Harabagiu, S.: The Impact of Selectional Preference Agreement on Semantic Relational Similarity. In: Proceedings of International Conference on Computational Semantics (IWCS) (2013)
15. Ritter, A., Mausam, Etzioni, O.: A latent dirichlet allocation method for selectional preferences. In: Proceedings of ACL (2010) 424–434
16. Sass, B.: The Verb Argument Browser. In: Sojka, P., Horák, A., Kopecek, I., Pala, K. (szerk.): 11th International Conference on Text, Speech and Dialog (TSD), Brno, Czech Republic. Lecture Notes in Computer Science 5246 (2008) 187–192

17. Tian, Z., Xiang, H., Liu, Z., Zheng, Q.: A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation. In: Proceedings of ACL (2013) 1169–1179
18. Tufiş, D., Cristea, D., Stamou, S.: Balka-Net: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, 7(1⁻2) (2004) 3–4
19. van de Cruys, T.: A non-negative tensor factorization model for selectional preference induction. In: Natural Language Engineering 16(4) (2010) 417–437
20. Váradi, T.: The Hungarian National Corpus. In: Zampolli, A. (szerk.) Proceedings of the Second International Conference on Language Resources and Evaluation. Las Palmas: ELRA (2002) 385–389
21. Vossen, P.: EuroWordNet General Document, Version 3. University of Amsterdam (1999)
22. Ye, P.: Selectional Preference Based Verb Sense Disambiguation Using WordNet. In: Proceedings of the Australasian Language Technology Workshop (2004)

Corpus-based Population of a Mid-level Business Ontology

András Kornai

MTA SZTAKI

Abstract. We describe the creation of a broad mid-level ontology, several thousand nodes, suitable for classification and analysis of business documents of the kind regularly kept in corporate document storage. The main claim of the paper is that we can populate a rich mid-level ontology by largely automatic, corpus-based methods.

1 Introduction

In Section 2 we begin by reviewing some standard notions, and describe the principles of what we will refer to as *Midlevel Business Ontology* (MBO). These principles guide the learning process that is used to extract an actual ontology of over 5k entities from a corpus of 20k documents of the kind found in corporate document storage: memos, activity plans, agendas, proposals, CVs, regulatory (legal) documents, accounting materials, bills, invoices, letters (including emails), etc. In Section 3 we describe the process of node selection, and in Section 4 we describe the data cleaning process. We believe our chief method, the iterative sharpening of linear classifiers, is also applicable to the problem of automatically building a rudimentary hierarchy among the entries, and we conclude the paper with some programmatic remarks on this.

2 Linking MBO to well-known upper ontologies

We assume, without argumentation, the standard tripartite division into high-, mid- and low-level ontologies. For the high level (also known as *upper, top,* or *foundational*) ontology, we use the 4lang ontology [1] now better called 40lang, inasmuch as bindings have been extended to 40 languages [2].

Perhaps the major division line among various ontologies is whether they are intended to capture knowledge about the world (e.g. about distinctions among various physical objects such as tools) or about conceptual entities. To put this another way, we must decide whether it is the difference between *hammers* and *nails* that we are intent on systematizing, or the difference between the concepts (words, mental/cognitive entities) 'hammer' and 'nail'. Since our interest is with the latter, our work is more closely related to ontologies like DOLCE than to word taxonomies like WordNet [3].

The very same object, say an MS Word file preserved on a particular floppy disk, can be a 'contract', a 'draft', or an 'exhibit', which means that very different rules apply to it – drafts can be modified at will, while tampering with evidence is a crime. At the same time, different objects, such as the file as it appears on the hard drive, in hardcopy, or in an email attachment preserved on a computer on another continent, may relevantly be called the same.

Cataloging physical objects remains a valid goal for ontologies, but to use MBO for this purpose it would have to be supplemented by some system of physical or logical coordinates which lies outside the business ontology proper. The main lesson we take away from physical objects is that none of them are true endurants: things have a beginning (creation process) and end (destruction process). This is evident for business objects like contracts or offers, but in MBO we treat more enduring abstract objects like laws and regulations the same way.

For a full ontology, we would need three kinds of information: pure generic, modified generic, and domain-specific. By *pure generic* information we mean the kind of very general statement that objects (typically, nouns) can be divided in two basic classes, 'physical' and 'conceptual', with mass, energy, and space-time coordinates easily attached to the former, but not the latter, while requirements, obligations, etc. are easily attached to the latter but not the former. Statements at this high level of generality apply within the business domain just as well as in any other domain we can think of, such as the medical or the legal domains, and thus belong in the top-level ontology.

With a thousand or so entries, 4lang is considerably richer than the philosophically inspired top-level ontologies, and contains many words that we call *modified generics*. For an example, consider *charge*, which is in a business context tied to fees 'merces', in a legal context to 'accusatio', both of which modify the general meaning of charge as some kind of attack 'impetus' quite substantially. The overlap between 4lang and the raw mid-level list is a rich source of examples of this phenomenon, but we find even more examples among words that are not considered basic and are thus not present in 4lang: consider for example the verb *to hedge*. Outside the business domain, this means 'to avoid giving a promise or direct answer' (Merriam-Webster), within the domain the prevalent meaning is 'to buy or sell commodity futures as a protection against loss due to price fluctuation'. The two meanings share the common element 'to evade the risk of commitment' but the technical means of carrying out the evasion are very different.

Finally, for an example of a *domain-specific* concept consider *budget* (both noun and verb). It is possible to use this word in another domain, e.g. a newspaper story about a boxing match may say that the loser didn't budget enough energy for the final round, but by doing so the writer invokes the business metaphor (a reversal of the more common 'business is war' trope). To capture the truly business-specific vocabulary we need to proceed top down, building out some core scripts, such as `retail business`, where products are sold to customers, `rental business`, where products are leased or rented to customers, `service`, `market`, and so forth. All these core scripts have the same typecasting power

over their components: we may normally think of surgical wards in the medical context, where an appendictomy is 'an urgent life-saving procedure', but we may also think of them as retail stores, where appendictomies are products, sold to customers. These customers happen to be called 'patients', but business is business, we first need to establish their capacity to pay.

Since our overarching goal is to establish the business-specific concepts, including the business-specific readings of generic concepts, with as little human intervention as possible, we need to divide the corpus into documents that are, ideally, reflective only of a single core category such as `retail`. For this we need to clean the corpus of material that belongs, according to human judgment, to two or more (sub)domains at the same type: typical examples would be a document that describes the procedure for testing financial software, as it belongs both to `banking` and `IT`, or plans for customer-facing operations (`retail`) for an organization that normally operates upstream (`wholesale`).

3 Automatic acquisition of nodes

Part of the challenge in building mid-level ontologies comes from above, linking with the top level, and part comes from below, in trying to link with specific low-level ontologies and knowledge bases. But there are two challenges that are intrinsic to the middle level: populating the domain ontology, and keeping it free from material from other mid- or low-level domains. One way to build MBO would be based on introspection, but it is hard to find experienced businessmen who are also sophisticated lexicographers, ontology builders, and knowledge engineers at the same time. Here we describe how we can select a rich mid-level ontology based on a corpus, and defer the issue of keeping out polluting material to Section 3.

We begin with the CS corpus of about 20k documents selected randomly from the servers of a well known multinational firm (over EUR 10 billion in annual revenues and over 100k employees) that offers professional services to other businesses, guaranteeing that the vocabulary extracted from it is not restricted to any vertical. (As it currently stands, the CS corpus is not available to the public, but efforts to suitably anonymize it are under way.) The 27m word tokens are in 453k types, of which 216,450 occur more than once (hapax legomena are discarded). The rest was compared to the Google 1T vocabulary [4] in several steps.

First, we considered the words unique to the CS corpus and order these by frequency. At the top we only find expressions such as *N/A* or *follow-up* which are missing from G1T only because Google is using a different tokenization algorithm, which splits on slash and hyphen. In fact, over 70% of the 103k words that do not appear in G1T are the result of such mismatches, and the remainder is dominated by token classes whose individual words are of little interest, such as numbers like *56101363*; SQL and other programming language keywords such as *VARCHAR25*; and table column headers like *StateIncluded*. Once these Information Technology (IT) words are discarded, by a data cleaning

process we defer to Section 3, we are left with only 1,722 words, the majority of which are foreign, typically French, Italian, German, and Spanish, reflecting the international nature of business. If these are removed and typos are discarded, we are left with only 85 words (in order of decreasing frequency):

subinventories, preadmit, autocash, promptable, billdate, tradelane, termdate, userviews, coverdoc, workrequest, substatuses, ratecode, preadmitted, megaprocesses, finaldoc, callbase, autoinvoicing, autoaccounting, acceptancetest, totalcharges, totalbarrels, recruitability, minispecs, invoiceless, soustotals, salesorders, workstructure, videocypher, sidemarking, preadmits, modelclass, modelcategory, desginator, wellnumber, prebonus, blueplate, waybilling, subnetworked, subledgering, subinstitutional, strawmans, stocknumbers, recoupability, rebillable, reapproves, prebilled, postbilling, outcomedoc, multifacilities, memodoc, intraoperation, hitchment, budgetxls, btuvalue, unissue, shipvendor, shiftwise, sheetmetals, settlementdoc, settlebatch, screenpainter, saleorder, retrieveability, reputs, reportxls, reportsdoc, palettization, nonclearable, newquantity, matrixtesting, matrixdoc, matricesdoc, materialsql, masterdoc, manweeks, knowledgeweb, jobchangeover, inputdoc, guidelinesdoc, detailable, dealsheets, bundletracker, autosourcing, autoscheduled.

Many of these are either whitespace errors or, more likely, also column headings: *term date, work request, settlement doc,* etc. With a high quality morphological analyzer we can find many others that appear in G1T in their citation (stem) form: *subinventory, substatus, preadmit,* etc., and once these are taken into account, we are left with a handful of compounds and latinate formations (particularly prefixes *pre-, sub-, un-, re-, intra-,* see [5]) that are truly characteristic of business vocabulary. Overall, words that are missing from G1T are not a significant source of MBO candidates.

Next we consider those 103k words that appear both in CS with absolute frequency $F > 1$ and in G1T with absolute frequency $G > 100$ (the cutoff of the Google count). We exclude proper names (since the corpus is not yet anonymized), which reduces the corpus to 30k word forms. Since the G1T corpus is much larger (by a factor of about 25,000), $\log(G/F)$ is on the average 9.93, with a variance of 2.22. Therefore, it makes sense to restrict attention to those words where this number is below average, i.e. those words that are used *at least as often* in business documents as in general English. Only 14k word forms meet this criterion, and a quarter of these are foreign. We can remove the bulk of these by prefiltering the corpus for language.

Of the remaining 10,227 words we consider only those 6,039 that appear at least in 9 documents. The publicly available mid-level entity list (for which see `http://hlt.sztaki.hu/resources`) is cleaned of typos (including proper names that were left uncapitalized) but not fully stemmed. Since this is a departure from standard lexicographic practice, let us briefly describe the reason for keeping non-stemmed (often derived, but sometimes even inflected) forms. Consider, for example, the adjective *yearly*, obtained from the stem *year* by an entirely regular, highly productive suffix. Since there is nothing business-specific about the word *year* or the way this word is used in business documents, it clearly doesn't belong

in MBO. But in the business context *yearly* carries a sense of obligation that is missing from the generic use – iceberg formation or stork migration happen yearly, but are not obligations. This is quite consistent with the fact that the relative frequency of *year* is the same in the business domain as in English in general, while the frequency of *yearly* is almost twice (1.92 times) as large.

In fact, higher than expected proportion of derived forms is a good predictor of domain-specificity. Consider a plural like *customizations* or a past tense like *architected* – these are far less likely in environments where *customization* is not a frequent noun and *to architect* is not a frequent verb to begin with. Though random spot-checks of material from other domains bear out the validity of this observation, we have something of a chicken-and-egg problem here, in that we cannot at the same time claim that our material proves the observation and use the observation to select the material. In this paper, we chose to take the validity of our observation on faith, and use it instrumentally to select the MBO nodes – independent verification must await the public availability of domain-specific corpora and their independently arrived at mid-level ontologies.

4 Cleaning the data

Ideally, we would want to begin with a few well understood scripts (in the sense of Schank and Abelson [6]) such as 'selling', 'investing', and other prominent business activities, but this would again lead us to the problem that we started out with, that there are very few domain experts who are also knowledge engineers. Thus we seek a less perfect automated or semi-automated solution, one that clusters the mid-level data in script- or frame-sized subdomains, ideally with minimal overlap. Of course eliminating overlap cannot be taken to the extreme: every business operates in some domain, often more than one, and if we omitted every `retail` document that is about `apparel` or `automotive` or similar verticals we'd be left with nothing.

Manual inspection of the raw entity list made clear that we have a significant number of documents containing terms that are highly specific to information technology (`IT`): not just programming terms like *alloc, atoi, fflush, fprintf, ...* but also expressions associated to the high-level planning stages such as *alphanum, autoexec, flowchart, gigabyte, groupware,* etc. Here we had to make a strategic decision, whether to treat IT as yet another business domain, or segregate the IT-specific vocabulary. Since our data was obtained via IT consulting, in the interest of a balanced ontology we chose the latter method, but we emphasize that the algorithm used for doing so is just as applicable to the `IT` versus `non-IT` decision as it is to `retail` versus `wholesale`.

In stage 0, we begin with a manually selected seed list of IT-specific words such as the ones listed above, and observe their probability in the corpus. We compute a simple but effective linear classifier (see [7]) that uses the *relevance* (defined as the difference between the log frequency in the positive set and the log frequency in the background model) of keywords and key phrases for weights, retaining only those keywords/phrases whose relevance exceeds some *threshold*

of significance τ, say $\tau = 3$. At this stage, we use the G1T count for background. The stage 0 classifier is thus a simple relevance-weighted word vector, which is multiplied with the TF vector of each document to obtain a *raw score* that gets normalized by dividing it by $n^{0.8}$, where n is the number of words in the document. (Here 0.8 is the Herdan-Heaps exponent, see [8] and [9]).

In stage 1, we rank the documents by the stage 0 classifier, and cut off the list by manual inspection so as to include only evidently IT-specific documents. Techniques for automating this step are of great interest, but would take us far from our immediate goals of populating the ontology and building the knowledge base. We now repeat the frequency count on the selected documents, and rerank the words, using either G1T or the overall corpus frequencies as background for establishing the relevances. The process can be iterated as many times as we wish, limited only by our ability to cut off the document lists (which is easy by binary search). A list of some 80 highly IT-specific terms obtained this way is included here:

abend abends alphanum autocreate autofill configurator customization customizations datafield datafiles datawindow datawindows dbase dbms deliverables dialer downtime esc fileserver flowchart flowsheet flowsheets fprintf functionality indirects inputters intercompany interfaced jobcode jpl keytab mainframes maint masterfiles matchcode matchcodes matl middleware mmddyy mmddyyyy parm parms pcs procs pseudocode redisplay redisplayed redisplays reformats routings rowid rqmt runscript signoff signoffs signon spoolfile sprintf sqlplus sqr strcat strcpy strncat strncmp strncpy strupr submenu subprocesses subsystem sybase systime tabbing tableset tablespace timestamp tinyint toupper userview varchar.
As we discussed in Section 2, such lists are likely to contain many terms like *redisplay redisplayed redisplays* that stemming would collapse in a single term, but this would not be desirable in that domain-specific terms like *indirects* would by such a process be reduced to terms like *indirect* that are no longer specific to the domain. Practical experience with these classifiers shows that removing all but the top d keys ($20 \leq d \leq 200$) by aggressive thresholding decreases the recall of the classifiers by very little and their precision even less, and that the key issue driving performance is the choice of words/phrases kept rather than their exact weight.

The algorithm is best analyzed in the frame of PAC-learning [10]. Our *sample space* S is the corpus, our *concept* C to learn was IT above, but could be any other mid-level concept like `insurance`. We are interested in learning the concept with $1-\delta$ probability and ϵ precision, with δ, ϵ in the 1-10% range, which is practically feasible, even though the theoretical bound based on VC dimension ($d+1$ for a linear classifier) falls short of what we want for this size ($N \approx 20,000$) data set.

5 Conclusions, future directions

The main claim of this paper was that from a raw corpus of some 20k business documents we can populate a sizeable mid-level ontology with minimal human intervention. While we cannot at present make the corpus publicly accessible

(anonymization is still under way), we make the the raw concept list of 5,779 entries downloadable from `http://hlt.sztaki.hu/resources`.

This list, for the reasons discussed in the paper, is still a mixture of morphologically complex (derived, inflected) and morphologically simplex (stem) forms. By automatic stemming we would lose both ontological insight (e.g. that *versioning* is not just the process of making versions) and discriminative power in classification tasks. Once the hierarchization is complete, we expect the list to shrink to half of its current size, still quite large for a mid-level ontology.

The next steps are building the hierarchy and attaching definitions to each concept. Our plan is to generalize PAC concept learning to the case of learning several concepts together. Broadly speaking, instead of a single linear classifier and the attached document set we try to bootstrap k classifiers such that the associated k document sets are largely disjoint and exhaustive. For each domain we start with small, manually created seed lists e.g. for `retail` we would have *customer discount price purchase retail seller store*, for `banking` we would have *account, atm, cd, checking, loan, savings* and so forth, for a few dozen subdomains.

In each iteration, we cluster the documents, and investigate how well the classifiers capture these. For now, we have now way of automating this manual supervision step, but we note that such spotchecks require a great deal less labor than manually classifying the entire corpus to different subdomains.

Acknowledgment

We thank Judit Ács and Attila Zséder for their help at various stages of the pipeline.

References

1. Kornai, A., Makrai, M.: A 4lang fogalmi szótár [The 4lang concept dictionary]. In Tanács, A., Vincze, V., eds.: IX. Magyar Számitógépes Nyelvészeti Konferencia [Ninth Conference on Hungarian Computational Linguistics]. (2013) 62–70
2. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
3. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Sweetening WordNet with DOLCE. AI Magazine **24**(3) (2003) 13–24
4. Brants, T., Franz, A.: Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia (2006)
5. Aronoff, M.: Word Formation in Generative Grammar. MIT Press (1976)
6. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures. Lawrence Erlbaum, Hillsdale, NJ (1977)
7. Kornai, A., Krellenstein, M., Mulligan, M., Twomey, D., Veress, F., Wysoker, A.: Classifying the Hungarian Web. In Copestake, A., Hajic, J., eds.: Proceedings of the EACL. (2003) 203–210
8. Herdan, G.: Quantitative linguistics. Butterworths, Washington (1964)

9. Heaps, H.S.: Information Retrieval – Computational and Theoretical Aspects. Academic Press (1978)
10. Valiant, L.G.: A theory of the learnable. Communications of the ACM **27**(11) (1984) 1134–1142

IV. Pszichológia

A nyelvi kategória modell kategóriáinak automatikus elemzése angol nyelvű szövegben

Pólya Tibor[1], Kővágó Pál[2], Szász Levente[2]
[1] Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu
[2] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
kovago.pal@ttk.mta.hu
szasz.levente@ttk.mta.hu

Kivonat: A nyelvi kategória modell a hétköznapi nyelvhasználat szociálpszichológiai kutatásának egyik leggyakrabban használt elemzési eszköze és elmélete. A modell az interperszonális cselekvés leírásában megjelenő absztrakció 5 kategóriáját különbözteti meg. A tanulmányban a modell által meghatározott kategóriák automatikus azonosítására képes eszközt mutatunk be. Az elemzés első lépéseként a szöveg szófaji és szintaktikai elemzését a coreNLP végzi el. A második lépésben az absztrakciós kategóriák felismerését a NooJ szoftverben írt gráfok végzik el. Végül az elemzés harmadik lépése lehetőséget ad arra, hogy a felhasználó különböző csoportokba sorolja a találatokat.

1 A nyelvi kategória modell

A hétköznapi nyelvhasználat szociálpszichológiai kutatásainak egyik leggyakrabban használt elmélete és elemzési eszköze a Semin és Fiedler nevéhez köthető nyelvi kategória modell [8] (angolul Linguistic Category Model, rövidítve: LCM). A nyelvi kategória modell az interperszonális cselekvések leírásának konkrét-absztrakt dimenzió mentén elhelyezhető változatait ragadja meg. A modell szerint az interperszonális cselekvéseket az absztrakció öt szintjén írhatjuk le. A legkonkrétabb fogalmazásmód a leíró cselekvő igével (descriptive action verb, rövidítve: DAV) történő leírás. Például: „Józsi *megüti* Gézát”. A leíró cselekvő igék mindig egy cselekvésre vonatkoznak. A cselekvés kezdete és vége egyértelműen azonosítható. A cselekvésnek van invariáns fizikai jellemzője. Végül önmagában a leíró cselekvő igéknek nincs értékelő jelentése.

Ennél absztraktabb az értelmező cselekvő ige (interpretative action verb, rövidítve: IAV) felhasználásával történő leírás. Például: „Józsi bántja Gézát”. Az értelmező cselekvő igék több azonos cselekvésre vonatkoznak. A cselekvés kezdete és vége szintén egyértelműen azonosítható, de a cselekvésnek nincs egyértelmű invariáns fizikai jellemzője. Az értelmező cselekvő igék esetében a negatív vagy pozitív irányú értelmező mozzanat már tetten érhető.

Az állapotot kifejező cselekvő igék (state action verb, rövidítve: SAV) a IAV-ok közeli rokonai, absztraktságuk szintje az értelmező cselekvő igékkel azonosnak tekinthető. Például: „Józsi *felbőszíti* Gézát." A cselekvés állapot igék egyedi eseményekre vagy események csoportjára vonatkoznak, de a leírás a cselekvés érzelmi következményeire irányítja a figyelmet. A leírt cselekvés ebben az esetben is egyértelmű kezdettel, illetve befejezéssel rendelkezik, de a cselekvés állapot igének önmagában értékelő jelentése van.

Az állapotjelző igék (state verb, rövidítve: SV) hosszan fennálló kognitív vagy érzelmi állapotot írnak le, így kezdetük és befejezésük nem azonosítható. Például: „Józsi *utálja* Gézát".

A legabsztraktabb kategória a cselekvés melléknévvel (ADJ) történő leírása. Például: „Józsi *agresszív*." Ilyenkor a leírás azt implikálja, hogy a cselekvés a célszemély állandó, belső személyes tulajdonsága miatt jött létre.

A nyelvi kategória modellnek két kódolási útmutatója létezik. Az egyiket Klaus Fiedler és munkatársai [7] készítették, a másikat Gün Semin és munkatársai [1] hozták létre. Az automatikus elemző kidolgozása során az elsőként említett leírást követtük.

A szociálpszichológiai vizsgálatok eredményei szerint az interperszonális cselekvés leírásának absztraktsága magyarázó erővel bír például az attribúciós következtetések [8], a sztereotípiák terjedésének módjával [9] és a csoportközi elfogultsággal kapcsolatban. Utóbbit Maass és munkatársai [4,5] tették vizsgálódásuk tárgyává. Kutatásuk eredményeként jött létre a nyelvi csoportközi elfogultság (linguistic intergroup bias, rövidítve: LIB) fogalma. Univerzális emberi jelenség, hogy önértékelésünk egyik fontos összetevőjét azok a csoportok adják, amelyeknek mi is a tagjai vagyunk [11]. A pozitív önértékelésre való törekvés elvéből következően a saját csoport tagjainak viselkedését úgy próbáljuk láttatni, hogy annak pozitív cselekedetei belső okokkal legyenek magyarázhatók, míg a negatív megnyilvánulásait külső, szituációs tényezőknek lehessen tulajdonítani. Ezt nyelvi szinten úgy érjük el, hogy a pozitív cselekedeteket absztraktabban fogalmazzuk meg a negatív cselekedetekhez képest. A külső csoport esetén is hasonló „logika" mentén járunk el, csak éppen fordítva. Azt szeretnénk, hogy a külső csoport rosszabb minőségben tűnjön fel a saját csoportunkhoz képest, ezért annak negatív tetteit absztraktabban, pozitív cselekvéseit pedig konkrétabban fogalmazzuk meg.

Az interperszonális cselekvések leírásában tetten érhető absztraktság szociálpszichológiai vizsgálatainak többsége úgy jár el, hogy az ingeranyagként adott mondatok absztraktságát variálva azonosítja annak hatásait. Hosszabb szövegek absztraktságának kódolása nagy kihívást jelent az empirikus vizsgálatok számára, mivel ehhez akár több száz igét kell kategorizálni. Az általunk kidolgozott elemzési eszköz célja az, hogy megbízhatóan képes legyen nagy terjedelmű szövegben előforduló interperszonális cselekvések absztraktságának megállapítására.

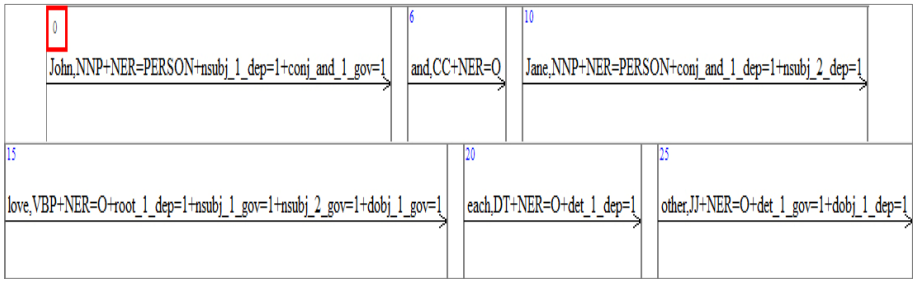
2 A nyelvi kategória modell kategóriáinak automatikus elemzése

Kézenfekvőnek tűnhet, hogy egy szófaji alapon nyugvó kategóriarendszer automatizálása egyszerűen megoldható szótár alapú keresőkkel. Ahhoz azonban, hogy az

elemzés szociálpszichológiai mondanivalóval is bírjon, nem elegendő tudnunk a nyelvi kategória modell kategóriáinak előfordulási gyakoriságát, hanem azt is tudnunk kell, hogy az adott absztrakciós szintű szóalak melyik szereplőhöz tartozik. Annak érdekében, hogy a megtalált ige vagy melléknév összeköthető legyen a cselekvő argumentumával vagy a minősített személlyel, ismernünk kell a szöveg szintaktikai szerkezeti jellemzőit. Egy ilyen elemző használata ráadásul minimalizálja a szavak azonos alakúságából fakadó hibákat is.

Az általunk elkészített angol nyelvű automatizált LCM elemző tehát nem egyszerűen szótár alapon keresi ki és kategorizálja a szövegben előforduló állítmányokat, hanem szintaktikai adatokra támaszkodva hozza összefüggésbe azokat alanyukkal. A melléknévi kategória esetén azt a tárgyat vagy személyt is képes azonosítani az elemző, amelyhez kapcsolódik az adott melléknév.

Az elemzés három lépésben történik. Az első lépésben a szöveg POS taggelését, a tulajdonnevek felismerését és a szöveg szintaktikai elemzését a coreNLP látja el [2, 13]. Az outputként kapott XML formátumú fájlt egy XSLT fordítóval[1] transzformáljuk, hogy a NooJ [10] külön tudja választani a szöveget és annak annotációit. A szöveg annotációi ebben az esetben szavanként tartalmaznak egy POS taget, egy NER értéket, illetve minden egyes függőségi kapcsolatot, amelyet az adott szó a coreNLP által megkapott. A coreNLP szintaktikai elemzője a mondat szerkezetét szópárok egymáshoz való viszonyának jelölésével képezi le. Az alany-állítmányi kapcsolatban például az állítmány ún. „nsubj governor”, az alany pedig „nsubj dependent” annotációt kap. Egy szó több ilyen kategóriát is kaphat, hiszen például egy állítmányhoz több alany is kapcsolódhat. A NooJ nyelvi elemzőben definiálható szabályok sajátosságai miatt ahhoz, hogy össze tudjuk kötni, mely szavak alkotnak egy szintaktikai szópárt, minden szintaktikai pár kap indexként egy számot. Amikor tehát két alanya van egy állítmánynak, az állítmány két nsubj governor szintaktikai kategóriát kap, melyeket 1 és 2 indexszel látunk el az XML fordítás során. Ugyanezt a két indexet fogja megkapni az első és a második alanya az állítmánynak (lásd. 1 ábra).



1. ábra: „John and Jane love each other.” mondat coreNLP általi elemzésének bemenete a NooJ szoftverben

1 Az XML fordításban közreműködött Matuszka Tamás és Rácz Gábor

A második lépésben a coreNLP-ben elemzett szöveget a NooJ-ban elkészített LCM nevű gráffal elemezzük tovább. Ahogy az az 1. ábrán látható a „John" és „love" szavak, illetve a „Jane" és „love" szóalakok nsubj dependency kategóriával kapcsolódnak össze. A példában szereplő „love" ige az állapotjelző ige (SV) LCM kategóriába kerül. Ez az információ egy háttérszótárnak köszönhetően áll rendelkezésre, melyet a fejlesztés korai szakaszában hoztunk létre. A szótár összeállításához a British National Corpus[2] adatbázisát használtuk fel. A leggyakoribb 6318 szótő listájából [3] kigyűjtöttünk az igéket. A listán 1281 ige szerepelt. A legtöbb igének több jelentése is van. A kódolás során az igék leggyakoribb jelentése alapján végeztük el a kategorizálást. Az igék leggyakoribb jelentését a The Longman Dictionary of Contemporary English Online [12] alapján választottuk ki. Az igéket két független kódoló kategorizálta be a nyelvi kategória modell 4 igekategóriájába. A kódolók közötti egyet nem értést egy harmadik kódoló bevonásával oldottuk fel. Tapasztalataink szerint a leggyakoribb igék használata önmagában magas találati arányok elérését teszi lehetővé, azonban a háttér szótárakat könnyedén bővíthetjük a vizsgálatunkban szereplő szövegekben előforduló speciális szavakkal. A melléknevek azonosítására a coreNLP POS taggerét alkalmaztuk.

Az általunk elkészített LCM NooJ gráf kategóriába sorolja a szövegben előforduló azon igéket, amelyek szerepelnek a háttérszótárban. A kategóriába soroláshoz az ige szótövén kívül felhasználjuk a POS tag-et és a szintaktikai elemzés eredményét is. A 2. ábrán látható példánál a mondat egyszerűségéből következően az állítmányi pozícióban levő igét kell megtalálnia a gráfnak, majd egy összekapcsoló gráf párba állítja az azonos indexszel szereplő állítmányokat és alanyokat, illetve jelzős szószerkezeteket egy mondaton belül.

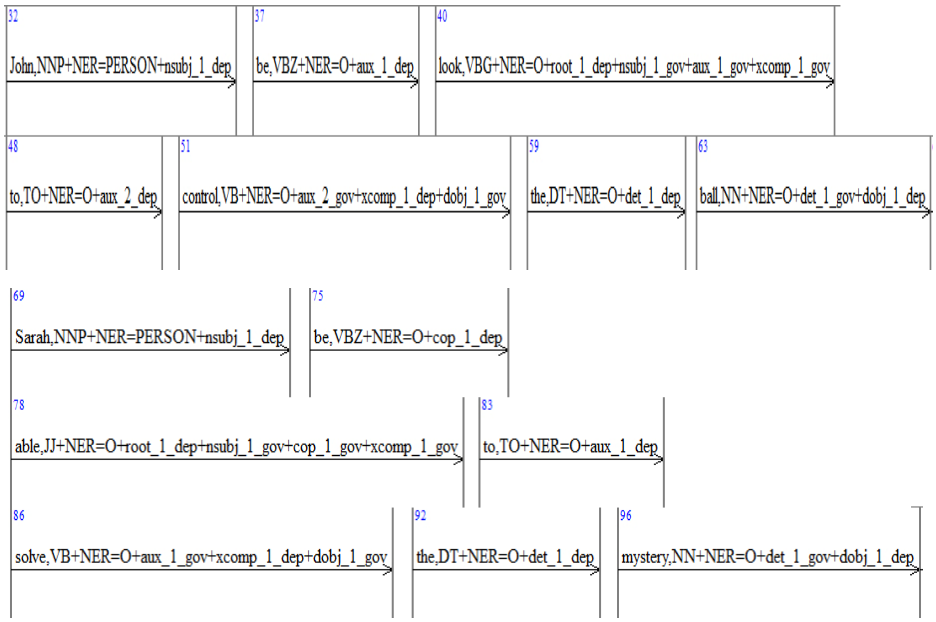
2. ábra: „John and Jane love each other." Az LCM gráf működése egy példán keresztül. A gráf először megtalálja a „love" állítmányt, majd összeköti azt a két alanyával.

Az elemzés harmadik és egyben utolsó lépése egy manuális elemzés a NooJ által megadott konkordancia lista segítségével. A konkordancia listában LCM kategóriába sorolva szerepelnek a találatok, illetve az azokhoz kapcsolódó alanyok vagy minősített entitások. A konkordancia adatok alapján manuálisan döntést hozhat az elemzés végzője arról, hogy tovább szűkíti-e a találatokat. Például elképzelhető, hogy az elemzés végzője csak azokat a találatokat veszi figyelembe, amelyek élő személyek

2 http://www.natcorp.ox.ac.uk

által végrehajtott cselekvéseket írnak le. A nyelvi kategória modellt alkalmazó szociálpszichológusok között nincs egyetértés abban, hogy általában a cselekvés vagy csak az interperszonális cselekvés az, ami elemzendő a szövegben. Szintén indokolt lehet az, hogy az elemzés végzője külön csoportba sorolva veszi figyelembe a saját és a külső csoport tagjainak cselekvésében megjelenő absztrakciót.

Az eddigiekben csak olyan esetekről szóltunk, amikor a megtalálandó ige állítmányi pozícióban van a mondatban. A következőkben két olyan példát mutatunk be, ahol a megtalálandó ige nem kap „nsubj depedency" kategóriát. Ez fakadhat a coreNLP elemzési sajátosságaiból vagy abból, hogy az adott ige valóban nem állítmányi pozíciót foglal el a mondatban. Ilyen esetekben az elemző célja összekapcsolni az igét azzal az entitással, amire vonatkozik, erre láthatunk példát a 3. ábrán.

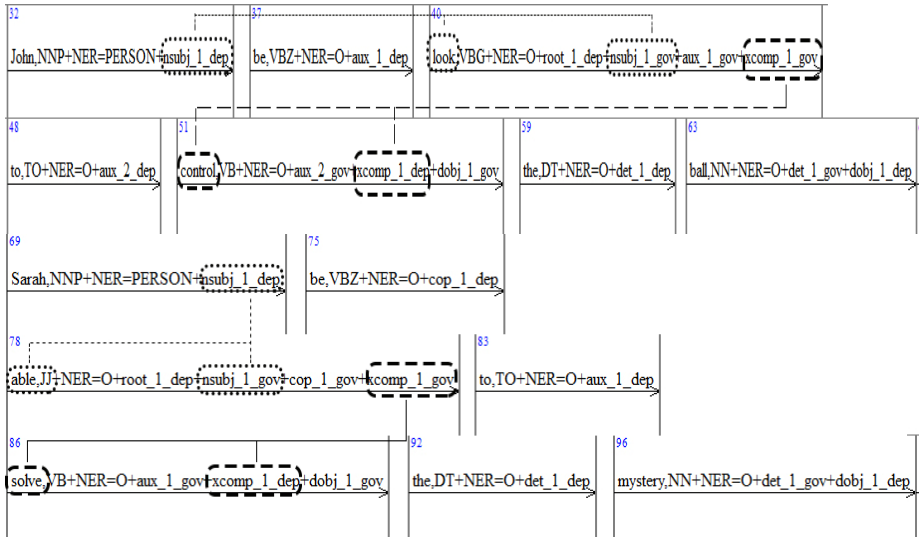


3. ábra: Két példamondat, ahol a megtalálandó ige nem közvetlenül kapcsolódik a cselekvőhöz. „John is looking to control the ball." és "Sarah is able to solve the mystery." mondatok coreNLP általi elemzésének bemenete a NooJ szoftverben

Az első mondatnál az „nsubj" kategóriát a „looking" „linking verb" [3] fogja megkapni, a másodiknál pedig nem egy ige, hanem egy melléknév: „able". Ezekben az esetekben az általunk elkészített LCM NooJ gráf megtalálja a nyelvi kategória modell szempontjából releváns igéket: az első mondat esetében a „control" IAV kategóriájú igét, a második mondat esetében a „solve" IAV kategóriájú igét. A „control" és a

[3] A „linking verb" olyan szó vagy kifejezés, amely egy mondatban az alanyt és a hozzá tartozó állítmányt kapcsolja össze.

„solve” igék az „xcomp dependency” [4] kategóriával kapcsolódnak a coreNLP által megjelölt állítmányokhoz. A gráf ebben az esetben összekapcsolja az „xcomp dependency” kategória segítségével a megtalálandó igéket a mondat állítmányával úgy, hogy az ideiglenesen megadott LCM kategória az állítmány indexét vigye tovább annak érdekében, hogy az alapesetnek vett alany-állítmányi szerkezetnek megfelelő módon összekapcsolható legyen a számunkra fontos ige azzal az entitással, amire vonatkozik (lásd 4. ábra).



4. ábra: „John is looking to control the ball.” és “Sarah is able to solve the mystery.” mondatok elemzése az LCM gráffal. A gráf elsőként a szaggatott vonallal jelölt elemeket találja meg a háttérszótárak segítségével. Második lépésben ezeket köti össze a pontozott vonallal jelölt entitásokkal

Fontos megemlíteni, hogy a gráf jelenlegi verziójában a „looking to control” szerkezet téves találatot is hozni fog, hiszen a „looking” igét meg fogja találni mint állítmányi pozícióban levő DAV kategóriájú igét. Ezt a típusú hibát az LCM elemzőnk több részletben történő futtatásával, illetve komplex kizárási szabályokkal el lehet hárítani. A téves találat elhárításával kapcsolatosan elméleti kérdések is felmerülnek, hiszen bizonyos szerzők [pl. 4,5] az elemzéseikben a „linking verb”-eket is figyelembe veszik mint találatot.

[4] Az xcomp dependency kategória az “open clausal complement” mondatszerkezetet jelöli. Magyarul ehhez a legközelebb azok az esetek állnak, amikor az állítmányt egy főnévi igenévvel rendelkező bővítmény követi.

3 A nyelvi kategória modell elemző reliabilitása

3.1 Szövegminta

Az általunk létrehozott nyelvi kategória modell elemző rendszer reliabalitásának méréséhez futballszurkolók internetes fórum bejegyzéseit használtuk fel. A választás mellett három érv is felhozható. Egyrészt azért döntöttünk sporttal kapcsolatos szövegminta alkalmazása mellett, mert a versengés könnyen kiválthatja a csoportközi elfogultság erőteljes megjelenését és annak nyelvi manifesztációját is. Másrészt az is fontos szempont volt, hogy természetes szöveget (spontán nyelvi megnyilvánulásokat) szerettünk volna elemezni, valódi kihívás elé állítva az elemző rendszerünket. Harmadrészt a fórumokra rendszerint több személy ír véleményt, ami heterogenitást biztosít az elemzett nyelvi mintának.

A Manchester City angol labdarúgó csapat internetes fórumáról [6] a 2013. szeptember és október hónapok legjelentősebb meccseiről szóló kommentárokat válogattuk be az elemzésbe. Ezek között győzelmek és vereségek egyaránt megtalálhatók. Két változó mentén csoportosítottuk a szövegminta mondatait: a saját vagy a külső csoportról (az ellenfél meccsről meccsre változik) mond véleményt, illetve pozitív vagy negatív véleményt fejez ki a kommentet író személy. A fentiek figyelembevételével négy alminta jött létre. Az elemző rendszerünket ezeken futtattuk le. Valamint a kézi kódolást is elvégeztük, melyet „gold standard”-nek tekintettük.

3.2 Eredmények

Az automatikus elemzés megbízhatóságát két módon mértük. A megbízhatóság egyik indikátora az, hogy az elemző rendszer által elvégzett és a kézi kódolás mennyire vezet hasonló kimenetekhez. Az 1. táblázat ad erre vonatkozó információkat. A magas találati és pontossági értékek azt mutatják, hogy az elemző eszközünk megbízhatóan azonosítja a nyelvi kategória modell kategóriáit.

1. Táblázat: A nyelvi kategória modell megbízhatósága: összesített értékek

Nyelvi kategória modell kategóriái	Kézi kódolás eredménye	Találat %	Pontosság %
Leíró cselekvő ige (DAV)	25	80,0	80,0
Értelmező cselekvő ige (IAV)	31	67,7	84,0
Állapotot kifejező cselekvő ige (SAV)	1	0	0
Állapotjelző ige (SV)	9	100	100
Melléknév (ADJ)	85	84,7	90,0
Összes kategória	161	81,9	80,6

A megbízhatóság másik indikátora a 4 alminta összesített absztrakciós mutatójának kiszámítása volt. A szöveg absztrakciós mutatója egy hányados segítségével adható meg [1]. A számláló kiszámításához minden LCM kategóriához egy súlyértéket rendelünk. Ez az érték a leíró cselekvő igék esetében 1, az értelmező cselekvő és az állapotot kifejező cselekvő igék esetében 2, az állapotjelző igék esetében 3, és végül a melléknevek esetében 4. A hányados számlálóját a súlyok és az egyes LCM kategóriák előfordulásának összegzett szorzatai adják. A hányados nevezőjében pedig az LCM kategóriák előfordulásának összege szerepel. A 2. táblázat tartalmazza a 4 almintában a kézi és az automatikus elemzés eredménye alapján kiszámolt absztrakciós mutatókat. Bár a mutató értékei azt jelzik, hogy az elemzett nyelvi mintában nem jelentkezik a csoportközi nyelvi elfogultság [4,5], azonban a gépi kódolás adataiból számított mutatók különbségének iránya azonos a kézi elemzés eredményéből számolt mutatókkal. Vagyis a saját csoport negatív cselekvése esetében számolt mutató értéke magasabb, mint a saját csoport pozitív cselekvése esetében számolt mutató mind a kézi, mind a gépi elemzés esetén. Hasonlóképpen a külső csoport pozitív cselekvése esetében számolt mutató értéke magasabb a külső csoport negatív cselekvésnél számolt értéknél a kézi és a gépi elemzés esetében is. A kézi és gépi elemzés alapján számított absztrakciós mutatók értéke nagyon közel van egymáshoz.

2. Táblázat: A nyelvi kategória modell megbízhatósága: absztrakciós mutatók

Szöveg almintái	**Absztrakciós mutató**	
	Kézi kódolás	**Gépi kódolás**
Saját csoport pozitív értékelése	3,095	3,064
Saját csoport negatív értékelése	3,184	3,197
Külső csoport pozitív értékelése	2,800	2,666
Külső csoport negatív értékelése	2,593	2,450

A megbízhatóság elemzésének eredményei azt mutatják, hogy az általunk létrehozott elemző eljárás megbízhatóan működik. Hangsúlyozzuk azonban, hogy ezeket a méréseket viszonylag kis terjedelmű szövegen végeztük el. A megbízhatóság megállapításához nagyobb terjedelmű szövegek elemzését is szükségesnek tartjuk.

Hivatkozások

1. Coenen, L. H. M., Hedebouw, L., Semin, G. R.: The Linguistic Category Model (LCM) Manual. (2006) Letöltve: http://www.cratylus.org/Text/1111548454250-3815/uploadedFiles/1151434300359-0007.pdf. Letöltés időpontja: 2013. 11. 03.
2. Finkel, J. R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), (2005) 363-370.
3. Kilgarriff, A.: BNC database and word frequency lists. http://www.kilgarriff.co.uk/bnc-readme.html. (2013).

4. Maass, A., Ceccarelli, R., Rudin, S.: Linguistic Intergroup Bias: Evidence for in-group-protective motivation. Journal of Personality and Social Psychology, (1996) Vol. 71(3), 512-526.
5. Maass, A., Salvi, D., Arcuri, L., Semin, G. R.: Language use in intergroup contexts: The linguistic intergroup bias. Journal of Personality and Social Psychology, Vol. 57(6). (1989) 981–993.
6. Manchester City futball csapat szurkolóinak fóruma: http://forums.bluemoon-mcfc.co.uk/
7. Schmid, J., Fiedler, K., Semin, G., Englich, B.: Measuring Implicit Causality: The Linguistic Category Model. (é.n.)
8. Semin, G. R., Fiedler, K.: The cognitive functions of linguistic categories in describing persons social cognition and language. Journal of Personality and Social Psychology, Vol. 54 (4) (1988) 558-568.
9. Semin, G.R.: Agenda 2000 – Communication: Language as an implementational device for cognition, European Journal of Social Psychology, Vol. 30(5), (2000) 595-612.
10. Silberztein, M.: Nooj Manual.: Letöltve: http://www.nooj4nlp.net/NooJManual.pdf (2003) Letöltés időpontja: 2013. 12. 02.
11. Tajfel, H.: Interindividual behaviour and intergroup behaviour. In H. Tajfel (ed), Differentiation between Social Groups. Studies in the social psychology of intergroup relations. Academic Press, London. (1978) 27-60.
12. The Longman Dictionary of Contemporary English Online http://www.ldoceonline.com/
13. Toutanova, K., Manning, C. D.: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), (2000) 63-70.

Narratív kategoriális tartalomelemzés: a NARRCAT[1]

Ehmann Bea[1], Csertő István[1], Ferenczhalmy Réka[2], Fülöp Éva[1], Hargitai Rita[2], Kővágó Pál[1,2], Pólya Tibor[1], Szalai Katalin[2], Vincze Orsolya[2], László János[1,2]

[1] MTA TTK Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
ehmann.bea@ttk.mta.hu, csertopi@gmail.com,
fülöp@mtapi.com, kovago.pal@mta.ttk.hu,
polya.tibor@ttk.mta.hu

[2] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
ferenczhalmy@gmail.com, hargitairita@freemail.com,
orsolyavincze@hotmail.com, szalaikati.sza@gmail.com
laszlo.janos@ttk.mta.hu

Kivonat: A Számítógépes Nyelvészeti Konferenciákon a korábbi évek során már számos alkalommal beszámoltunk a Narratív Pszichológiai Kutatócsoport tevékenységéről és korpusznyelvészekkel való együttműködéséről. Jelen dolgozatban a NarrCat narratív kategoriális elemzőt a tudományos narratív pszichológia elméletén alapuló egységes egészként kívánjuk bemutatni a nyelvész és a pszichológus közösségnek. Bemutatjuk a NarrCat Pszichotematikus moduljait, Hipermoduljait és Relációs moduljait, és ezek egymással való kapcsolatát. Ismertetjük a csoport által végzett új fejlesztéseket, s végül felvázoljuk az interdiszciplináris együttműködés jelenlegi helyzetét.

1 Tudományos narratív pszichológia

A narratív pszichológiai paradigma, mely szerint a való világban létező egyének és csoportok saját identitásukat és pszichológiailag érvényes valóságukat történetek révén alkotják meg, a múlt század nyolcvanas éveiben kezdett kibontakozni [2, 34].

A tudományos narratív pszichológia irányzatát László János indította el [22, 23, 24, 25], s az elmélet az ő vezetése alatt működő munkacsoport közreműködésével nyerte el mai formáját [11, 19, 27, 32].

A tudományos narratív pszichológia lényege az alábbiakban foglalható össze: az egyének és a csoportok a történeteiket különböző kompozíciós elvek révén hozzák létre, s e kompozíciós elvek tükrözik az egyének és a csoportok belső, pszichológiai kategóriákkal leírható állapotait. Az elmélet módszere a narratív kategoriális tartalomelemzés, mely a kompozíciós elveket és kategóriákat pszichológiai kategóriákkal párosítja és statisztikailag elemzi. Az elmélet és a módszer hozama, hogy az empirikus eredmények értelmezése révén diagnosztizálja és bejósolja az egyének és csoportok pszichológiai, valamint identitás állapotait és folyamatait.

[1] A kutatást az OTKA K 109009 számú pályázata támogatta.

2 A narratív kategoriális elemzés elméleti koncepciója

Az elmúlt húsz év során a Narratív Munkacsoport önálló alkalmazást fejlesztett ki empirikus narratív pszichológiai vizsgálatok végzésére, melyet több éve már NarrCatnak nevezünk, és NooJ platformon működtetünk [35].

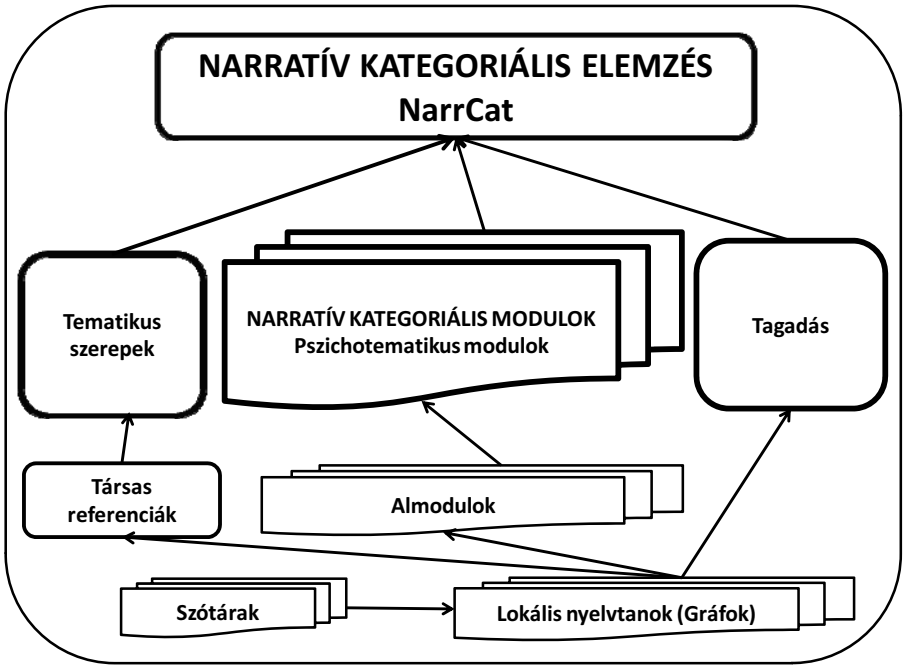
A narratív kategoriális elemzés gondolata csaknem két évtizede merült fel. 1994-ben Intézetünk felkérést kapott arra, hogy szenvedélybetegek angol nyelvű naplóinak tartalomelemzésével térképezze fel, milyen fajta nyelvhasználat jelezheti előre a gyógyulást. Egyik eredményünket a LIWC szoftver [28] segítségével kaptuk: a gyógyultnak tekintett szenvedélybetegek naplóikban szignifikánsan többször használták a 'gátlás' és a 'belátás' szavait. Ezen túllépve azonban – akkor még kézi vezérléssel – minden egyes mondatot bekategorizáltunk aszerint, hogy az magára a naplóíróra, a terápiára vagy a külvilágra vonatkozott-e; illetve, hogy a mondat tartalma negatív, semleges vagy pozitív volt-e. Eredményeink szerint a terápia akkor bizonyult sikeresnek, ha a résztvevők a terápia kezdetén mind önmagukhoz, mind a terápiához negatívan viszonyultak, s ez a viszonyulás a terápia előrehaladtával fokozatosan pozitívvá vált [26].

A NarrCat által végzett narratív kategoriális elemzés során gépi úton végzett, specifikus tematikájú adatredukció történik: az egyéni és csoportnarratívumok meghatározott egységeit (mondatokat/mondatrészeket) a későbbi narratív pszichológiai elemzés alapjául szolgáló kategóriákká transzformáljuk. Például: 'Büszke vagyok apámra' = *a Szelf mint Ágens absztrakt pozitív érzelme a Másik mint Recipiens iránt a jelenben*; illetve 'A törökök megtámadták a magyarokat' = *Külső csoport mint Ágens negatív aktivitása a Saját csoport mint Recipiens iránt a múltban*. Az alábbiakban felvázoljuk, miként történik ez a transzformáció.

3 A NarrCat szerkezeti felépítése

A NarrCat két nyelvtechnológiai előfeltétel révén jöhetett létre: az egyik a jelenleg platformul szolgáló NooJ korpusznyelvészeti fejlesztő környezet [35], a másik pedig a magyar nyelvű nemzeti korpuszok [3, 39]. A NarrCat kifejlesztésének intézményes feltételei az elmúlt évtized során több pillérre – a pszichológusok részéről az MTA TTK Kognitív Idegtudományi és Pszichológiai Intézete és a Pécsi Tudományegyetem, a nyelvészek részéről az MTA Nyelvtudományi Intézete, a Szegedi Tudományegyetem és a Morphologic Kft. közötti együttműködésre, s végül, de nem utolsósorban, a Magyar Számítógépes Nyelvészeti konferenciákon szerzett tapasztalatokra épültek.

A NarrCat szerkezeti felépítésének átfogó képét az 1. ábra mutatja.



1. ábra: A NarrCat szerkezete

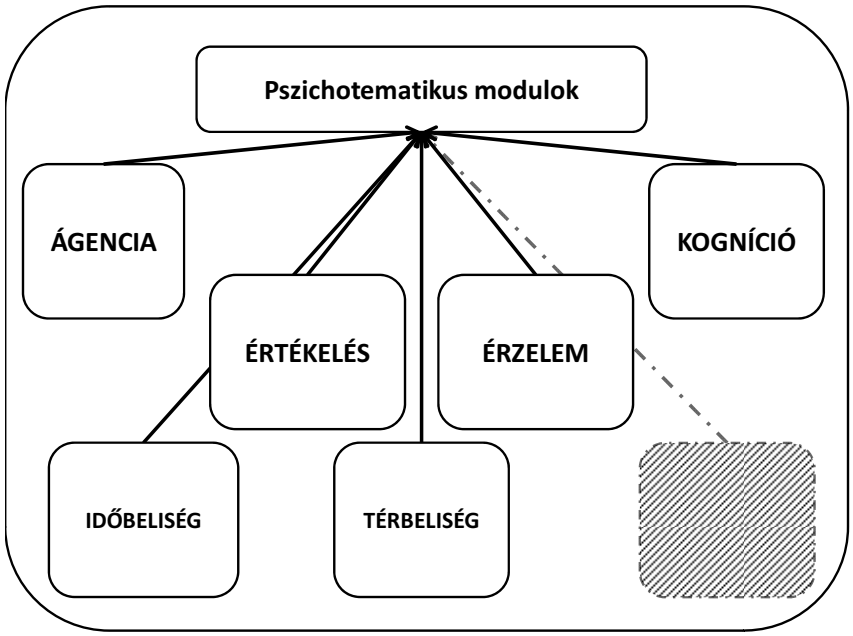
A moduláris felépítésű rendszer alapját szótárak képezik; ezek szókincsét egyfelől a magyar írott nyelv általános szókincsét reprezentáló szövegkorpuszokból (Magyar Nemzeti Szövegtár [39], Szeged Korpusz [3]), másfelől specifikus pszichológiai szövegekből álló korpuszból nyertük ki. Ez utóbbiban megtalálhatóak klinikai pszichológiai populációkkal (depressziós, borderline, droghasználó, krízisben lévő betegekkel) készített mélyinterjúk, többgenerációs traumatizált családinterjúk, normál populációkkal (teljesítmény-, veszteség-, párkapcsolati témában) felvett félig strukturált interjúk, valamint nemzeti, történelmi és etnikai vonatkozású szövegkorpuszok.

A lingvisztikailag annotált szótárak inputként szolgálnak a lokális nyelvtanokhoz. A lokális nyelvtanok két magasabb rendű modulrendszerhez adnak bemenetet: a Pszichotematikus modulokhoz és a Relációs modulokhoz – ez utóbbiak a Társas referenciák, a Tagadás és a Tematikus szerepek modulok. A szótárak, a lokális nyelvtanok és a modulok rugalmas kombinálhatósága ún. Hipermodulok létrehozását is lehetővé teszi – ilyen például a Pszichológiai perspektíva és a Téri-idői modul.

A NarrCat nyitott és bővíthető rendszer; általában és eredeti szándéka szerint jelen formájában is alkalmas a legkülönfélébb egyéni és csoportnarratívumok elemzésére, ám kisebb, projektspecifikus alkalmazásokat is megenged [21].

4 A NarrCat pszichotematikus moduljai

A rendszer magvát a Narratív kategoriális, avagy pszichotematikus modulok képezik. Ezek közül a négy legrobusztusabb az Érzelem, az Értékelés, az Ágencia és a Kogníció; kevésbé robusztusak, ám még mindig meglehetősen összetettek például az Időbeliség és a Térbeliség modulok. Minden modul rendelkezik almodulokkal, melyek bemeneteit szótárakon alapuló lokális nyelvtanok képezik. A pszichotematikus modulok áttekintését a 2. ábrán láthatjuk. A satírozottan jelölt üres modul a rendszer rugalmas bővíthetőségét kívánja jelképezni.

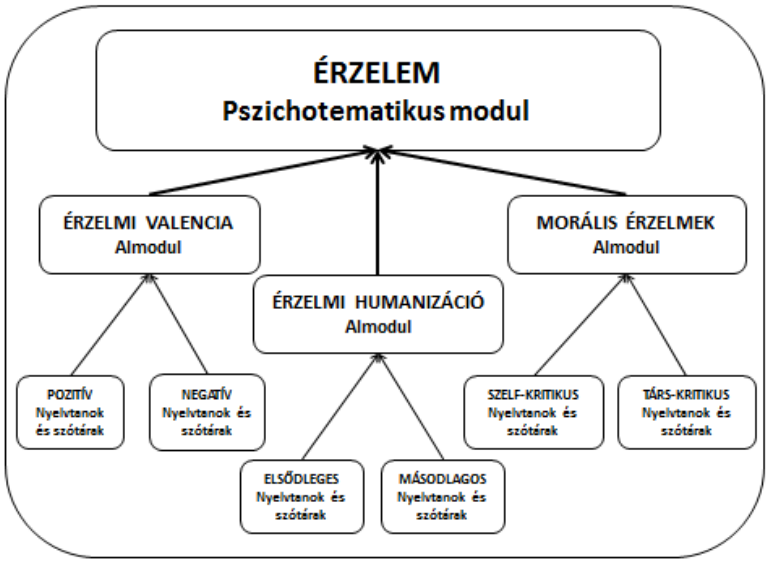


2. ábra: Pszichotematikus modulok

A következőkben részletesebben is bemutatjuk az egyes pszichotematikus modulokat.

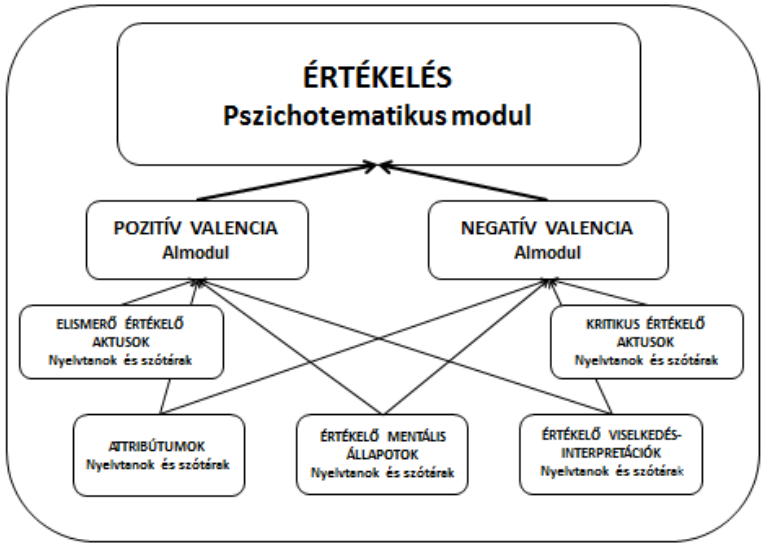
4.1 A NarrCat Érzelem modulja

A NarrCat Érzelem moduljának szerkezete a 3. ábrán látható [16]. Főbb összetevői az Érzelmi valencia (negatív és pozitív érzelmek), az Érzelmi humanitás (elsődleges és másodlagos érzelmek) és a Morális (szelf-kritikus és társ-kritikus) érzelmek által alkotott alrendszerek. Az érzelem modult a nemzeti identitás vizsgálatok keretében a történelmi pályaérzelmek feltárására, valamint a nemzeti identitás érzelmi szerveződésének és a kollektív traumák feldolgozottságának vizsgálatára alkalmaztuk [15].



3. ábra: A NarrCat Érzelem modulja

4.2 A NarrCat Értékelés modulja



4. ábra: A NarrCat Értékelés modulja

Az Értékelés modul a Pozitív és Negatív értékelés almodulokból tevődik össze, melyek a 4. ábrán látható szótárakból és lokális nyelvtanokból kapnak bemenetet [4]. A modult a nemzeti identitás vizsgálatok keretében csoportközi elfogultság vizsgálatokban alkalmaztuk [5].

4.3 A NarrCat Ágencia modulja

Az Ágencia modul almoduljai az Aktivitás [38] és az Intencionalitás [12]. Az előbbi az Aktivitás és a Passzivitás, az utóbbi az Intencionalitás és a Megszorítás nyelvtanokból és szótárakból kap bemenetet. A szereplők aktivitásának-passzivitásának kvantitatív mutatói segítségével feltérképezhető, hogy a vizsgált személy vagy csoport mennyire van hatással a környezetére; az Intencionalitás-Megszorítás almodul pedig az ágens gondolkodásának célirányosságát és hatékonyságát méri. Az Ágencia modult történelmi szövegek elemzésében is alkalmaztuk, a saját csoport és a külső csoport viszonyainak feltárására [13, 20, 37].

4.4 A NarrCat Kogníció modulja

A Kogníció pszichotematikus modul összetevői a mentális igék és főnevek, valamint a mentális idiómák [41]. A Kogníció modult a mentalizáció és a kognitív empátia kutatásában, valamint az egyéni és történelmi traumák feldolgozásának vizsgálatában alkalmaztuk [40, 42].

4.5 A NarrCat Térbeliség modulja

A Térbeliség modul részben térbeli deiktikus szavakat és névmásokat tartalmaz [30], valamint a Téri interperszonális kapcsolati mozgás almodult foglalja magában, mely utóbbi a Társas közelítés és a Társas távolodás szótárakból és nyelvtanokból kap bemenetet. A társas közeledés és távolodás a borderline páciensek narratívumainak vizsgálatában játszik szerepet [29].

4.6 A NarrCat Időbeliség modulja

Az Időbeliség pszichotematikus modult a Tartalmi és a Funkcionális almodulok alkotják [8]. A modult a szubjektív időélmény és a személyiségvonások összefüggésének vizsgálatára [8], valamint traumatizált személyek narratívumait jellemző szubjektív időélmény feltárására alkalmaztuk [9].

5 A NarrCat Hipermoduljai

A pszichotematikus modulok, illetve ezek almoduljai és lokális nyelvtanai különböző kombinációkban, rugalmasan és nagy variabilitással hipermodulokká is

összekapcsolhatók. A NarrCatnak ez a specifikus tulajdonsága az egyéni és a csoportidentitás összetett pszichológiai folyamatainak részletesebb feltárását teszi lehetővé. A rendszer jelenleg két Hipermodult tartalmaz.

5.1 A Pszichológiai perspektíva hipermodul

A Pszichológiai perspektíva Hipermodul az Érzelem, a Kogníció modulokból, valamint az Ágencia modul Intencionalitás almoduljából tevődik össze. A perspektíva alkalmazása számos pszichológiai jelenség vizsgálatában játszik szerepet, mint például a mentalizáció, az empátia vagy az irodalmi és történelmi szövegek befogadásának perspektíva felvétele [30, 42].

5.2 A Téri-idői perspektíva hipermodul

A Tér-idői perspektíva Hipermodult jelenleg lokális nyelvtanok alkotják. A konstruktum a narratívum tartalma és az elbeszélői pozíció közötti viszonyra vonatkozik. Három formáját írták le: a retrospektív, az újra-átélő és a metanarratív formát [30]. A Hipermodul a narrátor által elfoglalt téri-idői perspektíva azonosítására szolgál. A téri-idői perspektíva hipermodult a fenyegetett társas identitással és az érzelemszabályozással kapcsolatos vizsgálatokban alkalmaztuk [31].

6 A NarrCat Relációs moduljai

A NarrCat három Relációs modult tartalmaz: a Tagadást, a Társas referenciákat és a Tematikus szerepeket. Mindhárom modul pszichotematikus szerepet is betölt.

6.1 A NarrCat Tagadás modulja

A tagadás modul a tagadószavakat, a tagadó névmásokat, tagadó határozószavakat, tagadó névutókat és fosztóképzőket méri [18]. A modul részben pszichotematikus modulnak tekinthető, hiszen önmagában is rendelkezik pszichológiai korrelátumokkal – például sztereotipizálás jele lehet [1]. A tagadást pszichodinamikai szempontból az egészséges emberi környezethez és morális mércékhez való alkalmazkodásra, illetve a világ értéktelenítésére, a destrukcióra és öndestrukcióra való hajlamra vonatkozóan vizsgáltuk [18].

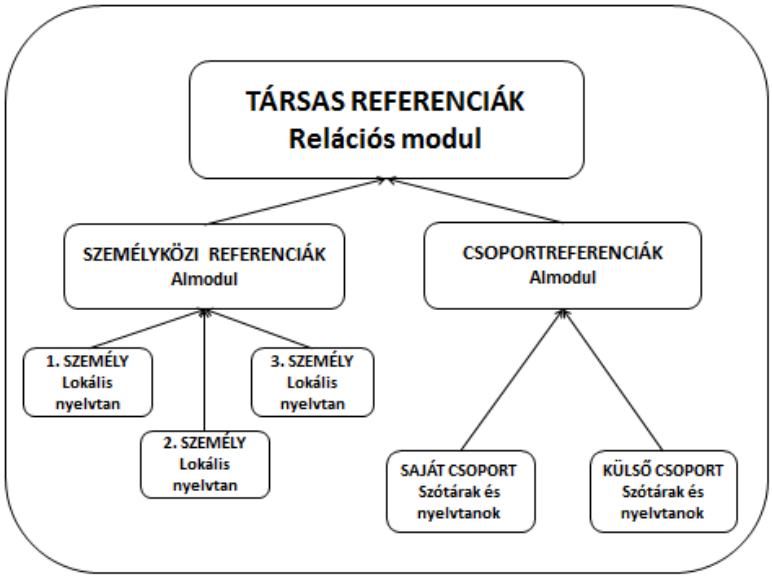
Egyúttal azonban relációs modulként is viselkedik, mivel módosíthatja egyes pszichotematikus modulok pszichológiai korrelátumait (például a túlságosan magas szelf-referencia arány önmagában nem feltétlenül jelez depressziót, magas tagadás aránnyal párosulva azonban már igen). A Tagadás modulnak ezért narratív kategorizációs mintázatelemzésben van – jelenleg még kevéssé vizsgált – relációs szerepe.

6.2 A NarrCat Társas referenciák modulja

A narratív szociálpszichológia két fő kategorizációs dimenziója az Én és a Másik, illetve a Saját csoport és a Külső csoport. A Társas referenciák modulja ennek megfelelően két almodulból áll.

A Személyközi referenciák almodul az én, te, ő, mi, ti ők referenciákat azonosítja [18]. A szelf-referencia pszichológiai tartalmát depressziókutatásban alkalmaztuk [18]; az én és a mi referenciák egymáshoz viszonyított arányát pedig izolált kiscsoport vizsgálatában arra használtuk, hogy nyomonkövessük a csapatszellem alakulását egy űranalóg szimulációs expedíció során [7].

A Csoportközi referenciák almodul összetettebb: minthogy az elbeszélésben a saját csoport és a külső csoport egyrészt annak függvénye, hogy milyen szereplők vesznek részt az eseményben, másrészt annak, hogy az elbeszélő melyik csoport tagja – ez az almodul projekt-specifikus szótárakon és lokális nyelvtanokon alapul.



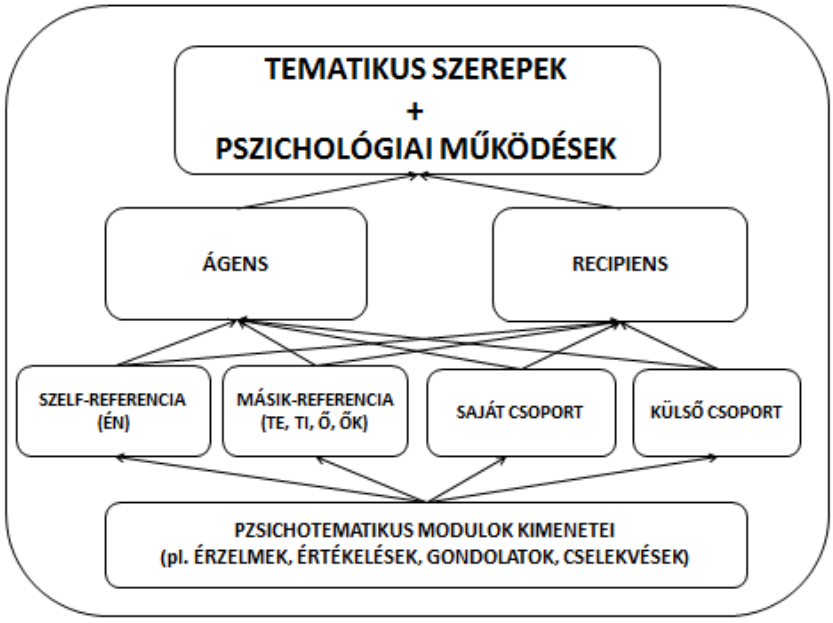
5. ábra: A NarrCat Társas referenciák Relációs modulja

A NarrCat specifikus újdonsága, hogy a pszichotematikus modulokkal kapott eredményeket képes a modul által azonosított találatokkal logikai kapcsolatban álló személyekkel és csoportokkal is összekötni, azaz kijelzi, hogy ki érez, ki értékel, ki gondolkozik, ki aktív, stb. Ekképp a Társas referenciák modulja a Pszichotematikus modulok felett a NarrCat második logikai szintje. Erre épül rá a NarrCat harmadik logikai szintje: a Tematikus szerepek modulja.

6.3 A NarrCat Tematikus szerepek modulja

A tematikus szerepek (Semantic Role Labeling, SRL) funkció célja annak azonosítása, hogy az Én és a Másik, illetve a Saját csoport és a Külső csoport vajon Ágense vagy Recipiense-e a cselekvésnek, az érzelemnek, a kogníciónak vagy az értékelésnek [10].

A tematikus vagy szemantikus szerepek nyelvészeti irodalma hosszú évszázadokra nyúlik vissza [17], napjainkban pedig a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe [26]. Az automatikusan feldolgozni kívánt tematikus szerepek száma a különböző szerzők felfogásában más és más – a leggyakoribbak az Ágens, a Páciens, a Téma, az Experiencer, a Kedvezményezett, az Eszköz, A Hely, a Forrás, a Cél és a Mód. Foley és van Valin az Actor és Undergoer, Dowty pedig a Proto-ágens és a Proto-páciens szerepeket javasolja [14, 6].



6. ábra: A tematikus szerepek elemzésének logikai szerkezete

A narratív pszichológusok számára is tökéletesen elegendő a kettős felosztás. A narratív kategoriális tartalomelemzés tehát a NarrCat három logikai szintjén a következő dolgokat kívánja megtudni az egyének és csoportok énelbeszéléseiből: vajon az Én-e avagy a Másik, a Saját csoport-e avagy a Külső csoport az az Ágens, aki érez, értékel, gondolkozik és cselekszik? Továbbá, vajon az Én-e avagy a Másik, a Saját csoport-e avagy a Külső csoport az a Recipiens, akire az Ágens érzései, értékelései, gondolatai és cselekedetei vonatkoznak vagy irányulnak? A NarrCat szintjeinek logikai szerkezetét a 6. ábra mutatja.

A tematikus szerepek finomítása és a pszichotematikus modulokkal egy platformon való összekötése jelenleg is folyamatban van.

A narratív kategoriális tartalomelemzés végül azzal válik teljessé, hogy az egyének és a csoportok énelbeszéléseiben szereplő, immár narratív kategóriákká transzformált megnyilatkozásokból statisztikai elemzés révén következtetéseket vonunk le az egyének és a csoportok identitásállapotairól, identitásfolyamatairól és identitásmintázataik változásairól.

7 NarrCat – Interdiszciplináris együttműködés és kitekintés

A NarrCat létrejöttében szinte a kezdettől fogva együttműködtek nyelvészek és pszichológusok. A Pszichotematikus modulok kimunkálásában az MTA Nyelvtudományi Intézete, a Szegedi Tudományegyetem és a Morphologic Kft. vett részt.

Jelenleg a NarrCat második és a harmadik logikai szintje képezi nyelvészek, informatikusok és pszichológusok közötti interdiszciplináris együttműködés tárgyát. A tematikus szerepek vizsgálatában és a szintaktikai elemzésekben külső eszköze korábban a MetaMorpho volt, jelenleg a magyarlancot használjuk [10, 33, 43, 44].

Hivatkozások

1. Beukeboom, C. J., Finkenauer, C., Wigboldus, D. H. J.: The negation bias: When Negations Signal Stereotypic Expectancies. Journal of Personality & Social Psychology, 99(6) (2010) 978–992
2. Bruner, J.: Actual Minds, Possible Worlds. Cambridge, Harvard University Press (1986)
3. Csendes, D., Alexin, Z., Csirik, J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005) 409–412
4. Csertő I., László, J.: A csoportközi értékelés mint a csoporttrauma érzelmi feldolgozásának indikátora a nemzeti történelem elbeszéléseiben. In: Tanács A. és Vincze V. (szerk.), VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2011) 212–222
5. Csertő, I., László, J.: Intergroup Evaluation as an Indicator of Emotional Elaboration of Collective Traumas in National Historical Narratives. Sociology Study, 3(3), (2013) 207–224
6. Dowty, D.: Thematic proto-roles and argument selection. Language 67. (1991) 547–619
7. Ehmann, B., Balazs, L., Fulop, E., Hargitai, R., Kabai, P., Peley, B., Polya, T., Vargha, A., Vincze, O. and Laszlo, J. (2011): Narrative Psychological Content Analysis as a Tool for Psychological Status Monitoring of Crews in Isolated, Confined and Extreme Settings. Acta Astronautica, 68 (9–10) (2011) 1560–1566
8. Ehmann, B., Garami, V., Naszódi, M., Kis, B. and László, J.: Subjective Time Experience: Identifying Psychological Correlates by Narrative Psychological Content Analysis. Empirical Culture and Text Research 3, (2007) 14–25
9. Ehmann, B., Garami, V.: Narrative Psychological Content Analysis with NooJ: Linguistic Markers of Time Experience in Self-Reports. In: Váradi, T., Kuti, J., Silberztein, M.: Applications of Finite-State Language Processing -- Selected Papers from the 2008 International NooJ Conference Cambridge Scolars Publishing, Newcastle upon Tyne, UK (2010) 186–196
10. Ehmann, B., Lendvai, P., Pólya, T., Vincze, O., Miháltz, M., Tihanyi, L., Váradi, T., László, J.: Narrative Psychological Application of Semantic Role Labeling. In: Vučković,

K., Bekavac, B., Silberztein, M. (eds.): Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference, Cambridge Scholars Publishing, Newcastle upon Tyne, UK (2011) 218–228
11. Ehmann, B.: A szöveg mélyén. A pszichológiai tartalomelemzés. Új Mandátum, Budapest (2002)
12. Ferenczhalmy, R., László J.: Az intencionalitás modul kidolgozása Nooj tartalomelemző programmal. In: Alexin, Z., Csendes, D.: (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged (2006) 285–295
13. Ferenczhalmy, R., László, J.: In-group versus out-group intentionality as indicators of national identity. Empirical Text and Culture Research 4 (2010) 59–69
14. Foley, W. A. and Van Valin, R. D.: Functional syntax and universal grammar. Cambridge University Press, Cambridge (1984)
15. Fülöp, É., Csertő, I., Ilg, B., Szabó, Zs., Slugoski, B. and László, L.: Emotional elaboration of collective trauma sin historical narratives. In László, J., Forgas, J., Vincze, O. (eds.), Social Cognition and Communication. Sydney Symposium of Social Psychology. New York, Psychology Press. (2013) 245–262
16. Fülöp, É., László, J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző program segítségével In: Alexin, Z., Csendes, D.: (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged (2006) 296–304
17. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics, Vol. 28, No. 3, (2002) 245-288
18. Hargitai R, Naszódi M, Kis B, Nagy L, Bóna A, László J. Linguistic markers of depressive dynamics in self-narratives: Negation and self-reference. Empirical Text and Culture Research, 3 (2007) 26–38.
19. Hargitai, R.: Sors és történet: Szondi Lipót sorsanalízise a narratív pszichológia tükrében, Új Mandátum, Budapest (2008)
20. László J., Szalai K., Ferenczhalmy R.: Role of Agency in Social Representations of History. Societal and Political Psychology International Review (1) (2010) 31–43
21. László, J., Csertő, I., Ferenczhalmy, R., Fülöp, É., Hargitai, R., Péley, B., Pohárnok, M., Pólya, T., Szalai, K., Vincze, O. & Ehmann, B.: Narrative language as expression of individual and group identity: The Narrative Categorical Content Analysis (NarrCat). Sage Open, http://sgo.sagepub.com/content/3/2/2158244013492084 (2013)
22. László, J.: A történetek tudománya. Bevezetés a narratív pszichológiába. Új Mandátum, Budapest (2005)
23. László, J.: Historical Tales and National Identity. An introduction to narrative social psychology. Routledge, London New York (2013)
24. László, J.: The Science of Stories: An introduction to Narrative Psychology. Routledge, London New York (2008)
25. László, J.: Történelemtörténetek. Bevezetés a narratív szociálpszichológiába. Akadémiai Kiadó, Budapest (2012)
26. Márquez, L., Carreras, X., Litkowsky, K. C. and Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue. Computational Linguistics, 34 (2) (2008) 145–159
27. Péley, B.: Rítus és történet. Beavatás és kábítószeres létezésmód. Új Mandátum, Budapest (2002)
28. Pennebaker, J.W., Booth, R.J., & Francis, M.E.: Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX: LIWC (www.liwc.net). (2007)
29. Pohárnok M, Naszódi M, Kis B, Nagy L, Bóna A, László J.: Exploring the spatial organization of interpersonal relations by means of computational linguistic analysis, Empirical Text and Culture Research 3: (2007) 39–49
30. Pólya, T., Kis, B., Naszódi, M., László, J.: Narrative perspective and the emotion regulation of a narrating person. Empirical Text and Culture Research, 7(3), (2007) 50–61

31. Polya, T., Laszlo, J., Forgas, J. P.: Making sense of life stories: the role of narrative perspective in perceiving hidden information about social identity. European Journal of Social Psychology 35(6) (2005) 785–796.
32. Pólya, T.: Identitás az elbeszélésben. Szociális identitás és narratív perspektíva. Új Mandátum, Budapest (2007)
33. Prószéky, G. and Tihanyi, L.: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference, 19–24. ASLIB, London, United Kingdom (2002)
34. Sarbin, T. R. (Ed.): Narrative Psychology. The Storied Nature of Human Conduct. Praeger, New York (1986)
35. Silberztein, M.: NooJ Manual. http://www.nooj4nlp.net/NooJManual.pdf (2003)
36. Stephenson, G.M., László, J., Ehmann, B., Lefever, R.M.H., Lefever, R.: Diaries of Significant Events: Socio-Linguistic Correlates of Therapeutic Outcomes in Patients with Addiction Problems. Journal of Community and Applied Social Psychology, 7, (1997) 389–411
37. Szalai, K., László, J.: Activity as a linguistic marker of agency: Measuring in-group versus out- group activity in Hungarian historical narratives. Empirical Text and Culture Research 4: (2010) 50–58
38. Szalai, K., László, J.: Az aktivitás-passzivitás modul kidolgozása NooJ tartalomelemző programmal. In: Alexin, Z., Csendes, D.: (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged (2006) 330–338
39. 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
40. Vincze O., Tóth J., László J.: Perspektíva-felvétel, csoportidentitás és sztereotípia A mentális igék szerepe a szövegben. In: Vincze O., és Bigazzi S. (szerk.): Élmény, történet – a történetek élménye. Új Mandátum, Budapest (2008) 52–60
41. Vincze O.: Mentális kifejezések jelentősége a perspektíva-felvételben a csoportidentitás tükrében. In: Tanács, a, Csendes,D. (szerk):V. Magyar Számítógépes Nyelvészeti Konferencia. (2007) 250–258
42. Vincze, O., Ilg, B., & Pólya, T.: The role of narrative perspective in the elaboration of individual and historical traumas. In László, J., Forgas, J., Vincze, O. (Eds.), Social Cognition and Communication. Sydney Symposium of Social Psychology. New York, Psychology Press. (2013) 229–244
43. Vincze, O., Kata G., Ehmann, B., László, J.: Technológiai fejlesztések a NooJ pszichológiai alkalmazásában. In: Tanács, A., Szauter, D., Vincze, V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2009) 285–294
44. Zsibrita, J., Vincze, V., Farkas,R.: Magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács A. és Vincze V. (szerk.), IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 368–374

Történetszerkezet mint az érzelmi intelligencia indikátora

Pólya Tibor

Magyar Tudományos Akadémia, Természettudományi Kutatóközpont,
Kognitív Idegtudományi és Pszichológiai Intézet
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.mta.hu

1 Az érzelmi intelligencia fogalma

1.1 Az érzelmi intelligencia meghatározása

Az érzelmi intelligencia fogalmának meghatározásában két egymással élesen vitázó elképzeléssel találkozhatunk. Az egyik elképzelés szerint az érzelmi intelligencia képességként [6], a másik elképzelés szerint személyiségjellemzőként [2] határozható meg. A képességként meghatározott érzelmi intelligencia 4 területet foglal magában. A területeket Oláh Attila [8] összefoglalása alapján mutatjuk be az 1. táblázatban.

1. táblázat: A képességként meghatározott érzelmi intelligencia meghatározása

Érzelmi percepció	Az érzelem észlelése, értékelése és kifejezése. Az érzelem azonosításának képessége önmagunknál (különböző testi és lelki állapotainkban). Az érzelem azonosításának képessége más személyeknél és különböző helyzetekben. Az érzelmek pontos kifejezésének képessége, és az érzésekhez kötődő igények kifejezése. Azon képesség, hogy különbséget tudunk tenni a pontos és pontatlan vagy az őszinte és nem őszinte érzelemkifejezési módok között.
Érzelmi integráció	A gondolkodás érzelmi serkentése. Azon képesség, hogy valaki az érzései alapján újrarendezze, fontossági sorrendbe állítsa gondolatait tárgyakkal, eseményekkel és más emberekkel kapcsolatosan. Azon képesség, hogy létrehozzunk olyan élénk érzelmeket, amelyek facilitálják az ítéletalkotást és az érzésekre vonatkozó emlékezést. Azon képesség, hogy tőkét kovácsoljunk a hangu-

	latváltozásainkból, amelyek több nézőpont felvételét segítik, és hogy integrálják a hangulat kiváltotta nézőpontokat. Azon képesség, hogy az érzelmi állapotokat arra használjuk, hogy azok serkentsék a problémamegoldást és a kreativitást.
Érzelmi megértés	Az érzelmi információ megértése és elemzése, az érzelmi tudás alkalmazása. Az a képesség, hogy megértsük, hogy a különböző érzelmek hogyan viszonyulnak egymáshoz. Az a képesség, hogy észleljük az érzések okait és következményeit. Az a képesség, hogy értelmezni tudjuk a komplex érzéseket, a kevert érzéseket és az ellentmondó érzéseket. Az a képesség, hogy megértsük és megjósoljuk az érzelmek közötti valószínű átmeneteket.
Az érzelem szabályozása	Az a képesség, hogy nyitottak legyünk az érzésekre, kellemesekre és kellemetlenekre egyaránt. Az érzelmek monitorozásának és mérlegelésének képessége. Az a képesség, hogy létrehozzunk, fenntartsunk egy érzelmi állapotot, vagy éppen eltávolodjunk tőle, attól függően, hogy az állapotot mennyire ítéljük informatívnak vagy hasznosnak. Az a képesség, hogy kezeljük mások érzéseit, illetve a saját érzéseinket.

Az érzelmi intelligenciát személyiségjellemzőként meghatározó elképzelések is több tényezőt sorolnak fel. Petrides és munkatársai [9] által kidolgozott modell például 15 területet sorol fel (lásd 2. táblázat).

1.2 Az érzelmi intelligencia mérése

A két álláspont jelentősen különbözik abban is, hogy milyen elképzeléseik vannak az érzelmi intelligencia mérésére vonatkozóan. A képesség megközelítés szerint olyan tevékenységben kell mérni az érzelmi intelligenciát, amelyben az meg tud nyilvánulni, hasonlóan például ahhoz, ahogyan a személy intelligenciája megnyilvánul a problémamegoldás során. Az érzelmi intelligencia esetében azonban kérdéses, hogy milyen tevékenységben jelenhet meg ez a képesség. Az érzelmi intelligenciát személyiségjellemzőként meghatározó elképzelés a mérést illetően sokkal könnyebb helyzet-

ben van, mivel a pszichológiai kutatásban elterjedten használt kérdőíves eljárások segítségével mérhetőnek gondolja az érzelmi intelligenciát.

2. táblázat: A személyiségjellemzőként meghatározott érzelmi intelligencia meghatározása

Személyiségjellemző	Tulajdonság
Alkalmazkodó képesség	Rugalmas és kész arra, hogy új helyzetekhez alkalmazkodjon
Asszertivitás	Őszinte és kész arra, hogy kiálljon a jogaiért
Érzelem kifejezés	Érzéseit másoknak kommunikálja
Mások érzelmeinek kezelése	Más személyek érzéseit befolyásolni tudja
Érzelem észlelése	Tisztán látja saját és mások érzéseit
Érzelemszabályozás	Kontrollálni tudja saját érzelmeit
Impulzivitás	Önreflektív és képes ellenállni késztetéseinek
Kapcsolatok	Számára kielégítő személyes kapcsolatokkal rendelkezik
Önértékelés	Sikeres és magabiztos
Önmotiváció	Motivált és céljait a nehézségek ellenére sem adja fel
Társas tudatosság	Jártasság a társas kapcsolatok kialakításában
Stressz kezelés	Képes ellenszegülni a nyomásnak, illetve szabályozni a stresszt
Vonás empátia	Képes más személyek perspektívájának felvételére
Vonás boldogság	Jókedvű és elégedett az életével
Vonás optimizmus	A dolgok jó oldalát látja

1.3 Az érzelmi intelligencia és a történet konstrukció kapcsolata

Javaslatom szerint a történet konstrukciója olyan tevékenység, amelyben közvetlenül megjelenik az elbeszélő személy érzelmi intelligenciája. Az érzelmi intelligencia és a történet konstrukciója közötti kapcsolat két szinten ragadható meg.

Egyrészt az elbeszélt események érzelmeket váltottak ki a történet szereplőiből a múltban. A történet rendszerint olyan eseménysort ír le, amelynek során az események eltérnek az adott helyzetre vonatkozó elvárásoktól [15]. Az elvárások meghiúsulása következtében a történet szereplői gyakran élnek át intenzív érzelmeket, így a legtöbb történet sok érzelemre vonatkozó információt tartalmaz [7]. Az érzelmek azonban nemcsak a történet tartalmának fontos részét adják, hanem emellett a történetszerkezet kialakításában is szerepet játszanak. A történetnyelvtanok [5] fogalmaiban használva a történetet egy bonyodalmat okozó esemény indítja el, amely arra készteti a történet főszereplőjét, hogy megpróbálja visszaállítani a kezdeti harmonikus állapotot. A főszereplő próbálkozásait értékelő folyamatok kísérik, amelyek arról adnak információt, hogy főszereplő közelebb került-e célja eléréséhez, vagy éppen távolodik attól. Az értékelő folyamatoknak számos formája van. Ezek egyike a főszе-

replő által átélt érzelem, amelyszintén azt jelzik, hogy a főszereplő közelebb jutott-e célja eléréséhez.

A történet konstrukciója így azt feltételezi, hogy az elbeszélő személy elegendő tudással rendelkezik a társas világ működésére vonatkozóan. Ezen belül fontos szerepe van az érzelmekre vonatkozó tudásnak. Mayer és Salovey [6] érzelmi intelligencia meghatározásából az érzelem percepciója és megértése komponensek azok, amelyek szükségesek a történetmeséléshez.

Másrészt az elbeszélés érzelmeket vált ki az elbeszélés során az elbeszélőben és a hallgatóban is. Az elbeszélés társas helyzetbe illeszkedik, amely lényegi módon meghatározza a történetszerkezetét. Ennek részeként az elbeszélőnek például világossá kell tennie azt, hogy milyen érzelmi viszonyban van a történetbe foglalt eseményekhez. A történetmesélés egyik fontos célja, hogy érzelmeket váltson ki. Az érzelem kiváltó hatást elsősorban a hallgató szempontjából vizsgálták [1], de igaz az elbeszélő esetében is [9, 12, 13].

A történet konstrukciója így azt is feltételezi, hogy az elbeszélő képes úgy formálni a történetet, hogy az igazodjon hallgatója sajátosságaihoz. Ezen belül szintén fontos szerepe van az érzelmeknek. Mayer és Salovey [6] érzelmi intelligencia meghatározásából az érzelmek integrációja és szabályozása komponensek is szükségesek a történetmeséléshez.

Mindezek alapján az a hipotézis fogalmazható meg, hogy az elbeszélő személy érzelmi intelligenciája megnyilvánulhat a történet konstrukciójában. Azt várhatjuk, hogy az érzelmileg intelligensebb személyek által elmesélt történetek szerkezete kidolgozottabb, mint az érzelmileg kevésbe intelligens személyek által elmesélt történetek szerkezete.

2 Az érzelmi intelligencia és a történet konstrukció közötti kapcsolat empirikus vizsgálata

Hipotézisünket empirikus vizsgálatban ellenőriztük. A vizsgálatban 60 személy idézett fel érzelemteli epizódokat hívószavas eljárással. A hívószavak érzelmi kategóriák címkéi voltak, mint például öröm és düh. A történetek szerkezetét automatikus nyelvi elemzési eljárásokkal elemeztük. Az elemzéshez felhasználtuk a NarrCat rendszert [4], amely a történetek kompozicionális szerkezetének elemzésére kidolgozott automatikus eljárások gyűjteménye. A NarrCat a történetszerkezet következő komponenseit elemzi: ágencia, személyközi értékelés, érzelem, kogníció, idő, tagadás, valamint téridői és pszichológiai perspektíva. Elemeztük a történetekben előforduló időri lokalizációt is [13]. Az elemzéshez felhasználtuk továbbá a Regresszív Képzeleti Szótárt is [11]. Ez az eljárás az elbeszélő személy regressziójának szintjét méri az elsődleges gondolkodáshoz kapcsolódó jelentésű szavak előfordulása alapján.

A résztvevők érzelmi intelligenciájának méréséhez a Vonás Metahangulat Skálát [14] használtuk. A kérdőív 48 tételt tartalmaz és az érzelmi élmények 4 dimenzióját méri. Ezek a következők: kifejezés, hangulatjavítás, figyelem és tisztaság.

3. táblázat: A történet szerkezeti jellemzői
és az érzelmi intelligencia dimenziói közötti korrelációk

Szerkezeti jellemzők	Érzelmi intelligencia dimenziói			
	Kifejezés	Hangulat-javítás	Figyelem	Tisztaság
Kognitív ige	.04	.01	**.10***	.03
Érzelem szó	.07	-.00	.04	-.01
Pozitív érzelem szó	.06	-00	.06	.00
Negatív érzelem szó	.03	-.00	.00	-.01
Passzív ige	**.20****	**.15****	**.21*****	**13***
Téri-idői perspektíva				
Visszatekintő forma	**.15****	.07	.08	**.11***
Átélő forma	**.21*****	**.11***	**.10***	**.17****
Metanarratív forma	**.22*****	**.14****	**.14****	**.17****
Idői lokalizáció	**.19****	**.11***	.02	**.10***
Én referencia: egyes szám	**.18****	**.18****	**.13***	**17****
Én referencia: többes szám	.03	-.03	.04	.02
Narratív értékelés	**.27*****	**.18****	**.16****	**.20*****

4. táblázat: Az elsődleges és másodlagos fogalmi gondolkodás összetevői
és az érzelmi intelligencia dimenziói közötti korrelációk

Fogalmi gondolkodás	Érzelmi intelligencia dimenziói			
	Kifejezés	Hangulat-javítás	Figyelem	Tisztaság
Elsődleges				
Drive	.08	.05	-.01	.08
Érzékelés	**.13***	.06	.03	**.12***
Regresszió	**.11***	.02	.02	.08
Védekezés	**.13***	.06	.05	**.10***
Ikarosz	-.02	.01	-.05	.00
Másodlagos				
Absztrakció	.06	.08	.02	.03
Erkölcs	.07	-.03	**.12***	.01
Eszköz	-06	.00	.02	-.02
Idő	**.16****	.05	.02	**.11***
Rend	**.14***	**.13***	.02	**.13***
Társas	.09	.04	.05	.07
Korlátozás	**.13***	**.12***	.07	.09

Az érzelmi intelligencia és a történetszerkezet közötti kapcsolat vizsgálatára korrelációs elemzést végeztünk. Az elemzés eredményét a 3. táblázat foglalja össze. Az eredmények azt mutatják, hogy a történetszerkezet jellemzői szoros kapcsolatban vannak az elbeszélő személy érzelmi intelligenciának szintjével. Az érzelmileg intel-

ligensebb személyek történeteinek szerkezete kidolgozottabb: sok narratív értékelést, passzív igét, egyes számú én referenciát, idői lokalizációt, és változatosabb téri-idői perspektívát használnak. Az érzelmi intelligencia és a fogalmi gondolkodás közötti kapcsolat vizsgálatára szintén korrelációs elemzést végeztünk. Az eredményeket a 4. táblázat foglalja össze. Ebben az esetben az összefüggések nem konzisztens mintázatát kaptuk, amely azt mutatja, hogy nincs egyszerűen megfogalmazható összefüggés az érzelmi intelligencia és a fogalmi gondolkodás konstruktumai között.

Köszönetnyilvánítás

A vizsgálatot az OTKA 67914-es számú pályázata és a Bolyai kutatási ösztöndíj támogatta.

Hivatkozások

1. Brewer, W. F., Lichtenstein, E. H.: Stories are to entertain: A structural affect theory of stories. Technical Report, No. 265. Washington, National Institute of Education. (1982).
2. Bruner, J. S.: Two modes of thought. In: Actual minds, possible worlds. Harvard University Press, Cambridge. (1986) 11–43
3. Labov, W., Waletzky, J.: Narrative Analysis: Oral Versions of Personal Experience. In J. Helms (ed), Essays on the Verbal and Visual Arts. Seattle, University of Washington Press (1967) 4–44
4. László, J., Csertő, I., Ferenczhalmy, R., Fülöp, É., Hargitai, R., Péley, B., Pohárnok, M., Pólya, T., Szalai, K., Vincze, O. Ehmann, B.: Narrative language as expression of individual and group identity: The Narrative Categorical Content Analysis. (2013) Sage Open. http://sgo.sagepub.com/content/3/2/2158244013492084
5. Mandler, J.M., Johnson, N.S.: Remembrance of things parsed: story structure and recall. Cognitive Psychology, 9 (1977) 111–151
6. Mayer, J. D., Salovey, P.: What is emotional intelligence? In P. Salovey; D. Sluyter. (Eds.), Emotional development and emotional intelligence. New York, Basic Books. (1997) 3–31
7. Oatley, K. Why fiction may be twice as true as fact: fiction as cognitive and emotional simulation. Review of General Psychology, 3(2) (1999) 101–117
8. Oláh A.: Érzelmek, megküzdés és optimális élmény. Budapest, Trefort Kiadó (2005)
9. Petrides, K. V. Ability and trait emotional intelligence. In Chamorro-Premuzic, T., Furnham, A., & von Stumm, S. (Eds.), The Blackwell-Wiley Handbook of Individual Differences. New York: Wiley (2011) 656–678
10. Pólya T., Kovács I.: Történetszerkezet és érzelmi intenzitás. Pszichológia. 31(3) (2011) 273–294
11. Pólya T., Szász L.: A Regresszív Képzeleti Szótár magyar nyelvű változatának kidolgozása. IX. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2013) 124–132
12. Pólya, T., Kis, B., Naszódi, M. László, J. Narrative perspective and the emotion regulation of a narrating person. Empirical Text and Culture Research, 7(3) (2007) 50–61
13. Pólya, T., Gábor, K.: Linguistic Structure, Narrative Structure and Emotional Intensity. Third International Workshop on Emotion, Corpora for Research on Emotion and Affect. LREC, Valetta (2010) 20–24

14. Salovey, P., Mayer, J.D., Goldman, S.L., Turvey, C., Palfai, T.P.: Emotional attention, clarity, and repair: Exploring emotional intelligence using the Trait Meta-Mood Scale. In J.W. Pennebaker (Ed.), Emotion, disclosure, and health. Washington: American Psychological Association (1995) 125–154
15. Schank, R.C., Abelson, R.P.: Knowledge and memory: The real story. In R. S. Wyer, Jr., (ed), Knowledge and memory: The real story. Advances in social cognition. Lawrence Erlbaum, Hillsdale (1995) Vol. 8. 1–85

A magabiztosság-krízis skála gyakorlati alkalmazása

Puskás László

Pécsi Tudományegyetem Bölcsészettudományi Kara, Pszichológia Doktori Iskola
laszlopuskas@gmail.com

Kivonat: Tanulmányunkban egy korábbi kutatásunk elméleti eredményeinek gyakorlati tapasztalatait kívánjuk bemutatni. 2011-ben a Magyar Számítógépes nyelvészeti Konferencián egy új narratív pszichológiai eljárást ismertettünk, melyben összekapcsoltuk a narratív pszichológiai tartalomelemzést és a vokális mintázatok pszichológiai „tartalomelemzését". Feltételezésünk szerint a krízishelyzet nyelvi-fonetikai mintázata jól körülhatárolható, és ezen jegyek alapján a közlő lelkiállapotára vonatkozóan pszichológiailag értékelhető megállapítások tehetők. A nyelvi tartalmi elemek és a megszólalás fonetikai szerkezetének mintázata alapján létrehoztuk a magabiztosság-krízis indexet, melynek értékéből következtetést vonhatunk le a beszélő lelkiállapotára vonatkozóan. Vizsgálatunk nyelvi anyagát akkor Shakespeare Lear királyának első és utolsó nagymonológja alkotta, melyekben a krízishelyzet előtti és utáni állapotot vizsgáltuk. Meggyőződésünk volt, hogy a színészi játék modellálta helyzet olyan társas tapasztaláson alapul, amely alkalmas lehet arra, hogy a spontán megszólalások mintázatát követve, a való világban előforduló krízishelyzetekre vonatkozóan is értékelhető információt adjon. Jelenlegi kutatásunkban a magabiztosság-krízis indexet spontán megszólalásokon teszteltük.

1 Bevezetés

Az emberek és a csoportok a történetek révén alkotják meg saját identitásuk és pszichológiailag érvényes valóságuk számos lényeges vonását. Ezek a történetek a kompozíciós és élményminőségek alapján vallanak az elbeszélők várható viselkedési adaptációjáról és megküzdési kapacitásáról. A tudományos narratív pszichológia az elbeszélést komplex pszichológiai tartalmak hordozójának tekinti, melynek tanulmányozása révén eredményesen tanulmányozható az emberi társas alkalmazkodás. A pszichológiai folyamatok, az elbeszélés és az identitás közötti szoros kapcsolatot hangsúlyozza. [7]

A narratológia az elbeszélések véges számú alkotóelemét (időviszonyok, nézőpontok, szereplők érzelmi viszonyai stb.) írta le, valamint ezek véges számú kombinációját. Ezek az alkotóelemek a szövegben jól beazonosíthatók, és az elbeszélés így meghatározott komponenseihez pszichológiai jelentéstartalmak társíthatók. [4, 6, 7]

A narratív pszichológiai kutatások az utóbbi évekig nem foglalkoztak az elhangzott szöveg fonetikai paramétereivel. [5] Az elmúlt néhány év kutatási eredményeinek köszönhetően került kifejlesztésre az az eljárás, amely első ízben tett kísérletet a narratív pszichológiai tartalomelemzés és a fonetikai paraméterek vizsgálatának össze-

kapcsolására. A kidolgozott eljárással lefolytatott első gyakorlati vizsgálatok közlésére 2011-ben és 2012-ben került sor. [10, 12] A tanulmány az első vizsgálatok eredményeit követő gyakorlati tapasztalatokkal, és az eljárás továbbfejlesztésének lehetőségeivel foglalkozik. Míg az első vizsgálatok még az érzelmek színész, illetve színműíró általi megfogalmazását vizsgálták, addig a mostani tanulmány már a spontán megnyilatkozásokkal foglalkozik, összehasonlítva a kapott eredményeket a korábbi vizsgálatok tapasztalataival.

2 A tudományos narratív pszichológia és a fonetikai elemzések összekapcsolásának előzményei

A szöveg nyelvi tartalmi jegyeinek és fonetikai struktúrájának párhuzamos vizsgálata, a tudományos narratív pszichológiai megközelítés kereteinek kibővítése, illetve a fonetikai vizsgálatok beépítése a kutatásokba 2004-re nyúlik vissza. Ekkor kezdtem el foglalkozni László János szakmai vezetésével és segítségével egy olyan eljárás kidolgozásán, amely lehetőséget ad a fonetikai elemzések felhasználására a tudományos narratív pszichológiában. Kezdetben egy olyan komplex számítógépes program kidolgozása volt a cél, amely mind a tudományos narratív pszichológiai tartalomelemzés, mind pedig a fonetikai vizsgálatok lefolytatására, és együttes kezelésére alkalmas. A kidolgozott koncepció első komoly bemutatására 2005-ben, az Alpok-Adria Pszichológiai Konferencián került sor Zadarban. A konferencián elhangzott előadás anyaga később egy tanulmánykötetben is megjelent. [11]

2007-ig a koncepció megvalósítása egy olyan eszköz kifejlesztésére irányult, amely a nyelvi tartalmi elemek és a fonetikai jegyek együttes és korlátlan vizsgálatára ad lehetőséget, egy könnyen kezelhető program segítségével. A cél elérésében hangsúlyosabb szerep jutott az eszköz kifejlesztésének, mint egy minden részletre kiterjedő eljárás kidolgozásának. Elsősorban a szemlélet volt a fontos, az, hogy a narratív pszichológia tartalomelemzési módszerei és az elhangzott szöveg fonetikai struktúrájának elemzése integrálható. Nagyjából erre az időre azonban már a tartalomelemzés területén a NooJ program fejlettsége [14], és a hozzá kapcsolódó modulok fejlesztése olyan szintre jutott, amely megkérdőjelezte egy önálló program megvalósításának létjogosultságát, amely ezzel a programmal már nem volt képes felvenni a versenyt; miközben a fonetikai elemzések a Praat programmal is elvégezhetők voltak [1]. Ennek az lett az eredménye, hogy 2008-tól a program kidolgozásról egyre inkább áttevődött a hangsúly egy olyan eljárás kidolgozására, amely képes integrálni a nyelvi és a fonetikai struktúrát. (Később a programfejlesztést végleg elvetettük.)

2011-ben a Magyar Számítógépes Nyelvészeti Konferencián kerültek bemutatásra az újonnan kidolgozott eljárás első eredményei, amelyek Shakespeare Lear királyának első és utolsó monológját elemezték a színész modellálta helyzet fonetikai, és a színműíró szövegének nyelvi tartalmi jegyei alapján. [10] A magabiztosság és a krízis nyelvi jegyeit a Pennebaker és Ireland [9], valamint a László János és munkatársai [6] által kidolgozott módszer segítségével vizsgáltuk. Míg a krízis fonetikai jegyeihez részben Scherer korábbi összefoglaló tanulmánya került felhasználásra, melyben 39 korábbi tanulmány fonetikai vizsgálatait összegezte, részben pedig saját felvetésein-

ket vizsgáltuk, melyek a Scherer-féle koncepcióból is levezethetők. Létrehoztuk a magabiztosság-krízis indexet, amelyben a nyelvi és fonetikai markerek együttesen jelzik a krízis jelenlétét, illetve mértékét. Ez a tanulmány részben a konferencián bemutatásra került eredmények [10], részben pedig a 2012-ben László Jánossal és Fülöp Évával közösen írt tanulmány [12] eredményeinek továbbfejlesztésével, illetve spontán megszólalásokon történő vizsgálatával foglalkozik. A cél továbbra is annak bemutatása, hogy a fonetikai elemzéseknek létjogosultsága van a tudományos narratív pszichológiában, de ennek az ismertetett eljárás csak az egyik lehetséges módja, a fonetikai elemzések felhasználása, integrálása még számos kiaknázatlan lehetőséget rejt magában.

3 A vizsgálat

3.1 A vizsgálati anyag

Vizsgálatunk nyelvi anyagát Stohl Andrással, a neves magyar színésszel készült interjúk alkotják, melyek kutatási célból történő felhasználásához a színész hozzájárult. 2010 májusában a színész személyi sérüléssel járó közúti balesetet okozott. A balesetet követő vizsgálat során vérében a megengedettnél nagyobb mértékű alkoholszintet, valamint kábító hatású anyagot találtak. A színészt – aki 2002-ben már egyszer ittasan balesetet okozott – tíz hónap szabadságvesztésre ítélték, amelyből ötöt kellett letöltenie. Két interjúrészletet hasonlítottuk össze. Az egyik a börtönbe vonulást megelőzően készült, és annak elbeszélését tartalmazza, mi történt a színésszel közvetlenül a baleset után. A másik interjúrészlet a letöltött büntetést követően egy „lakástalkshowban” készült.

3.2 Módszer és eredmények

3.2.1 A fonetikai paraméterek vizsgálata

Az érzelmi állapotok fonetikai paraméterekre gyakorolt feltételezett hatását részben Scherer tanulmányának felhasználásával [13], korábbi vizsgálataink alapján tanulmányoztuk. [10, 12] A lelkiállapot-változásokhoz kapcsolódó, feltételezett akusztikai változásokat az 1. táblázat tartalmazza.

Scherer tanulmánya összefoglalta a fonetikai paraméterek és az érzelmi állapotok közötti kapcsolatot vizsgáló több évtizedre visszatekintő kutatásokat. Nemcsak összegezte a harminckilenc korábbi tanulmány tapasztalatait, de egységes fogalmi keret szerint rendszerezte, mivel a tanulmányok az érzelmi állapotok címkézésére más-más fogalmat használtak.

Az interjúrészletek akusztikai változásait a Praat [1] fonetikai programmal vizsgáltuk, amit az Amszterdami Egyetemen fejlesztettek ki.

1. táblázat: A lelkiállapot-változásokhoz és a közlő pillanatnyi lelkiállapot-változásához kapcsolódó, feltételezett akusztikai változások

	Artikulációs tempó	Hangerő	Hangerő-intervallum	Beszédszakasz hossza	Szünet hossza
Élvezet/boldogság	csökken	csökken	csökken/=	-	-
Jókedv/öröm	nő	csökken/nő	nő	rövid	rövid
Nemtetszés/undor	-	nő	-	?	?
Megvetés/lenézés	-	nő	-	?	?
Szomorúság/levertség	nő	csökken/nő	csökken	-	-
Bánat/kétségbeesés	nő	nő	-	rövid	rövid
Szorongás/aggodalom	-	nő	-	-	-
Félelem/rettegés	nő!	nő!	nő	rövid	rövid
Ingerültség/hideg düh	-	nő!	nő	-	-
Őrjöngés/forró düh	csökken	nő!	nő	rövid	rövid
Unalom/közömbösség	-	csökken/nő	-	-	-
Szégyen/bűntudat	-	nő	-	-	-

A „!” jel megnövekedett erejű változást jósol.

3.2.2 A magabiztosság-dominancia és a krízishelyzet skálázása, a magabiztosság-krízis index

Caplan meghatározása szerint a krízis olyan lelkiállapot, amely külső események hatására alakul ki, amikor az egyének olyan problémákkal találják magukat szemben, amelyek mindennél fontosabbá válnak számukra, és amelyeket sem elkerülni, sem pedig a szokásos eszközökkel megoldani nem tudnak. [2] A krízishelyzetek a meghatározás alapján igen sokfélék lehetnek: közeli hozzátartozó elvesztése, baleset, válás, szakítás, munkahely elvesztése stb.

A krízishelyzet vizsgálatánál figyelembe kell vennünk annak időbeli elhúzódását, illetve feldolgozásának időtartamát is. A krízishelyzet feldolgozásával, kihatásainak elmúlásával, illetve azok kezelhetővé válásával a krízishelyzetre utaló jelek a közlő elbeszélésében jelentősen csökkenhetnek, amiből már legfeljebb a krízis feldolgozottságának mértékére tudunk következtetni.

A magabiztosság és a krízis nyelvi markereinek vizsgálatával kifejlesztett magabiztosság-krízis indexet Shakespeare Lear monológjainak vizsgálatához használtuk fel először. [10, 12] A Stohl Andrással készült interjúrészletek ennek a módszernek a felhasználásával kerültek elemzésre. Az eljárás lefolytatása nemcsak a korábbi vizsgálat megismétlését jelentette, hanem az indexhez felhasznált nyelvi tartalmi elemek, illetve fonetikai paraméterek érvényességének vizsgálatát is. Az indexhez felhasznált

paramétereken kívül módszertani kérdésekkel is kellett foglalkozni. A beszélgetéseknél számolni kell azzal, hogy a beszélgetőpartnerek időnként hangosan helyeselhetnek, közbevághatnak, ami a vizsgált beszédszakaszok akusztikai jellemzőinek mérését megnehezíti. Ennek kezelésére külön módszert dolgoztunk ki. Az index kiszámításához, a korábbi vizsgálatoknak megfelelően, hat arányszámot használtunk fel, melyek értékét egymással összeadtuk [10, 12]:

1. A kettő másodperc alatti beszédszakaszok száma osztva a vizsgált szöveg szószámával – rövid beszédszakaszok.
2. A hangerőcsúcsokat tartalmazó beszédszakaszok száma osztva a vizsgált szöveg szószámával. (Ebbe a kategóriába tartozik minden nyolcvan dB-t meghaladó beszédszakasz, de a megnyilatkozótól függően ennek mértéke a beszélőhöz mérten csökkenthető.) – Magas hangerő.
3. Az alacsony hangerő-intervallumokat tartalmazó beszédszakaszok száma (amelyek nem haladják meg a húsz dB-t) osztva a vizsgált szöveg szószámával – monoton beszéd.
4. A szelf-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával – szelf-referencia.
5. A tagadásra vonatkozó szavak száma osztva a vizsgált szöveg szószámával – tagadás.
6. Negatív korrekciós index: a mi-referenciára vonatkozó szavak száma osztva a vizsgált szöveg szószámával, negatív előjellel – mi-referencia.

Az interjúrészletek vizsgálata technikailag nem okozott nehézséget, mert csak a lakástalkshowban elhangzott interjúrészben van egy olyan beszédszakasz, ahol a beszélgetőpartner az érdeklődés fenntartásának jelzésére szolgáló közbevetést tesz, ugyanakkor olyan általános eljárást kellett találni ennek a problémának a kezelésére, amely alkalmas arra, hogy kezelje a közbevetésekkel, közbeszólásokkal, megakasztásokkal terhelt beszédszakaszokat. Általános szabályként, ha a közbevetés, helyeslés nem akasztja meg a beszélőt, akkor a beszélgetőpartner megszólalásával „terhelt” rész hangerőcsúcsát és hangerőminimumát összehasonlítjuk a beszédszakasz hangerőcsúcsával, illetve hangerőminimumával. Amennyiben a hangerőcsúcs, illetve a hangerőminimum nem éri el a beszédszakaszét, akkor a közbevetés fonetikai paraméterei nem befolyásolják a magabiztosság-krízis indexhez vizsgált hangerőcsúcsot, illetve hangerő-intervallumot, így ebben az esetben a párhuzamos megszólalás nem befolyásolja a vizsgálati eredményeket. Ha a közbevetés tartalmazza a beszédszakasz valamelyik vizsgált kiugró értékét, akkor azt az adott paraméterre, paraméterekre célszerű figyelmen kívül hagyni.

3.2.3 A magabiztosság-krízis indexszel kapott eredmények

A kapott eredmények azt mutatták, hogy – bár feltételezhető, hogy a színész élete a szabadulást követően sem volt stresszmentes – a magabiztosság-krízis indexszel nyert eredményekkel a krízishelyzet egyértelműen kimutatható. Bár a hangfelvétel nem közvetlenül a kritikus életeseményt követően készült, egy olyan lélektani helyzetben adott számot a színész a történtekről, amelyre az autóbalesettel összefüggésben kisza-

bott börtönbüntetés letöltését megelőzően került sor, tehát közvetlen összefüggésbe hozható a krízishelyzetet kiváltó életeseménnyel. A kapott eredményeket összehasonlítottuk a Lear királynál nyert eredményekkel. Azt tapasztaltuk, hogy Lear esetében sokkal nagyobb volt krízishelyzetben az index értéke, ami egyértelműen annak tulajdonítható, hogy Lear esetében egy olyan szélsőséges krízishelyzetről beszélhetünk, ahonnan már nincs tovább: elvesztette a lányait, a sorscsapások következtében elméje erősen megbomlott, és koránál fogva esélye sincs arra, hogy új életet kezdjen. Összehasonlítottuk a „magabiztos megnyilatkozásokat" is. Ebben az esetben azt tapasztaltuk, hogy a színésznél magasabb az index értéke. A kapott eredmény alapján feltételezhető, hogy továbbra is stresszes az élete, de az eredmény nem olyan értékű, hogy abból krízishelyzetre lehetne következtetni. Az index értéke tehát valóban egy skálán helyezhető el, amely összefüggésben áll a krízishelyzet, illetve a magabiztosság mértékével. Vizsgálatunk során az indexhez használt paraméterek használhatóságát is ellenőriztük. Azt tapasztaltuk, hogy korábbi alapvetéseink helytállóak, és az index meghatározása nem szorul komoly kiigazításra. Az index segítségével kapott eredményeket a 2. táblázat mutatja.

2. táblázat: A magabiztosság-krízis index kiszámítása a hat felhasznált mérőszám alapján, Stohl Andrásnál és a Lear monológokban

Mérőszámok	**1**	**2**	**3**	**4**	**5**	**6**	**Összesen**
Stohl András – lakástalkshow	0,0893	0,0000	0,0000	0,0893	0,0476	-0,0119	0,2143
Stohl András – krízisinterjú	0,2010	0,0637	0,0392	0,0490	0,0098	0,0000	0,3627
Lear első monológja	0,0540	0,0270	0,0135	0,0000	0,0135	-0,2162	-0,1082
Lear utolsó monológja	0,3200	0,2533	0,1333	0,1200	0,0133	0,0000	0,8399

A táblázat adataiból látszik, hogy a Stohl Andrással készített „krízisinterjú" magabiztosság-krízis index értéke közel hetven százalékkal nagyobb, mint a lakástalkshownál kapott érték. Ez az eltérés annak fényében különösen jelentős, hogy a lakástalkshow MBK-indexe eleve nem olyan alacsony, és a szelf-referenciára vonatkozó érték nemcsak hogy magas, de jócskán meghaladja a krízisinterjúnál tapasztaltat. Ennek egyértelműen az az oka, hogy a kiválasztott interjúrészletben a színész arról beszél, hogy milyennek látják őt az emberek, illetve ő milyennek látja magát, tehát a téma önmagában is feltételezi a nagyobb szelf-referencia használatot. Az egyes szám első személy használata csak valószínűsítheti a krízist, és annak is csak az egyik eleme. Az indexnél a felhasznált elemek együttes értékéből lehet következtetni a közlő lelkiállapotára, és még ha lehetnek is kilengések egy-egy értékben a végeredmény alapján jó eséllyel következtethetünk a közlő lelkiállapotára. A másik olyan érték, ami meghaladja a krízisinterjúnál mértet, a tagadás, illetve a tagadószavak nagy arányú használata, amely nélkül a lakástalkshowban mért eredmény a krízisinterjúénak kevesebb mint fele lenne. Lényegében a tagadásnál kapott mérőszámmal lépi át

az index azt a határt, ami alapján feltételezhető, hogy a színész élete nem mentes a stressztől, de krízishelyzetről nem beszélhetünk.

A magabiztosság-krízis index kiszámításánál a Lear monológok vizsgálatánál felhasznált módszert használtuk fel a következőképpen [10, 12]:

A fonetikai paraméterek vizsgálatánál a kiválasztott beszédszakaszok számát elosztottuk a vizsgált szöveg szószámával, mert úgy tekintettük, mintha ezen beszédszakaszok mindegyikében kijelöltünk volna egy szót. Ha egy beszédszakaszban több, a vizsgálatnak megfelelő fonetikai paraméter is megjelent, akkor úgy tekintettük, a beszédszakasz szószámától függetlenül, mintha egy másik szót is megjelöltünk volna benne. Ezt a módszert azért alkalmaztuk, mert ha a teljes beszédszakaszt kijelöljük, akkor az olyan lett volna, mintha a kiválasztott szöveg beszédszakaszainak átlagos szószámát jelöltük volna ki benne (ráadásul annyiszor, ahány vizsgált paraméternek a beszédszakasz megfelel). Ez azonban olyan aránytalanságot idéz elő, ami lehetetlenné teszi a fonetikai és a nyelvi paraméterek összeillesztését, mivel a fonetikai paraméterek mérőszámát a hozzájuk képest alacsony értékű nyelvi paraméterek lényegesen nem befolyásolhatnák.

A kettő másodperc alatti beszédszakaszok számának relatív gyakoriságából következtethetünk a közlő gondolatainak összetettségére, arra, hogy az adott helyzettel kapcsolatban milyen korábban konstruált sémával rendelkezik, mekkora a fájdalma. A kiegyensúlyozott megnyilatkozásokban is előfordulhatnak rövidebb beszédszakaszok, de krízishelyzetben jóval nagyobb lehet az előfordulási gyakoriságuk, mivel a helyzet újdonságértékéből adódóan a válaszreakció kevésbé automatikus.

A hangerőcsúcsok a krízishelyzetek fontos indikátorai. Ahogy az 1. táblázatban már ismertetésre került, bánat/kétségbeesés, szorongás/aggodalom és szégyen/bűntudat esetén növekszik a hangerő; félelem/rettegés, ingerültség/hideg düh és őrjöngés/forró düh esetén pedig fokozottan nő.

Az alacsony hangerő-intervallumok gyakorisága egyfajta monotonitást ad a megnyilatkozásnak, amely az erő és a magabiztosság hiányára, rossz lelkiállapotra utalhat.

A szelf-referencia és a tagadás előfordulási gyakoriságának vizsgálatával Pennebaker és Ireland [9], valamint László és munkatársai [4] is foglalkoztak. Az egyes szám első személyű névmások, illetve a személyragok relatív előfordulási gyakoriságát vizsgálták. A túlzott énre utalás a befelé fordulás jele, míg a 'mi'-re utalás a mások irányába való nyitást fejezi ki. Patológiás esetben a magas én-referencia kapcsolatba hozható a depresszióval, a szuicid tendenciákkal. A tagadás pszichodinamikai szempontból az egészséges emberi környezethez és morális mércékhez való alkalmazkodás zavaraira, illetve a világ értéktelenítésére, a destrukcióra és öndestrukcióra való hajlamra utalhat. [3] Krízishelyzetben problémás a megváltozott környezethez való alkalmazkodás, fokozottan fordulhat elő tagadás az elbeszélésben.

A mi-referencia a magabiztosság-krízis index negatív korrekciós mérőszáma, mivel a mi-referencia relatív gyakorisága az indexnél használt többi változóval szemben ellentétes irányú hatást mér. Ezenkívül ennek a változónak a negatív értéke jelentősen csökkentheti a „véletlenszerűen" a megnyilatkozásba került, vizsgált paraméterek relatív előfordulási gyakoriságának értékét, az erőteljes krízisnél kapott eredményeket viszont nem, vagy elhanyagolható mértékben befolyásolja.

4 Összegzés

Tanulmányunk élőszóban elhangzó, spontán megnyilatkozásokban vizsgálta a krízishelyzetet, a szöveg nyelvi tartalmi elemeinek és fonetikai struktúrájának párhuzamos tanulmányozásával. A megnyilatkozásokat a magabiztosság-krízis index segítségével vizsgáltuk. Célunk nemcsak az index használhatóságának ellenőrzése, pontosítása volt, hanem egy olyan új szemlélet meghonosítására is kísérletet tettünk, amely integrálja a tudományos narratív pszichológiai tartalomelemzést és a szöveg fonetikai jegyeinek vizsgálatát. Fontos hangsúlyozni, hogy a magabiztosság-krízis indexszel lefolytatott vizsgálat nem az egyetlen módja a két vizsgálati módszer összekapcsolásának, csupán egyik lehetséges módszere, amelyet a szemlélet meghonosításával remélhetőleg más típusú vizsgálatok is követnek majd.

Az új szemléletmód a tudományos narratív pszichológia és a fonetikai vizsgálatok korábban felhalmozott tudásanyagára épít, figyelembe véve ezeknek a területeknek új kutatási eredményekkel való gazdagodását, illetve gazdagítását is. A két terület összekapcsolását ahhoz lehetne hasonlítani, amikor Puskás Tivadar feltalálta a telefonközpontot: az addig felhalmozódott ismereteket integrálta olyan módon, hogy egy új struktúrát alakított ki, és az új szemlélettel az akkori lehetőségeket a korábbihoz képest jelentősen megnövelte.

A tudományos narratív pszichológia már nemcsak a szavak és témák szintjén vizsgálja az elbeszélések pszichológiai jelentéseit, hanem a narratívum szintjén is. Olyan narratív minőségek mentén vizsgálja a pszichológiai jelentéseket, mint a struktúra, a szervezettség, a perspektíva, az időviszonyok és a koherencia. [8] Ezzel már eleve vizsgálja a történetek nyelv feletti tartalmát is. Amikor tehát a megnyilatkozás fonetikai struktúráját vizsgáljuk, lényegében a tudományos narratív pszichológia elméleti keretein belül maradunk, csupán annak korábbi eszköztárát bővítjük.

Összehasonlítottuk az indexszel kapott eredményeket a Lear monológok esetében a színész, illetve drámaíró modellálta helyzetben, valamint Stohl András spontán megnyilatkozásaiban, mind a magabiztosság, mind pedig a krízishelyzet esetében. Az index értéke jelentős eltérést mutatott krízishelyzetben a „hétköznapi" beszédhelyzetekhez képest, mind a színész modellálta helyzetekben, mind pedig a spontán megszólalásoknál. Összehasonlítottuk a két megnyilatkozónál kapott értékeket is, amelyek mind a magabiztosság, mind pedig a krízishelyzet esetében eltértek egymástól, de mindkét esetben a két pólus „megfelelő" oldalán helyezkedtek el. Az index értéke tehát egy skálán helyezhető el, amelyből nemcsak a közlő krízishelyzetére, illetve magabiztosságára következtethetünk jó eséllyel, hanem annak mértékére, illetve feldolgozottságára is.

A fonetikai struktúra vizsgálatának elsődleges célja a tudományos narratív pszichológiai eljárás eredményeinek gazdagítása, illetve pontosítása volt élőszóban is elhangzó megnyilatkozásoknál. A két vizsgálati módszer összekapcsolása a magabiztosság-krízis indexben – amely a két párhuzamos, de egymással meg nem feleltethető struktúrát vizsgálja – eredményesnek bizonyult, és igazolta a fonetikai struktúra vizsgálatának létjogosultságát a narratív pszichológiai tartalomelemzésekben. A megnyilatkozások nyelvi tartalmi elemeinek és fonetikai jegyeinek párhuzamos vizsgálatát meghonosító szemlélet alapja lehet egy *összetett tudományos narratív pszichológiai eljárás* alkalmazásának, amely remélhetőleg más kutatók figyelmét is felhívja az új

vizsgálati lehetőségekre, és új tanulmányok születnek majd ennek az eljárásnak, illetve ennek a szemléletmódnak a segítségével!

Hivatkozások

1. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer [computer program]. Forrás:http://www.praat.org/ (2013)
2. Caplan, G.: Principles of preventive psychiatry. New York, Basic Books (1964)
3. Hargitai, R. Naszódi, M., Kis, B., Nagy, L., Bóna, A., László, J.: A depresszív dinamika nyelvi markerei az én-elbeszélésekben. A LAS VERTIKUM tagadás és szelfreferencia modulja. Pszichológia, 2 (2005) 181–199
4. László J.: Előszó. Forrás: László J., Thomka B. (szerk.): Narratív pszichológia. Narratívák 5. Budapest, Kijárat Kiadó (2001) 7–15
5. László, J.: Narratív pszichológia. Pszichológia, 28, 4 (2008) 301–317
6. László, J.: The science of stories.: An introduction to narrative psychology. London; New York: Routledge (2008)
7. László, J.: Történelemtörténetek – Bevezetés a narratív pszichológiába. Budapest, Akadémiai Kiadó (2012)
8. László, J., Ehmann, B., Péley, B., Pólya, T.: A narratív pszichológiai tartalomelemzés: elméleti alapvetés és első eredmények. Forrás: Pszichológia, 20. évfolyam, 4. szám (2000) 367–390
9. Pennebaker, J. W., Ireland, M.: Analyzing Words to Understanding. In: Jan Auracher, William van Peer (Eds.): New Beginnings to Literary Studies. Cambridge Scholar Publishing (2008) 24–48
10. Puskás L.: Paralingvisztikai jegyek a narratív pszichológiai tartalomelemzésben: a magabiztosság-krízis skála. Forrás: Takács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (JATE Press) Szeged (2011)
11. Puskás, L., Karsai, B.: A New Method in Narrative Psychology. In: Cognition and Interpretation. Pécs Studies in Psychology. Edited by Beatrix Lábadi. PTE BTK Pszichológiai Intézet (2008)
12. Puskás, L., László, J., Fülöp, É.: Lear király lelkiállapot-változása első és utolsó monológjának szövegbeli és akusztikai jegyei alapján. Pszichológia 2012/2
13. Scherer, K. R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99 (1986) 143–165. Magyarul: Vokális érzelemkifejezés. Áttekintés és egy modell az eljövendő kutatásokhoz. Fordította: Bodor Péter. Forrás: Barkóczi I., Séra L. (szerk.): Érzelmek és érzelemelméletek. Budapest, Tankönyvkiadó (1989)
14. Silberztein, M.: NooJ manual. Forrás: http://www.nooj4nlp.net (2003)

V. Orvosi NLP

Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben

Siklósi Borbála[1], Novák Attila[1,2]

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

Kivonat Az orvosi szövegek feldolgozása ma a nyelvtechnológia egyik legaktívabban kutatott részterülete. Az általános szövegekre ma már jól működő eszközök helyes, normalizált bemenetet feltételeznek. Orvosi szövegek esetén ez a feltétel nem teljesül, ezért az ezekre jellemző nagy mennyiségű zaj miatt kész eszközök alkalmazása nem lehetséges. A normalizálás egyik lépése a rövidítések észlelése és feloldása. Ebben a cikkben egy nem felügyelt automatikus módszert mutatunk be rövidítéssorozatok feloldására magyar nyelvű klinikai dokumentumokban. Három módszert ismertetünk, melyek különböző mértékben támaszkodnak külső erőforrásokra, illetve magára a klinikai korpuszra.

1. Bevezetés

Az orvosi szövegek feldolgozása ma a nyelvtechnológia egyik legaktívabban kutatott részterülete. Olyan programcsomagokból, melyek orvosi szövegekből nyernek ki a felszínen nem elérhető információkat és összefüggéseket, angol nyelven számos jól működő megvalósítás létezik. Az orvosi protokollok helyi jellegzetességeire vonatkozó adatok, illetve a helyi közösségeket érintő járványügyi információk azonban csak az adott nyelven lejegyzett szövegekben fedezhetők fel. A magyar nyelvű klinikai szövegek feldolgozására alkalmas eszközök létrehozása tehát nemcsak érdekes kihívás a nyelvi nehézségek miatt, hanem szükséges feladat is.

Azok a dokumentumok, amelyek kórházi körülmények között, nyelvi ellenőrzés nélkül jönnek létre, jellemzően sok helyesírási hibát tartalmaznak, tele vannak magyar-latin szakkifejezésekkel, illetve következetlenül használt rövidítésekkel [1,2]. Ezeknek a rövidítéseknek a használata sok esetben követ valamilyen szabályrendszert, de legtöbbször mégsem felelnek meg a rájuk vonatkozó hivatalos szabályzatnak, nem is beszélve a nem szándékos, ámde gyakori elírásokról. Ezért a rövidítések feloldása nem oldható meg egy lépésben, azoknak egy lexikonra való egyszerű illesztésével. Továbbá, a klinikai körülmények között létrejött dokumentumokban sokszor hosszabb, sok szóból álló szerkezetek szinte minden szava rövidítve van, nem csupán elvétve találunk egy-egy rövidítést az

egyébként teljes szavakat tartalmazó mondatokban [3]. Az egy egységet alkotó kifejezések elhatárolása az ilyen rövidítéssorozatokban sokszor még emberi szakértők számára is kihívást jelent, eltekintve az adott szöveg szerzőjétől, akinek remélhetőleg teljesen érthetőek a saját maga számára készített feljegyzések. Ha a hosszabb rövidítéssorozatokban az egyes szavakat tokenenként próbálnánk feloldani az egyes rövidítéseket egymástól függetlennek tekintve, az nagyon nagy számú feloldási kombinációt eredményezne. Ez az ilyen kijelentések jelentésének egyértelműsítése helyett a feloldásból eredő zajt növelné a szövegekben. A klinikai dokumentumok feldolgozása tehát nem egyszerű feladat, melynek részeként a rövidítések feloldása minden további lépés előfeltétele.

A rövidítések feloldásához használt külső lexikonok használata azért sem vezet önmagában megoldáshoz, mert ilyen erőforrások magyar nyelvre csupán korlátozott mértékben és minőségben állnak rendelkezésre. A BNO-kódrendszer hivatalos leírása az egyik ilyen elérhető adatbázis, azonban ennek használatakor is külön feladat a rövidítésekből a megfelelő, a leírásokra illeszthető minták előállítása. Ez tehát korántsem alkalmazható olyan közvetlen módon, mint az angol nyelven elérhető UMLS (Unified Medical Language System) rövidítéstára, amely a legtöbb angol orvosi rövidítést, azok változatait és lehetséges feloldásait tartalmazza [4].

Ha lenne is magyar nyelven elérhető ilyen erőforrás, az csupán a lehetséges feloldási javaslatok kigyűjtésére lenne alkalmas. A javaslatok megfelelő rangsorolásához, melynek során a szövegkörnyezetben is helytálló feloldásnak kellene első helyezettként megjelenni, megfelelően egyértelműsített nyelvmodellre lenne szükség. Mivel azonban nincsen olyan orvosi korpusz, amiben a rövidítések helyett azok kifejtett formája szerepelne, ezért ilyen nyelvmodell sem áll rendelkezésre. Egy ilyen korpusz létrehozása pedig olyan nagy mennyiségű és drága szakértői munkát igényelne, ami jelen kutatás keretei között nem volt megvalósítható.

A bemutatott kutatás célja a több tokenből álló rövidítéssorozatok automatikus feloldása külső erőforrások és a rendelkezésünkre álló klinikai korpusz felhasználásával. Bemutatjuk, hogy ebben a folyamatban szükséges, de nem elégséges a kész lexikonok és az általunk készített kisebb, korpuszspecifikus lexikon használata. Módszerünk hatékonyságát ugyanakkor jelentős mértékben növelte az algoritmus kiegészítése egy olyan lépéssel, amelynek során a rövidítéseket a korpusz szövegére illesztve is keresünk feloldásjelölteket. Ezzel biztosítható továbbá a doménfüggetlenség, hiszen a korpuszt a maga nyers formájában használjuk fel, tehát módszerünk a nem felügyelt algoritmusok körébe tartozik.

2. Az orvosi korpuszban előforduló rövidítések jellemzői

A rövidítések sorozatából álló jegyzetelési stílus bevett szokás a klinikai jegyzetek, dokumentumok létrehozása során. Ez a tömörített írásmód számos hivatalos és egyedi rövidítést vagy jelölést tartalmaz, amelyek nagy részének használata csak az adott szakterületre, esetleg csak egy orvosra vagy asszisztensre jellemző. A rövidítések jelölhetnek az adott orvosi szakterület körében releváns fogalmakat, vagy olyan hétköznapi szavakat és kifejezéseket, amelyek a klinikai szövegekben

gyakran fordulnak elő, ezért bevett szokás a rövidített alak használata. A szakértő olvasó számára az ilyen rövidített alakok jelentése általában éppen annyira egyértelmű, mint a szabványos rövidítések esetén, hiszen kellő ismerettel és gyakorlattal rendelkezik, valamint tisztában van a szövegkörnyezet jelentésével is. Az 1. táblázatban látható néhány példa a különböző rövidítéstípusokra. Vannak közöttük elterjedt, gyakran használt, egyértelmű alakok, melyek általában latin eredetűek. Másoknak azonban még az orvosi szakterületen belül is több jelentése lehet.

1. táblázat. Példák a korpuszban előforduló rövidítésekre.

Domén	Rövidítés	Feloldás	Magyarul
szabványos	o. d.	oculus dexter	jobb szem
	med. gr.	mediocris gradus	közepes fokú
doménspecifikus	o. (ophthalmology)	oculus	szem
	o. (general anatomy)	os	csont
általános szóhasználatú	sü	saját szemüveg	saját szemüveg
speciális kifejezések	fén	fényérzés nélkül	fényérzés nélkül
	n	normál	normál
általános szavak	köv	következő	következő
	lsd	lásd	lásd

A folyó szövegekben található rövidítésekkel kapcsolatos első probléma azok felismerése. Mivel ezek a szövegek nem követik a helyesírási és központozási szabályokat a rövidítések jelölésének a területén sem, ezért ezek felismeréséhez nem elegendő a rájuk vonatkozó helyesírási szabályok formalizált alkalmazása. A rövidítést jelző pontok általában hiányoznak a rövidített szóalakok végéről, a rövidítésekben vegyesen szerepelnek kis- és nagybetűk, jellemző, hogy a betűszavakat is csupa kisbetűvel írják, valamint ugyanannak a szónak vagy kifejezésnek számtalan különböző hosszú rövidítése lehet. A következő formák például mind ugyanazt a fogalmat jelölik: *vvf*, *vvfény*, *vörösvfény* - ezek mind a "vörös visszfény" kifejezés rövidített alakjai.

A 600792 tokenből álló klinikai korpuszban 3154 különböző rövidítést azonosítottunk automatikus módszer alkalmazásával (l. a 4.2 bekezdést). Egy rövidítésnek tekintettük azokat a rövidítéssorozatokat is, amelyeket nem tör meg semmilyen teljes szóalak. Természetesen ezeknek a szekvenciáknak az egyes tagjai különálló rövidítések is lehetnek. A következő példamondatban tehát négy rövidítés(sorozat) található.

Dg : Tu. pp. inf et orbitae l. dex. , Cataracta incip. o. utr. , Hypertonia.

A rövidítések:

Dg,
Tu. pp. inf,
l. dex.,
incip. o. utr..

A példában szereplő utolsó minta félrevezető, hiszen az *incip.* token az őt megelőző szóhoz (*Cataracta*) kapcsolódik szemantikailag, ami viszont nem része a mondatban felismert rövidítések halmazának. A kifejezések ilyen vegyes formában való leírása igen gyakori, továbbá változó a rövidített és a teljes alakban kiírt szavak megválasztása is.

A rövidítéssorozatok egyes tagjainak a jelentése a szövegkörnyezet figyelembevétele nélkül általában nem határozható meg egyértelműen az egyes rövidítések nagyfokú többértelműsége miatt. Így ha létezne is a magyar orvosi nyelvre vonatkozó rövidítések teljes és jól használható listája, az egyes rövidítések erre való illesztése nem oldaná meg a problémát, csupán javaslatokat tudna tenni a lehetséges feloldásokra. Sok esetben egyetlen önmagában álló rövidítésre nagyon nagy számú javaslat érkezhet. (További problémát jelent a klinikai dokumentumok keverék magyar-latin nyelvezete, ezért már a rövidítéseknél is fontos azok nyelvének megkülönböztetése.)

Annak ellenére azonban, hogy az egyes rövidítések önmagukban állva erősen többértelműek, gyakran fordulnak elő rövidítéssorozatok részeként, ahol biztosabban meghatározható az egyértelmű jelentésük. Például, az "o." rövidítés bármely o-val kezdődő magyar vagy latin szó rövidítése is lehet. Még az orvosi szaknyelvre szűkítve is igen nagy a lehetőségek száma. Az általunk vizsgált klinikai korpusz szemészeti részében azonban az "o." rövidítés csak elvétve fordul elő önmagában, sokkal inkább olyan szerkezetekben, mint például "o. s.", "o. d.", vagy "o. u.", melyek jelentése *oculus sinister* (bal szem), *oculus dexter* (jobb szem), illetve *oculi utriusque* (mindkét szem). Az ilyen összetételekben az "o." jelentése már egyértelműen meghatározható. Természetesen az ugyanazzal a jelentéssel bíró rövidített alakoknak is számos variációja előfordulhat, így az "o.s." gyakori változatai például az "o. sin.", "os", "OS" stb. A példában szereplő kifejezések változataira vonatkozó gyakorisági adatokat tartalmaz a 2. táblázat.

2. táblázat. Három gyakori kifejezés: *oculus sinister*, *oculus dexter* és *oculi utrisque* néhány rövidített alakja, azok korpuszbeli gyakoriságával.

oculus sinister	freq	oculus dexter	freq	oculi utriusque	freq
o. s.	1056	o. d.	1543	o. u.	897
o.s.	15	o.d.	3	o.u.	37
o. s	51	o. d	188	o. u	180
os	160	od	235	ou	257
O. s.	118	O. d.	353	O. u.	39
o. sin.	348	o. dex.	156	o. utr.	398
o. sin	246	o. dex	19	o. utr	129
O. sin	336	O. dex	106	O. utr	50
O. sin.	48	O. dex.	16	O. utr.	77

Elsődleges célunk az olyan rövidítéssorozatok felismerése volt, melyek egyben vizsgálva egyértelműen feloldhatóak. Mivel sok esetben teljes kijelentések,

vagy akár mondatok vannak csak rövidített alakokkal megfogalmazva, ezért az első lépés a hosszabb rövidítéssorozatok önálló jelentéssel bíró partíciókra való optimális felosztása. A fenti példában szereplő "incip. o. utr." sorozat optimális felbontását az "incip." és "o. utr." különválasztásával kapjuk, akkor is, ha az "incip." szó jelentése önmagában nem értelmezhető, azonban az nem része az "o. utr." tokenek által rövidített kifejezésnek sem.

3. Módszerek

A feladat során szemészeti osztályon keletkezett magyar nyelvű klinikai szövegekkel foglalkoztunk. Először a rövidítéseket azonosítottuk, majd három módszert alkalmaztunk a jelentéssel bíró egységek felismerésére és feloldására. Az utóbbi két problémát mindhárom módszer esetén egy lépésben oldottuk meg, ezzel érve el egyszerre az optimális lefedettséget és a jelentés meghatározását.

3.1. Rövidítések felismerése

A rövidítések azonosítása során nem támaszkodhatunk olyan felszíni tulajdonságokra, amik általános esetben egy token rövidítés mivoltára utalnának (pont a szó végén, csupa nagybetűs mozaikszavak, stb.). Ezért néhány heurisztikus szabályt alkalmaztunk, többek között a következő jellemzők figyelembevételével: pont jelenléte, vagy hiánya a szóalak végén; a szóalak hossza; magánhangzók és mássalhangzók aránya a szóalakon belül; a kis- és nagybetűk aránya a szóalakon belül; a HuMor morfológiai elemző [5,6] ítélete az adott szóalakról. A rövidítéseket azonosító algoritmus részletes ismertetésére ebben a cikkben nincs módunk. A rövidítések felismerésére alkalmazott módszerünkkel magasabb fedés és alacsonyabb pontosság garantálható, ami a további feldolgozás szempontjából előnyös. Célunk nem az egyes rövidített alakokat jelölő tokenek kinyerése, hanem a rövidítéssorozatok megtalálása, amiket később bontunk szemantikailag releváns részekre. Így, ha egy önmagában álló szót tévesen rövidítésként jelölünk úgy, hogy egyik szomszédja sem rövidítés, akkor az nem kerül a feloldandó rövidítések közé. Másrészt viszont, ha egy tokent tévesen beveszünk egy rövidítéssorozatba, akkor a feloldó algoritmus fogja biztosítani azt, hogy ne kerüljön feloldásra, hiszen nem tud majd rá optimálisan illeszthető feloldást találni. Például, az *Exstirp. tu. et reconstr. pp. inf. l. d.* sorozatban az *et* latin szó nem rövidítés. A sorozat feldarabolása során nem is lesz semmilyen szemantikailag összetartozó csoport része, sem az őt megelőző, sem a rákövetkező szóalakokhoz nem csatolható.

3.2. Rövidítések feloldása

Feloldási lehetőségek keresése külső erőforrásokban. Miután kinyertük a lehetséges rövidítéssorozatokat, egy maximális lefedést biztosító feloldási javaslatokat generáló rendszert alkalmaztunk. Az algoritmus a következő: egy rövidítéssorozat esetén annak összes lehetséges, nem átfedő felosztására reguláris

kifejezéseket generálunk, amiket aztán a rendelkezésünkre álló lexikonokra illesztünk. A reguláris kifejezések a rövidítés szabályai alapján jönnek létre, mint például minden egyes betű a rövidített kifejezés egy szavának kezdőbetűje, vagy többtagú rövidítések esetén, az egyes tagok felelnek meg egy-egy szó kezdetének. A 3. táblázat tartalmaz néhány ilyen szabályt leíró mintát. A létrejött minták száma és komplexitása arányos a vonatkozó rövidítéssorozat hosszával.

3. táblázat. Két rövid rövidítésből generált illesztendő reguláris kifejezések.

rövidítés	regexp	illeszkedő feloldás	regexp	illeszkedő feloldás
o. s.	`o[^ ]* s[^ ]*`	oculus sinister		
os	`os[^ ]*`	osteoporosis	`o[^ ]* s[^ ]*`	oculus sinister

A reguláris kifejezések illesztésére használt lexikonok egyike a BNO kódrendszer szemészeti szekcióinak leírásaiból és az Orvosi helyesírási szótárból [7] készített, 3329 elemet tartalmazó szólista volt. A másik lexikon egy kisebb méretű, szakterületi szakértő segítségével kézzel készített doménspecifikus kifejezéslista volt. Ebbe a listába olyan kifejezések kerültek be, amelyek olyan hétköznapi kifejezések rövidítései, melyek a szemészeti leírásokban egyedi feloldással, jelentéssel bírnak (például: “mou”, azaz *méterről olvas ujjat*). A feloldási javaslatok rangsorolásánál figyelembe vettük, hogy a javaslat melyik lexikonból származik.

A feloldási javaslatok meghatározása után azok mindegyike egy pontszámot kap. A legnagyobb lefedettséget és a legjobb feloldást egyszerre előnyben részesítő pontszámot három tulajdonság alapján határozzuk meg: 1) a feloldott alak hány tokent fed le az eredeti rövidítéssorozatból, 2) hány tokenből áll a rövidítéssorozat feldarabolása során keletkezett legnagyobb partíció, 3) hány tokenből áll a rövidítéssorozat feldarabolása során keletkezett legkisebb partíció. A fenti példában szereplő *Exstirp. tu. et reconstr. pp. inf. l. d.* sorozat esetén, annak az *Exstirp. tu. – et – reconstr. pp. inf. – l. d.* felbontására vonatkozó három szám a sorrendnek megfelelően: 7, 3 és 2.

Feloldás keresése a korpusz alapján. A fent ismertetett módszer legnagyobb hátránya az, hogy csak a hivatalos leírásokra illeszthető rövidítések oldhatóak fel a segítségével. A klinikai szövegekben azonban szabadon rövidítenek mindent, az ezeknek megfelelő kifejezések pedig nem találhatóak meg a hivatalos erőforrásokból épített lexikonokban. A viszonylag szűk domén miatt azonban feltételezhetjük, hogy az ilyen rövidített kifejezéseknek (vagy azok egyes részeinek) a kifejtett alakjai is legalább egyszer szerepelnek a korpuszban. Ezért ebben az esetben is a rövidítéssorozatok összes lehetséges feldarabolása során kapott egységekből képzett reguláris kifejezéseket illesztjük magára a korpuszra. Azzal a különbséggel tesszük ezt, hogy az egytagú darabokra kapott eredményeket nem vesszük figyelembe (ezek az általános szavak miatt a feloldási javaslatok listájában csak a zajt növelnék), illetve a korpuszban adott gyakoriságnál ritkábban előforduló illeszkedő kifejezéseket sem vesszük hozzá az eredményekhez.

Korpuszillesztés és külső erőforrások együttes alkalmazása. A harmadik esetben a fenti két módszert együtt alkalmaztuk, ezáltal érvényesítve előnyeiket és egymás által pótolva hiányosságaikat. A rövidítéssorozatok összes lehetséges felbontására először a korpuszban való keresést végezzük el, majd az így megkapott, részlegesen feloldott sorozatokra a külső lexikonokban is elvégezzük a reguláris kifejezések illesztését. Ezáltal a korpusz alapján való illesztésből maradt "lyukak" pótolhatóak.

A korpusz ilyen módon való felhasználása nem felügyelt módszerrel történik, ezáltal bármilyen más olyan aldoménre alkalmazható, amire egy nyers korpusz rendelkezésre áll. Ahogy a kiértékelés során később részletezzük, a korpuszban való kereséssel a rendszer robosztusabbá tehető, a teljesítmény sokkal kisebb mértékben esik a kézzel készített lexikon méretének csökkentése esetén.

4. Eredmények

A klinikai korpuszból 23, előre tokenizált dokumentumot különítettünk el tesztelési célra (összesen 4516 token). Ebben a teszthalmazban az automatikus felismerő 323 különböző rövidítést, illetve rövidítéssorozatot azonosított. Ezek közül kézzel választottuk ki azt a 44 egyedi rövidítéssorozatot (összesen 140 token), melyre a kiértékelést végeztük. Ezek a sorozatok legalább kétszer előfordulnak a tesztkorpuszban, és legalább két token hosszúságúak. A 4 táblázatban ezek közül szerepel néhány példa, a különböző rendszerek által generált feloldásokkal és a szakértő segítségével meghatározott tényleges feloldással együtt. Az automatikus rendszereknél az első helyre rangsorolt javaslatot tekintettük a végleges feloldásnak.

A kiértékelést két szinten végeztük. Megvizsgáltuk az egyes rendszerek teljesítményét a teljes, többszavas feloldások szintjén, és az egyedi tokenek szintjén is. Az első esetben, egy sorozat feloldása akkor és csak akkor helyes, ha annak minden tagját sikerült helyesen meghatározni. A második esetben a helyesen feloldott tokenek számát mértük, ami nyilvánvalóan jobb mérési eredményeket adott minden rendszer esetén. A teljesítmény mérésére a fedés, pontosság és F-mérték metrikákat használtuk a következő definíciókkal: a pontosság a helyes feloldások száma osztva az összes, bármilyen minőségű feloldás számával (sorozat- vagy tokenszinten), a fedés a helyes feloldások száma osztva az összes elem számával (sorozat vagy token szinten). Az F-mérték pedig a kettő harmonikus közepe. Az 5. táblázatban szerepelnek az automatikus kiértékelés számszerű eredményei.

Az eredményekből világosan látszik, hogy a korpusz önmagában való használata nem kielégítő. Ez nem is meglepő, hiszen éppen a leggyakoribb rövidítések azok, amelyek sosem szerepelnek kifejtett formában a szövegekben. A külső erőforrásokra tehát szükség van, de a kizárólag ezekre építő rendszer teljesítménye is rosszabb, mint a korpuszt és a lexikonokat együtt használóé.

A lexikonokat használó rendszerek esetén külön megvizsgáltuk a kézzel készített szótár jelentőségét. Ehhez a kiértékelést elvégeztük ennek a lexikonnak egy csökkentett verziójának használata mellett is. Az eredetileg 97 rövidítést, és azok feloldását tartalmazó listát 70-re csökkentettük, ami 28%-os méretváltó-

4. táblázat. Néhány példa az egyes rendszerek által automatikusan feloldott rövidítéssorozatokra, összehasonlítva a szakértő által megadott tényleges feloldással.

	Cat. incip. o. utr.
1. módszer	cat. incip. oculi utriusque
2. módszer	cat. incip. o. utr.
3. módszer	cataracta incipiens oculi utriusque
gold standard	cataracta incipiens oculi utriusque
	Myopia c. ast. o. utr.
1. módszer	myopia kritikus fúziós frekvencia ast. oculi utriusque
2. módszer	myopia cum ast. o. utr.
3. módszer	myopia cum astigmia oculi utriusque
gold standard	myopia cum astigmia oculi utriusque
	myop. maj. gr. o. u.
1. módszer	myop. maj. gr. oculi utriusque
2. módszer	myop. maj. grad. o. utr.
3. módszer	myopia maj grad. oculi utriusque
gold standard	myopia major gradus oculi utriusque
	med. gr. cum
1. módszer	med. gr. cum
2. módszer	med. gr. cum
3. módszer	med. gr. cum
gold standard	medium gradus cum

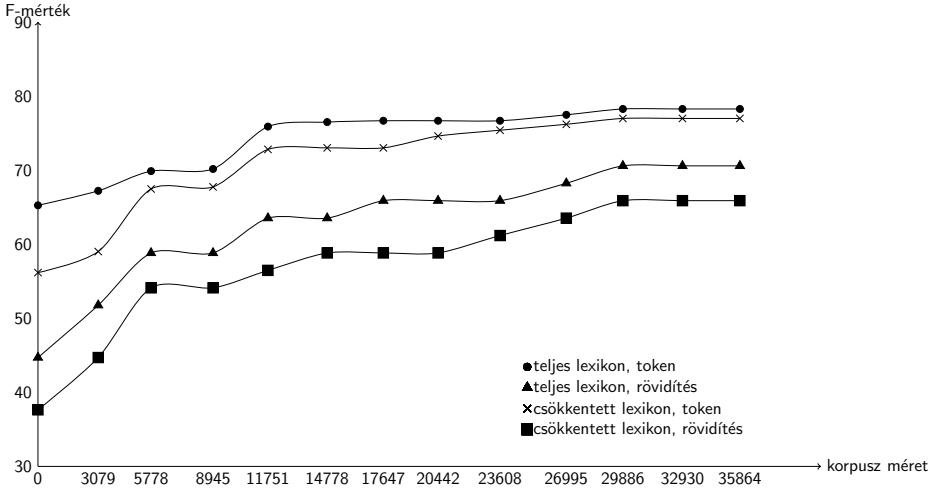
zást jelent. A másik lexikon mérete minden mérés során ugyanakkora volt. Bár a méretcsökkentés során mindegyik rendszer teljesítménye romlott, ez a romlás a korpuszt is felhasználó esetben sokkal kisebb mértékű. Ebben az esetben azokat a kifejezéseket, amelyeket a lexikonból töröltünk, a rendszer automatikusan pótolta a korpuszból. Az 1. ábrán a korpuszt használó rendszerek tanulási görbéje látszik a felhasznált korpusz méretének függvényében. Az x tengely 0 pontja megfelel a csak lexikont használó rendszernek (0 méretű korpusz). Ebben a pontban a saját lexikon méretének csökkentésével járó teljesítménybeli különbség még jelentős mértékű, de ezt a lemaradást a felhasznált korpusz növelésével a rendszer automatikusan behozza.

5. Konklúzió

Bemutattuk, hogy a magyar nyelvű klinikai dokumentumokban található rövidítések feloldása során szükség van ugyan külső lexikai erőforrásokra, a feloldás minősége azonban jelentős mértékben javítható a korpuszra alapuló nem felügyelt tanulási algoritmus használatával. Ezáltal megspórolható a drága és nagy erőfeszítésekkel járó, szigorúan doménspecifikus lexikonok kézi összeállítása. A rövidítések többértelműségének problémáját pedig azok sorozatokban való kezelésével oldottuk meg, így elkerülhetővé vált az egytagú rövidítések lehetséges

5. táblázat. A kiértékelés eredményei tokenek és teljes rövidítések szintjén, teljes és csökkentett saját lexikon használata mellett

	pontosság		fedés		F-mérték	
	abbr.	token	abbr.	token	abbr.	token
1. módszer (teljes lexikon)	46.34%	78.57%	43.18%	55.79%	44.70%	65.25%
1. módszer (csökkentett lexikon)	39.02%	68.04%	36.36%	47.82%	37.64%	56.17%
3. módszer (teljes lexikon)	73.17%	86.08%	68.18%	71.73%	70.58%	78.26%
3. módszer (csökkentett lexikon)	68.29%	85.08%	63.63%	70.28%	**65.88%**	**76.98%**
2. módszer (lexikon nélkül)	6.66%	41.79%	4.54%	20.28%	5.4%	27.31%



1. ábra. Az egyes rendszerek tanulási görbéje a felhasznált korpusz méretének függvényében

feloldásaiból kialakuló kezelhetetlen méretű keresési tér generálása. A mérési eredményekből kiderült, hogy bár van még lehetőség a minőségbeli javulásra a pontosság szempontjából, azonban látható az is, hogy a rendszer könnyen adaptálható bármilyen szűk domén rövidítéseinek feloldására.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014 pályázatok támogatásával készült.

Hivatkozások

1. Siklósi, B., Orosz, G., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.M., Forcada, M., M. Tyers, F., Waiganjo Wagacha, P., eds.: 8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages. (2012) 29.–34.
2. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. In Dediu, A.H., Martin-Vide, C., Mitkov, R., Truthe, B., eds.: Statistical Language and Speech Processing. Volume LNAI 7978., Springer Verlag (2013)
3. Barrows, J.R., Busuioc, M., Friedman, C.: Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proceedings of the AMIA Annual Symposium (2000) 51–55
4. Liu, H., Lussier, Y.A., Friedman, C.: A study of abbreviations in the UMLS. Proceedings of the AMIA Annual Symposium (2001) 393–397
5. Novák, A.: What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
6. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 261–268
7. Fábián, P., Magasi, P.: Orvosi helyesírási szótár. Akadémiai Kiadó, Budapest (1992)

Hol a határ?
Mondatok, szavak, klinikák

Orosz György, Prószéky Gábor

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

Kivonat Napjainkban egyre több elektronikusan rögzített dokumentum keletkezik klinikai környezetben, melyek egyik közös jellemzője, hogy létrehozásuk során a klinikai dolgozók nem fordítottak figyelmet a dokumentumok struktúrájának kialakítására, illetve a helyesírási normák betartására. Bár a mondat- és szóhatárok megállapítása egy olyan alapvető feladat, mely a feldolgozási lánc legelején helyezkedik el, irodalma mégsem jelentős, mivel ezt gyakran mérnöki munkának tekintik a kutatók. Jelen írásunkban ismertetjük a klinikai dokumentumok sajátosságait, különös tekintettel a mondat- és szóhatárok kérdésére. Részletesen bemutatunk egy hibrid szegmentáló algoritmust, mely szabályalapú részek mellett nem felügyelt gépi tanulást is használ. Az ismertetett módszer eredményességét részletesen megvizsgáljuk, másrészről összevetjük azt más magyar nyelvre elérhető rendszerekkel. Megmutatjuk, hogy a komplex eljárás teljesítménye jelentős mértékben meghaladja az alapjaként szolgáló szabályalapú rendszerét. Összevetve más mondatszegmentáló (és tokenizáló) metódusokkal, megállapítjuk, hogy csak az ismertetett új algoritmus képes oly mértékben mondat- és tokenhatárok azonosítására, hogy az a gyakorlatban is használható legyen.

1. Bevezetés

Magyarországon a napról napra keletkező nagy mennyiségű klinikai dokumentumok jelentős hányada csak archiválási célból készül és nem kerül feldolgozásra. Ezek nyelvtechnológiával támogatott újrafelhasználása, más nyelvekhez hasonlóan, nagy mértékben képes lenne segíteni a klinikákon praktizáló orvosokat jobb diagnózisok vagy új terápiák kifejlesztésében. A feldolgozó- és információkinyerő-eljárások legtöbbje a bemeneti szöveget mondatokra és/vagy szavakra bontva várja, így ezek pontos elvégzése szükségszerű. Bár az általános nyelvre léteznek nagy teljesítményű szegmentáló eszközök, de ezek alkalmazhatósága klinikai szövegeken nem bizonyított.

Írásunkban megvizsgáljuk a klinikai környezetben készült rekordokat, rávilágítva azok különleges tulajdonságaira. Bemutatunk egy kis méretű korpuszt, melyet az eszközök fejlesztése céljából hoztunk létre, majd ismertetünk

egy nagy teljesítményű szegmentáló algoritmust. Az eljárás szabályalapú komponenseken túl gépi tanuló (GT) algoritmusokat is foglalkoztat. Az utóbbi módszer alapja, hogy a nyers szövegekben pontra végződő tokenekről meghatározza, hogy a pont és a szó egybeírása csak a véletlen műve (mondathatár) vagy pedig szisztematikus használat eredménye (rövidítés). A pontosabb és teljesebb feldolgozás érdekében az eljárás számos más jellemző mellett morfológiai elemzéseket is használ.

A tesztkorpuszon végzett kiértékelésünkben megmutatjuk, hogy a klinikai szövegeken egyetlen szabadon elérhető eszköz sem teljesít megfelelően, míg az általunk fejlesztett algoritmus a gyakorlatban is jól használható.

2. Kapcsolódó munkák

2.1. Mondatok és tokenek azonosítása

A szövegek alkotóelemeinek keresése két részfeladatból tevődik össze: mondathatárok azonosítása és tokenekre bontás. Nagyon gyakran egy mondathatárkereső algoritmus feltételezi a rövidítések ismeretét, vagy magában foglalja azok azonosítását is. Míg a tokenizálást gyakran mérnöki feladatként kezeljük, ezzel szemben a mondathatárok felismerésének bővebb irodalma van. Read et al. összefoglaló írásában [1] az alábbi csoportokba osztja az ezzel foglalkozó kutatások: 1) szabályalapú rendszerek, amik domén- vagy nyelvspecifikus tudást használnak; 2) felügyelt gépi tanuláson (FGT) alapuló algoritmusok; 3) felügyelet nélküli gépi tanulást (FNGT) használó módszerek.

A gépi tanulást (GT) alkalmazó megoldások közül az egyik első Riley [2] algoritmusa volt, melyben döntési fákat használt mondatvégi írásjelek osztályozására. Analóg megközelítéssel bír a SATZ [3] keretrendszer, melyben számos FGT módszer érhető el, ami ezeken túl a szófaji címkék mint jellemzők használatára is képes. Az első eredmények, melyek maxent tanulást használtak mondatok szegmentálására, Reynar és Ratnaparkhi nevéhez fűződnek [4]. Másrészről a Gillick által bemutatott algoritmus [5] hasonló jellemzőket használva SVM módszeren alapul. Ismeretesek még Mikheev munkái, melyek közt szerepel egy szabályalapú eszköz [6], illetve ennek integrált használata egy szófaji egyértelműsítő keretrendszerben [7]. Az általunk ismert egyetlen FNGT-on alapú módszert Kiss és Strunk készítette, mely többszavas kifejezéseket azonosító algoritmust használ annak eldöntésére, hogy egy szó és egy pont rövidítést alkot-e.

Magyarra az ezidáig publikált alkalmazások szabályalapú megközelítést használnak: a huntoken [8] eszköz Mikheev rendszerén [6] alapul, míg a `magyarlanc` [9] hasonló modulja a MorphAdorner projekt [10] eredményeire épít.

2.2. Orvosi szövegek feldolgozása

Magyar nyelvű orvosi szövegek feldolgozásának irodalma ezidáig nem jelentős: Siklósi et al. [11,12] megoldása automatikus módon képes klinikai szövegek helyesírásának javítására, míg Orosz et al. egy morfológiai egyértelműsítő

rendszer teljesítményének növeléséről számolnak be [13]. Orvosi szövegek automatikus szegmentálásának kérdését egyik mű sem érinti.

Magyartól eltérően, az angol nyelvű orvosi szövegek szegmentálásának irodalma bővebb: mondatra bontó eljárásokként leginkább szabályalapú (pl. [14]) vagy FGT-t használó módszereket [15,16,17,18,19] használnak. Ezek közül is a legnépszerűbbek a maximum entrópián és CRF-en alapulók. A felügyelt tanuló algoritmusok egyik előnytelen tulajdonsága, hogy nagy mennyiségű manuálisan annotált adatra van szükségük. Ezek közül a doménspecifikus tanító anyagot használók általában jobban teljesítenek, de egyes kutatók, mint Tomanek et al. [20] az általános nyelvi adatok használata mellett érvelnek.

3. Erőforrások és metrikák

Az elkészült módszer fejlesztése és kiértékelése céljából szükséges volt létrehozni egy megfelelő méretű etalon korpuszt, illetve meghatározni azokat a metrikákat, amik a kiértékelés alapját képezték. Ebben a fejezetben ismertetjük az etalon létrejöttének lépéseit, jellemző tulajdonságait, majd pedig bemutatjuk azon mértékeket, melyek a méréseink alapját képezték.

3.1. Az etalon korpusz

A korpusz egy szemészeti klinikai rekordjainak véletlenszerűen kiválasztott bekezdéseit tartalmazza, melyeket először automatikusan tokenekre és mondatokra bontottunk, majd az így kapott szövegeket manuálisan javítottuk és ellenőriztük. Az így kapott etalon a helyesen szegmentált bekezdéseken túl tartalmazza még azok eredeti formáját is. A tesztkorpusz mintegy 2300 mondatot tartalmaz, melyből 1200 az egyes algoritmusok kiértékeléséhez, míg a maradék azok optimalizálására került felhasználásra.

Mivel az orvosi rekordokból kinyert bekezdések zajosak, így azok szegmentálása előtt szükség volt egy normalizáló modul alkalmazására is. Ennek a szabályalapú komponensnek az alábbi hibákkal kellett megküzdenie:

1. duplán konvertált karakterek, mint pl. '>',
2. „írógépproblémák": az '1' és '0' gyakran 'l' és 'o' betűkként szerepeltek,
3. dátumok nem konvencionális használata pl. '2011.01.02.', vagy '06.07.12.',
4. központozási hibák pl. '1.23mg', Törőközegek.Fundus :ép.'.

Hogy teljesebb képet kapjunk az orvosi szövegek karakterisztikájáról, összevetettük az etalont a Szeged Korpusszal (SZK) [21]. Az összehasonlítás az alábbi jelentős különbségeket fedte föl:

1. A rövidítések aránya az általunk vizsgált klinikai szövegekben mintegy 2,68%, míg ez az általános nyelvi korpuszban kevesebb mint 0,01% volt.
2. A SZK mondatai szinte mindig (98,96%) mondatzáró írásjellel végződnek, míg ez az orvosi szövegek mondataiban csak az esetek 51,72%-ban igaz.

3. Hasonlóan az előzőekhez, a mondatkezdő nagybetűk használatának aránya is nagymértékű eltérést mutat: a klinikai rekordokban ez csupán 87,19% míg az általános nyelvi szövegekben 99,58%.
4. A tokenizálást érintő jelentős különbség még a numerikus adatokat tartalmazó mondatok aránya, mely a klinikai rekordokban 13,50%, míg a SZK esetében ez az arány elhanyagolható.

3.2. Kiértékelési módszerek

A szakirodalomban nincs egyetértés afelől, hogy milyen metrikát érdemes használni a mondatrabontás és tokenizálás feladataiban: a GT módszereket alkalmazók gyakran F-mértéket, pontosságot és fedést használnak, míg beszédfelismerési feladatok esetén ugyanerre pl. a NIST metrikát alkalmazzák. Sokszor a fedés, illetve pontosság használata esetén sem egyértelmű, hogy mik az osztályozandó entitások, és azok milyen kategóriákba kerülhetnek.

Írásunkban a Read et al. [1] által bemutatott módszernek egy módosított változatát használjuk. Így a szegmentálást egy egységes osztályozási problémaként értelmezzük, amiben minden karaktert, illetve a köztük lévő üres sztringeket egy-egy címkével illetünk aszerint, hogy az entitás két token határán áll-e, egy mondatot zár-e le vagy esetleg az előzőek egyike sem. Ezt a sémát használva az eredmények elemzéséhez a bevett fedés- és pontosságalapú mértékekre támaszkodunk. A kiértékelés során az F_β-értéket is kalkulálunk: míg a tokenizálás feladatában az általános F_1 vizsgálatát megfelelőnek találtuk, a mondatokra bontás esetén a pontosságot előnyben részesítve a $\beta = 0,5$-t találtuk optimálisnak. Az utóbbi döntés mögött az a megfontolás áll, hogy a nyelvtechnológiai feldolgozási lánc rákövetkező moduljai még képesek lehetnek két szét nem választott mondat helyes elemzésére, de fals mondattöredékek feldolgozása a hibák további keletkezését szolgálja.

4. A szegmentáló lánc

Ebben a fejezetben ismertetjük azt az összetett algoritmust, mely nagy pontossággal végzi a klinikai szövegek mondatokra bontását. Az alábbiakban bemutatott algoritmus első eleme egy olyan szabályalapú komponens, ami elsősorban a tokenizálásért felelős. Ennek leírása után ismertetjük még azokat módszereket is, melyek tovább növelik a szegmentáló lánc teljesítményét.

4.1. A baseline algoritmus

Eljárásunk első lépésként egy olyan szabályalapú modult használ, melynek célja, hogy tokenekre bontsa a bekezdések szövegeit. A komponens ezen működését itt nem részletezzük, mivel algoritmusa tokenizálási feladatokban jól ismert szabályokra támaszkodik. Ez a komponens a tokenizáláson túl magában foglalja még olyan mondatvégek felismerését is, melyekre a tokenhatárok megállapítása során lehetőség nyílik. Erre a következő esetekben van mód:

1. ha egy létrejött token mondatvégi írásjel, ami egy nem írásjelet tartalmazó token előtt szerepel,
2. vagy ha egy sor egy teljes dátumkifejezéssel vagy egy vizsgálati eredménnyel kezdődik.

Megvizsgálva a fenti eljárás eredményességét azt találtuk, hogy így a mondatvégek mindössze felét lehetséges felfedni, ami az algoritmus magas pontossága mellett is túl alacsony összesített teljesítmény. A hibák részletes elemzése megmutatta még, hogy a fel nem ismert tokenhatárok jelentős része egybeesik a nem azonosított mondathatárokkal, ami szükségessé teszi a pontra végződő tokenek osztályozását. Így tehát úgy döntöttünk, hogy egy olyan komponenssel egészítjük ki az algoritmust, mely képes megkülönböztetni a rövidítéseket a mondatvégi szavaktól.

4.2. Eredményesebb mondathatár-felismerés gépi tanulás használatával

Általános nyelvi szövegekben kétfajta indikátor létezik, amik mondathatárokat jelezhetnek. Ez egyik ilyen az írásjelek jelenléte, a másik pedig a nagybetűk használata. Esetünkben az írásjelek közül csak a pont igényel további vizsgálódást, hiszen ez esetben áll csak fenn többértelműség. Hasonlóan a kapitalizált szavak elemzésével is körültekintően kell eljárni, hiszen a tulajdonneveken kívül az orvosi szövegekben bizonyos rövidítések és latin szavak is tévesen nagy kezdőbetűvel vannak írva. A fentieken felül nehezítik még a feladatot az olyan mondathatárok, amiknél mindkét jellemző egyszerre hiányzik.

Az indikátorokra építve is lehet automatikus eljárásokat építeni anélkül, hogy doménspecifikus rövidítéslista vagy tulajdonnév-szótár a rendelkezésünkre állna. Ugyanis egy feldolgozó algoritmusnak elégséges megfelelő bizonyítékot találnia egy szó (w), és az őt követő pont ($\bullet$) szeparáltságára, ami pedig Kiss és Strunk algoritmusához [22] vezet. Így tehát a kollokációk azonosítására használt log-likelihood arány egy megfelelő módszer a feladat megközelítésére. Esetünkben ez a (3)-ban formalizálható, ami statisztikai tesztre épülve felhasznál egy null és egy alternatív hipotézist.

$$H_0 : P(\bullet|w) = p = P(\bullet|\neg w) \tag{1}$$

$$H_A : P(\bullet|w) = p_1 \neq p_2 = P(\bullet|\neg w) \tag{2}$$

$$log\lambda = -2log\frac{L(H_0)}{L(H_A)} \tag{3}$$

A (1) formula a *(szó, $\bullet$)* pár függetlenségét fejezi ki, míg (2) teljesülése esetén feltételezhetjük, hogy ezek együttállása nem csupán véletlenszerű, mivel rövidítést jelölnek. Kiss és Strunk kutatása megmutatta, hogy a (3)-ban számolt $log\lambda$ értékek eloszlása χ^2-tel aszimptotikus, így statisztikai tesztként is használható. Ezzel együtt azt is megállapították, hogy ennek a módszernek a pontossága önmagában alacsony, így szükséges további skálázó faktorok alkalmazása.

Kutatásunkban ezekre az eredményekre támaszkodva alkalmazzuk a $log\lambda$ kalkulust, viszont szemben az eredeti munkával egy inverz pontozási módszert használunk ($iscore = 1/log\lambda$). Tesszük ezt azért, mert nem célunk az összes orvosi rövidítés azonosítása, sőt éppen ellenkezőleg, csak azon párok fellelése, amikről nagy biztonsággal feltételezhetjük, hogy nem összetartozóak, így tehát nem rövidített szóalakok. A fejlesztés során szükségesnek találtuk még a skálázó faktorok adaptálását is, melyet az alábbiakban részletezünk.

Hasonlóan [22]-hoz, az első tényező a tokenek hosszára épülve (len) jutalmazza a rövideket és bünteti a hosszúakat. A faktor számítása során felhasználtuk még a korpusz általános jellemzőit: az optimalizációs adatokból kinyert és manuálisan ellenőrzött rövidítéslista elemeinek a 90%-a legfeljebb 3 hosszúságú, míg az ettől hosszabb rövidített tokenek csak elvétve fordulnak elő. Így formalizáltuk ezeket a megfigyeléseket a (4) tényezőben.

$$S_{length}(iscore) = iscore \cdot \exp\left(len/3 - 1\right) \tag{4}$$

Mint azt [13]-ben ismertettük, a HuMor tőtárát orvosi doménen használatos szavakkal bővítettük, így ennek elemzéseit is felhasználtuk az osztályozási feladatban. Mivel az elemző számos rövidítést is ismer, így erre a tudásra alapozva tovább szűrhetjük a mondatvégi tokenek listáját. Az (5) indikátorfüggvény a HuMor elemzései alapján jelez, hogy az adott szónak létezik-e rövidítésre visszavezethető felbontása. A lexikális tudás nagyobb biztonsági foka miatt, nagyobb súlyt társítottunk ehhez a faktorhoz, továbbá (6) úgy került kialakításra, hogy képes legyen ellensúlyozni a rövid mondatvégi szavak hibás osztályozását.

$$indicator_{morph}(w) = \begin{cases} 1 & \text{ha } w \text{ szó elemzései között nincsen rövidítés} \\ -1 & \text{ha } w\text{-nek van rövidítés elemzése} \\ 0 & \text{egyébként} \end{cases} \tag{5}$$

$$S_{morph}(iscore) = iscore \cdot \exp\left(indicator_{morph} \cdot len^2\right) \tag{6}$$

A harmadik és egyben utolsó tényező a kötőjelek használatára épül. Vizsgálataink során azt tapasztaltuk, hogy ezek jelenléte nem jellemző a rövidítésekben, viszont annál inkább előfordulhatnak az összetett szavak képzésekor. Ezt a megfigyelést formalizálva a szó hosszával arányos tényezőként készítettük (7)-et, melyben a $indicator_{hyphen}$ akkor és csak akkor vesz fel 1 értéket, ha a szó tartalmaz kötőjelet, egyéb esetben az értéke 0.

$$S_{hyphen}(iscore) = iscore \cdot \exp\left(indicator_{hyphen} \cdot len\right) \tag{7}$$

A fentiek módosítók használatával számoljuk az összesített pontozást, amit (8) mutat be. Az $sscore$-t minden ponttal végződő tokenre kalkulálja az algoritmus, majd összeveti ezt egy empirikusan meghatározott küszöbértékkel ($< 1,5$), mely alapján rövidítésnek azonosítható egy entitás.

$$sscore = S_{hyphen} \circ S_{morph} \circ S_{length}(iscore) \tag{8}$$

4.3. További kapitalizáción alapuló szabályok

Munkánkban létrehoztunk még egy olyan komponenst is, mely szavak kapitalizációjára támaszkodik. Ez a modul is épít a HuMorra: ha egy szó analízisei között nem szerepel egy tulajdonnévi elemzés sem, és a szó nagy kezdőbetűvel van írva, akkor a szóban forgó entitás mondatkezdő jelöltté válik. Ezek további szűrésére is szükség van, mivel fennáll még a veszélye annak, hogy egy több tagból álló tulajdonnév egyik elemével van dolgunk. Így a kontextusok figyelembevételével, csak azokat a szavak kerülnek a mondatkezdő osztályba, amik biztosan nem tulajdonnevek.

5. Eredmények

Az algoritmus egészének teljesítményére egy mutató az összesített pontosság. Az 1. táblázatban közreadjuk az előfeldolgozott és a szegmentáló metódusok eredményeinek megfelelő értékeit. Itt a pontosság értékek magas volta azzal magyarázható, hogy a kiértékelő módszer a leggyakoribb jelenséget (*nincs módosítás*) egyformán jutalmazza a legnehezebbekkel. Közelebbi képet kapunk a komponensek egyenkénti teljesítményéről a 2. táblázatban, amiben a hibarátájuk csökkenését prezentáljuk.

1. táblázat. Az egyes feldolgozási fázisok összesített pontossága

	Összesített pontosság
Előfeldolgozott adat	97,55%
Baseline algoritmus	99,11%
Teljes lánc	99,74%

2. táblázat. Az egyes rendszerek hibaarányának csökkenése a baselinehoz viszonyítva

	Hibaráta csökkenés
$(w, \bullet)$ párok osztályozásával	58,62%
Kapitalizáción alapuló szabályokkal	9,25%
A teljes lánc	65,50%

Tüzetesebben megvizsgálva az egyes modulok teljesítményét a hagyományos pontosság, fedés és F-értékeket is számolunk. A mondathatárok azonosítását tekintve a 3. táblázat értékei jelentős teljesítménynövekedésről számolnak a fedést illetően, míg pontossági értékek csak kis mértékben csökkennek.

Eredményeinket érdemes tanulmányozni más magyar nyelvre szabadon elérhető szegmentáló eszközök teljesítményének fényében is. Vizsgálatunkban a

3. táblázat. Az egyes mondatrabontó modulok eredményességének vizsgálata

	Pontosság (P)	Fedés (R)	$F_{0,5}$
Baseline	96,57%	50,26%	81,54%
$(w, \bullet)$ párok osztályozásával	95,19%	78,19%	91,22%
Kapitalizáción alapuló szabályokkal	94,60%	71,56%	88,88%
A teljes lánc	93,28%	86,73%	91,89%

teszthalmaz adatain kiértékeltük a `magyarlanc` megfelelő modulját, a huntoken eszközt, az OpenNLP[1] mondatrabontó komponensét, illetve Punkt nyelvfüggetlen rendszert. A huntoken rendszer a működéséhez rövidítéslistákat használ, mely lehetőséget adott működésének testreszabásához. Így vizsgálatunk kiterjedt az általános tokenizáló (HTG) teljesítményén túl, egy orvosi rövidítésekkel adaptált (HTM) verziójára is. Mivel az OpenNLP FGT algoritmusokat használ mondatvégek azonosítására, így ehhez tanítóanyagként a Szeged Korpuszt mondatait használtuk.

4. táblázat. Szabadon elérhető mondatrabontó alkalmazások teljesítményének kiértékelése

	Pontosság (P)	Fedés (R)	$F_{0,5}$
`magyarlanc`	72,59%	77,68%	73,55%
HTG	44,73%	49,23%	45,56%
HTM	43,19%	42,09%	42,97%
Punkt	58,78%	45,66%	55,59%
OpenNLP	52,10%	96,30%	57,37%
A hibrid lánc	93,28%	86,73%	91,89%

A 4. táblázat adatai azt sugallják, hogy a zajos orvosi szövegeken az általános nyelvhasználatra optimalizált szoftverek sikertelennek bizonyulnak. Bár az OpenNLP kiemelkedő fedéssel rendelkezik, de cserébe a mondatok majd felét hibásan vágja szét, ami végeredményben alacsony F-pontot eredményez. Robusztus teljesítményt mutat még a `magyarlanc`, mely eredmény a jól felépített, doménfüggetlen szabályok használatának köszönhető. Ezekkel szemben a huntoken egyes változatai nyújtják a legalacsonyabb pontosságot és F-pontokat is. A Punkt eredményeit vizsgálva azt találjuk, hogy a felügyelet nélküli tanuló algoritmus doménadaptációja mintegy kétszeres teljesítménynövekedést eredményezett.

Bár munkánkban főleg a mondatok szegmentálására koncentrálunk, de vizsgáltuk még a tokenizáló rendszerek pontosságát is. Az elvégzett mérések (5. táblázat) összhangban állnak azzal a feltételezésünkkel, hogy a baseline algoritmus által fel nem fedezett tokenhatárok jelentős része egyben mondathatár is.

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

5. táblázat. A tokenizálás feladatára vonatkozó eredmények

	Pontosság (P)	Fedés (R)	F_1
Baseline	99,74%	74,94%	85,58%
A teljes lánc	98,54%	95,32%	96,90%

6. Összegzés

Írásunkban ismertettünk egy hibrid algoritmust, mely kiemelkedő eredményességgel képes mondat- és tokenhatárok azonosítására klinikai rekordok bekezdéseiben. Vizsgálatunk célja elsősorban a mondatvégek helyes detektálása volt, melyhez egy három lépésből álló eljárást készítettünk. A készített feldolgozási lánc szabályalapú komponensek mellett felügyelet nélküli gépi tanulásra is támaszkodik. Az algoritmus első lépésben mintaillesztés használatával elvégzi az alapszintű tokenizálást, majd ennek eredményében az egyes *(szó, •)* párok eloszlását figyelembe véve azonosítja a mondathatárok nagy részét, melyet az utolsó szabályalapú komponens tovább finomít. A bemutatott algoritmus különlegessége, hogy a határkeresési feladatokhoz egy morfológiai elemző tudását is sikerrel használja.

A létrehozott rendszer teljesítménye, összehasonlítva más szabadon elérhető szoftverekkel szemben is, kiemelkedően magas. Vizsgálatunk megmutatta, hogy a létrejött hibrid algoritmuson kívül nincsen más olyan szabadon hozzáférhető eszköz, mely hasonló eredményességgel végezné orvosi szövegeken a szegmentálás feladatát.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP – 4.2.1.B – 11/2/KMR-2011-0002 és TÁMOP – 4.2.2/B – 10/1–2010–0014 pályázatok támogatásával készült.

Hivatkozások

1. Read, J., Dridan, R., Oepen, S., Solberg, L.J.: Sentence Boundary Detection: A Long Solved Problem? In: 24th International Conference on Computational Linguistics (Coling 2012). India. (2012)
2. Riley, M.D.: Some applications of tree-based modelling to speech and language. In: Proceedings of the Workshop on Speech and Natural Language, Association for Computational Linguistics (1989) 339–352
3. Palmer, D.D., Hearst, M.A.: Adaptive sentence boundary disambiguation. In: Proceedings of the fourth conference on Applied natural language processing, Association for Computational Linguistics (1994) 78–83
4. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the fifth conference on Applied natural language processing, Association for Computational Linguistics (1997) 16–19

5. Gillick, D.: Sentence boundary detection and the problem with the US. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Association for Computational Linguistics (2009) 241–244
6. Mikheev, A.: Periods, capitalized words, etc. Computational Linguistics **28**(3) (2002) 289–318
7. Mikheev, A.: Tagging sentence boundaries. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Association for Computational Linguistics (2000) 264–271
8. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of Language Resources and Evaluation Conference. (2004)
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of Recent Advances in Natural Language Processing 2013, Hissar, Bulgaria, Association for Computational Linguistics (2013) 763–771
10. Kumar, A.: Monk project: Architecture overview. In: Proceedings of JCDL 2009 Workshop: Integrating Digital Library Content with Computational Tools and Services. (2009)
11. Siklósi, B., Orosz, Gy., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for hungarian clinical records. In De Pauw, G., De Schryver, G.M., Forcada, M.L., M Tyers, F., Waiganjo Wagacha, P., eds.: 8th SaLTMiL Workshop on Creation and use of basic lexical resources for lessresourced languages. (2012) 29.–34.
12. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in hungarian medical documents. In Dediu, A.H., Martín-Vide, C., Mitkov, R., Truthe, B., eds.: Statistical Language and Speech Processing. Volume 7978 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 248–259
13. Orosz, Gy., Novák, A., Prószéky, G.: Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése. In: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 159–169
14. Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., Denny, J.C.: Medex: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association **17**(1) (2010) 19–24
15. Apostolova, E., Channin, D.S., Demner-Fushman, D., Furst, J., Lytinen, S., Raicu, D.: Automatic segmentation of clinical texts. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE (2009) 5905–5908
16. Cho, P.S., Taira, R.K., Kangarloo, H.: Text boundary detection of medical reports. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2002) 998
17. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Schuler, K.K., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5) (2010) 507–513
18. Taira, R.K., Soderland, S.G., Jakobovits, R.M.: Automatic structuring of radiology free-text reports. Radiographics **21**(1) (2001) 237–245
19. Tomanek, K., Wermter, J., Hahn, U.: Sentence and token splitting based on conditional random fields. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. (2007) 49–57

20. Tomanek, K., Wermter, J., Hahn, U.: A reappraisal of sentence and token splitting for life sciences documents. Studies in Health Technology and Informatics **129**(Pt 1) (2006) 524–528
21. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora. (2004) 19–23
22. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Computational Linguistics **32**(4) (2006) 485–525

A magyar beteg

Siklósi Borbála[1], Novák Attila[1,2]

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

Kivonat A klinikai szövegek feldolgozása aktív kutatási terület, melynek során az egyik legnagyobb kihívás az ilyen szövegek azon sajátosságainak a kezelése, amelyek tekintetében ezek az általános szövegektől jelentősen eltérnek. Ezek között szerepel többek között a sok szakszó és rövidítés, a szinte csak rövidítésekből és numerikus adatokból álló „mondatok", valamint a jelentős számú helyesírási és központozási hiba, amelyből többek között a mondathatárok felismerésének rendkívül nehéz volta is következik.
Cikkünkben bemutatjuk a rendelkezésünkre álló magyar klinikai korpusz jellemzőit, különös tekintettel az előbb említett tényezőkre, összevetve azt egy általános tartalmú magyar szövegeket tartalmazó korpusszal. A szövegek felszíni tulajdonságai mellett összehasonlításokat végeztünk a leggyakoribb szavak disztribúciós szemantikai viselkedése alapján is, melynek során a jelentésbeli különbségek is kimutathatóak a különböző korpuszok között.

1. Bevezetés

A klinikai dokumentumok olyan szövegek, melyek kórházi körülmények között, mindennapi eseteket dokumentálva a kezelések során jönnek létre. Minőségük tehát nem összehasonlítható az elsősorban angol nyelven szintén aktívan vizsgált orvosi-biológiai szakirodalom nyelvezetével, amelyek többszörös ellenőrzésen keresztülmenve, szigorú nyelvi szabályok betartása mellett keletkeznek [1,2]. A klinikai orvosi szövegek ezzel szemben sietve, minden nyelvi segédeszköz, vagy emberi ellenőrzés nélkül, általában strukturálatlan formában jönnek létre. Jellemző továbbá, hogy keletkezésük során ezeknek a dokumentumoknak a címzettje általában az azt leíró orvos maga, tehát az eredeti célját nem befolyásolja a sajátos nyelvezet, egyedi rövidítések, utalások használata. Ezek a dokumentumok azonban nagyon sok olyan információt és tudást tartalmaznak, amelyeket ezen az elsődleges célon túl, az orvostudomány több területén alkalmazni lehetne. Ehhez arra lenne szükség, hogy a szövegekben leírt tényállásokat olyan formára hozzuk, amely lehetővé teszi ezeknek az információknak a hatékony kinyerését.

Több kísérlet született már a természetes nyelvű szövegek feldolgozásához általánosan használt eszközök orvosi szövegeken való alkalmazására, azonban ezek teljesítménye általában messze elmarad attól a szinttől, amit általános szövegeken elérnek. Ahhoz, hogy a már bevált módszerek, vagy azoknak egy része

adaptálható legyen az orvosi szövegekre, ismernünk kell ez utóbbinak a jellemzőit, illetve az általános szövegektől való főbb eltéréseket.

Ehhez több vizsgálatot végeztünk. A korpusz alapján először a felszíni alakok statisztikai eloszlását, majd ugyanezek egy feldolgozási lépéssel későbbi szintű (szótő, szófaj, névelemek, rövidítések) előfordulását vizsgáltuk, összehasonlítva a kapott mintákat az általános korpuszból kinyert adatokkal. Általános szövegként a Szeged Korpuszt használtuk. Jól elkülöníthetővé váltak a két szövegtípusban jellemzően előforduló nyelvi szerkezetek. Az eredmények elemzése során kimutathatóak azok a szerkezeti bizonytalanságok, amelyek miatt a klinikai szövegek jóval nehezebben értelmezhetőek az általános szövegeknél. Ilyen jellemzők nemcsak a rengeteg szakkifejezés jelenléte, hanem a szövegek gyakran rendkívül pongyola megformálása és az azonos fogalmak jelölésére konkrétan használt írott alakok rendkívüli változatossága is.

Természetesen a lexikai alakok vizsgálata során azok összehasonlítása nem vizsgálható érdemben, hiszen a szakkifejezések előfordulási aránya nyilvánvalóan nagyobb a szakszövegekben. A klinikai dokumentumokra azonban jellemző, hogy az esetleírásoknál, különösen a panaszok felvétele során egészen hétköznapi történetek leírása is szerepel. Ennek a kevert orvosi nyelvnek a statisztikai jellemzői is felismerhetők a korpusz önmagában való vizsgálata során.

Tanulmányunk célja a részletes statisztikai vizsgálatok alapján azon jelenségek bemutatása, amik igazolják az orvosi-klinikai szövegek feldolgozásának nehézségeit, illetve irányadók lehetnek a különböző eszközök fejlesztése során, melyek paraméterei így a specifikus problémákhoz hangolhatóak.

2. Korpuszok

Vizsgálataink során általános nyelvezetű korpuszként a Szeged Korpusz 2-t használtuk. Orvosi korpuszként pedig a rendelkezésünkre álló nyers klinikai dokumentumokat. Ezek 29 különböző osztályról származó kezelési lapok, zárójelentések, egyéb klinikai dokumentumok. A klinikai korpuszon belül külön foglalkoztunk a szemészeti dokumentumokkal, hiszen azok feldolgozottsági állapota a folyamatban lévő kutatásaink miatt sokkal előrébb tart, a már meglévő eszközeink adaptálása a többi osztály dokumentumaira még nem valósult meg. Így a következő három domént vetettük alá összehasonlító vizsgálatainknak: általános szövegek a Szeged Korpusz alapján (SZEG), vegyes orvosi szövegek (MED), illetve szemészeti szövegek (SZEM). A korpuszok méretére vonatkozó részletes adatok az 1. táblázatban találhatóak. A Szeged Korpuszra vonatkozó adatok itt és a továbbiakban is [3]-ból származnak.

A szófajok eloszlását illetően eltérő a két fő domén (SZEG és MED) összetétele. Míg a Szeged Korpuszban a leggyakoribb szófajok közül az első három a főnév, ige, melléknév, addig az orvosi szövegekben a főnevek mellett a melléknevek és a számnevek a leggyakoribbak, míg az igék száma az utóbbi két, közel azonos mennyiségben előforduló szófajhoz képest csak harmadannyiszor szerepel. Jelentős különbség még, hogy az orvosi szövegekben a névelők, kötőszavak és névmások is a rangsor második felében helyezkednek el. Ezek az előfordulási

1. táblázat: A három vizsgált korpusz mérete (tokenek és mondatok száma), az őket jellemző átlagos mondathossz

	tokenszám	mondatszám	átlagos mondathossz
Orvosi korpusz (MED)	7 119 841	734 666	9,69
Szemészeti korpusz (SZEM)	334 546	34 432	9,7
Szeged Korpusz (SZEG)	1 194 348	70 990	16,82

arányok nem is meglepőek, hiszen az orvosi feljegyzések legnagyobb része arról szól, hogy egy állapotot ír le (*valami valamilyen* (FN, MN)), vagy valamilyen vizsgálat eredményét (*valami valamennyi* (FN, SZN)). Az orvosi szövegekben előforduló számnevek túlnyomó része numerikus adat. A részletes szófaji eloszlásokat tartalmazza a 2. táblázat.

2. táblázat: A Szeged Korpusz és az Orvosi korpusz tokenjeinek szófaji eloszlása, illetve rangsora

	FN	MN	SZN	IGE	HAT	NM	DET	NU	KOT
MED	43,02%	13,87%	12,33%	3,88%	2,47%	2,21%	2,12%	1,03%	0,87%
SZEG	21,96%	9,48%	2,46%	9,55%	7,60%	3,85%	9,39%	1,24%	5,58%

	FN	MN	SZN	IGE	HAT	NM	DET	NU	KOT
MED	1	2	3	4	5	6	7	8	9
SZEG	1	3	8	2	5	7	4	9	6

3. Helyesírási különbségek

A klinikai dokumentumok jellegzetessége, hogy gyorsan, utólagos lektorálás, ellenőrzés, illetve automatikus segédeszközök (pl. helyesírás-ellenőrző) nélkül készülnek, ezért a leírás során keletkezett hibák száma igen nagy, valamint sokféle lehet [4]. Így nem csupán a magyar nyelv nehézségeiből eredő problémák jelennek meg, hanem sok olyan hiba is felmerült a szövegekben, melyek a szakterület sajátosságaiból erednek. A legjellemzőbb hibák az alábbiak:

- elgépelés, félreütés, betűcserék,
- központozás hiányosságai (pl. mondathatárok jelöletlensége) és rossz használata (pl. betűközök elhagyása az írásjelek körül, illetve a szavak között),
- nyelvtani hibák,
- mondattöredékek,
- a szakkifejezések latin és magyar helyesírással is, de gyakran a kettő valamilyen keverékeként fordulnak elő a szövegekben (pl. *tensio/tenzio/tensió/tenzió*); külön nehézséget jelent, hogy bár ezeknek a szavaknak a helyesírása

szabályozott, az orvosi szokások rendkívül változatosak, és időnként még a szakértőknek is problémát jelent az ilyen szavak helyességének megítélése,
– szakterületre jellemző és sokszor teljesen ad hoc rövidítések, amelyeknek nagy része nem felel meg a rövidítések írására vonatkozó helyesírási és központozási szabályoknak

A fenti hibajelenségek mindegyikére jellemző továbbá, hogy orvosonként, vagy akár a szövegeket lejegyző asszisztensenként is változóak a jellemző hibák. Így elképzelhető olyan helyzet, hogy egy adott szót az egyik dokumentum esetén javítani kell annak hibás volta miatt, egy másik dokumentumban azonban ugyanaz a szóalak egy sajátos rövidítés, melynek értelmezése nem egyezik meg a csupán elírt szó javításával.

A Szeged Korpusszal összehasonlítva két fő különbséget állapíthatunk meg. Az egyik a rövidítések aránya: míg a Szeged Korpuszban a rövidítések a tokenek 0,08%-át teszik ki, addig az általunk vizsgált anyag 7,15%-a rövidítés [5], tehát a rövidítések gyakorisága két nagyságrenddel nagyobb. Ezt a számítást az orvosi szövegekre egy 15 278 token méretű részkorpusz alapján végeztük. Szintén ebből számítottuk a helyesírási hibákat, amiket kézzel jelöltünk meg az orvosi szövegeknek ebben a részhalmazában. Ezért az ebben előforduló helyesírási hibák típusairól részletesebb statisztikát is tudtunk készíteni, melyet a 3. táblázat tartalmaz. A helyesírási hibák aránya az orvosi korpuszban 8,44%, ezzel szemben a Szeged Korpuszban csupán 0,27%. Ezen belül is az iskolai fogalmazásokat tartalmazó részkorpuszban is mindössze 0,87%, tehát tízszer kevesebb helyesírási hibát ejtettek a Szeged Korpuszban szereplő fogalmazásokat író iskolás tanulók, mint a klinikai szövegeket író orvosok. Az orvosi szövegek esetén ezek a hibák az esetek felében ponthibák (leginkább a pont hiánya a rövidítések végén). Az egybeírási és különírási hibák pedig közel azonos mértékben fordulnak elő, összesen a hibák 10%-át teszik ki. Amellett, hogy a rövidítések végéről gyakran hiányzik a pont, az orvosi szövegekre egyébként is jellemző a központozási hibák magas aránya. Míg a Szeged Korpuszban csak a mondatok 1,04%-a nem végződik pontra (címek), addig az orvosi dokumentumokban ez az arány 48,28%. Hasonló problémák vannak a mondatkezdő nagybetűhasználattal: míg a Szeged Korpuszban csak a mondatok 0,42%-a nem kezdődik nagybetűvel, addig az orvosi korpuszban a mondatok 12,81%-a. Ez teszi a mondatokra bontás látszólag triviális feladatát is rendkívül nehézzé [6].

3. táblázat: Az orvosi szövegek egy részkorpuszában előforduló helyesírási hibák típusai

	hibás	ponthiba	egybeírás	különírás	egyéb
Szeged Korpusz	0,27%	-	-	-	-
Szeged Korpusz – iskolás	0,87%	-	-	-	-
Orvosi korpusz	8,44%	46,55%	5,66%	5,59%	42,2%

4. Szemantikai különbségek

A szófaji és nyelvhasználati különbségek mellett az általános és az orvosi szövegek között gyakran jelentős eltérés mutatkozik meg azoknak a szavaknak a jelentésében is, amelyek mindkét korpuszban előfordulnak, tehát az egyes szavak szemantikája mást fed le a különböző korpuszokban. Ezt a jelenséget a disztribúciós szemantika módszerével vizsgáltuk. A disztribúciós szemantika lényege, hogy a szemantikailag hasonló szavak hasonló környezetben fordulnak elő. Tehát két szó jelentésének hasonlósága meghatározható a környezetük hasonlósága alapján. A szavak környezetét olyan jellemzőhalmazokkal reprezentáltuk, ahol minden jellemző egy relációból *(r)* és az adott reláció által meghatározott szóból *(w')* áll. Ezek a relációk más alkalmazásokban általában függőségi relációk, azonban a klinikai szövegekre ilyen elemzés a zajos mivoltuk miatt nem végezhető el kellően jó eredménnyel. [7] szintén klinikai szövegekre alkalmazva csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt. Mivel a mi esetünkben a morfológiai elemzés is rendelkezésre állt, ezért a következő jellemzőket vettük figyelembe:

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

Minden egyes jellemzőhöz meghatároztuk a korpuszbeli gyakoriságát. Ezekből a gyakoriságokból határozható meg a *(w,r,w')* hármas információtartalma *(I(w,r,w'))* maximum likelihood becsléssel. Ezután a két szó (*w* és *w'*) közötti hasonlóságot a következő metrikával számoltuk [8] alapján:

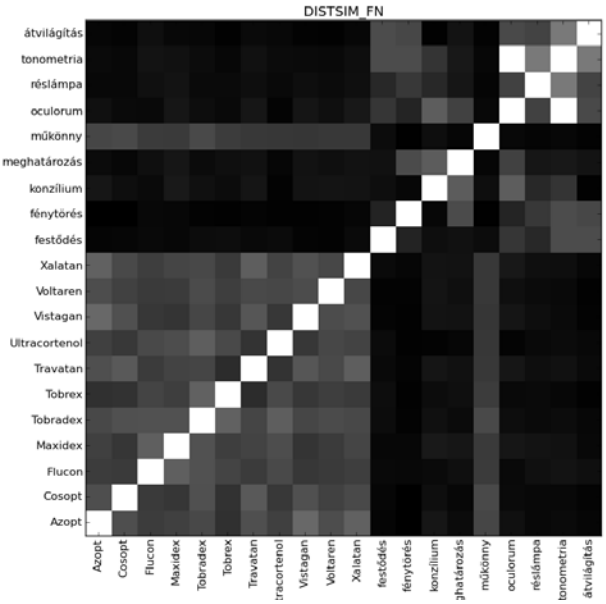
$$\frac{\sum_{(r,w)\in T(w_1)\bigcap T(w_2)}(I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w)+\sum_{(r,w)\in T(w_2)} I(w_2,r,w)},$$

ahol *T(w)* azoknak az *(r,w')* pároknak a halmaza, ahol az *I(w,r,w')* pozitív. Ennek a metrikának a használatával korpuszonként kiszámoltuk a leggyakoribb főnevekre, igékre és melléknevekre a páronkénti disztribúciós hasonlóságukat.

4.1. A szemészeti korpusz disztribúciós szemantikája

A gyakori főnevek vizsgálata során olyan szópárokat kerestünk, melyeknél jól kimutatható a más főnevekhez való viszonyuk. Az 1. ábrán egy ilyen részlet látható. A világosabb mezők jelzik az erősebb szemantikai kapcsolatot az adott két szó között. Az ábrán jól elkülönülő szemantikai terek láthatóak. Például a különböző szemcseppek nevei és a *műkönny* kifejezés egy behatárolható csoportot

alkotnak. Ezek fölé rendelhető a *szemcsepp* fogalom. Hasonlóan, az egyes szemészeti vizsgálatok is egy csoportba kerültek *(átvilágítás, tonometria, réslámpa)*, illetve az ezek által vizsgált jelenségek *(fénytörés, festődés)*.



1. ábra: A szemészeti korpusz leggyakoribb főneveinek hasonlósági mátrixa

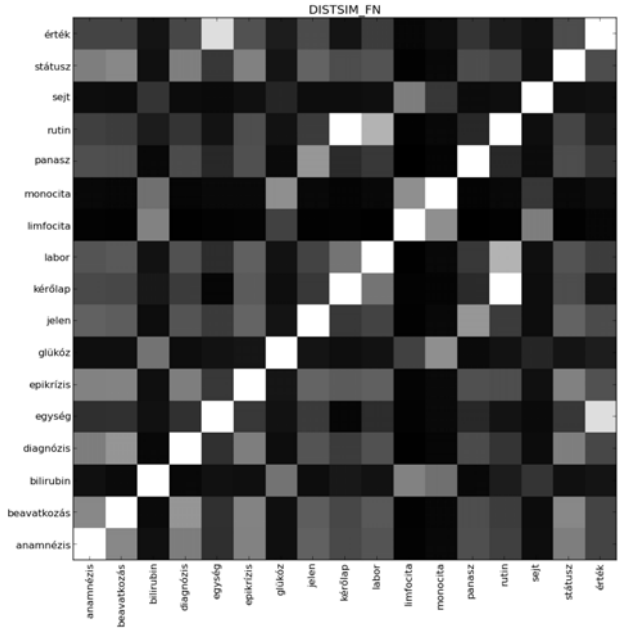
Az igei eloszlásra vonatkozóan is meghatározhatóak a szemantikai együttállások. Így a szemészeti korpusz esetén releváns csoportot alkotnak a *fáj* igéhez tartozó, hozzá hasonló kifejezések: *könnyezik, szúr, viszket, beragad.* Az orvosi korpusz eredményeinél látni fogjuk, hogy a *fáj* igéhez tartozó igék a *köhög, érez, romlik.*

A hasonlóságok kiértékelése során sok esetben nem tudtuk megítélni az egyes szakkifejezések közötti kapcsolat helyességét. A melléknevek esetén olyan hasonlóságokat találtunk, mint a *szélű* és a *határú*, valamint a *bal* és a *jobb* közötti kapcsolat, amelyek mindenképpen helytállóak. Az utóbbi párral kapcsolatban megjegyzendő, hogy a szemészeti szövegekben a *bal szem* és a *jobb szem* vizsgálatai miatt ezek kapcsolata sokkal erősebb és sokkal jobban elkülönülő csoportot alkotnak, mint az *alsó*, vagy *felső* mellékneveké, amik szintén az irányultságot jelzik. Az általános orvosi szövegekben a négy irány már egy csoportot alkot. A szinonímák és antonímák mellett a módszer kollokációkat is kimutat, pl. a *széli* és a *vesszős* hasonló disztribúciója abból adódik, hogy ezek leginkább a *(vaskos) széli vesszős homály* kifejezésben szerepelnek együtt.

4.2. Az orvosi szövegek disztribúciós szemantikája

Az általános orvosi szövegekből álló korpusz tartalma sokszínűbb, mint a szemészeti részkorpusz, ezért a szemantikai csoportok sem annyira kifinomultak,

mint egyetlen szűk domén esetén. Az azonban itt is megállapítható, hogy a létrejött relációk, illetve szemantikai csoportok helytállóak és relevánsak. A 2. ábrán szintén a leggyakoribb, legnagyobb hasonlóságot mutató főnevek szemantikai mátrixa látható. Az ábrán is élesen kiugrik a *limfocita–monocita* szópár, illetve a hozzájuk kapcsolódó *bilirubin*, *glükóz* és *sejt* szavak, melyekkel együtt élesen elhatárolódnak a többi fogalomtól. Hasonlóan jól elhatárolódnak az orvosi feljegyzések egyes részeit jelölő kifejezések: *anamnézis*, *diagnózis*, *epikrízis*, *státusz*. Ezekkel kapcsolatban látszik az, hogy bár a szemészeti dokumentumokban is ugyanezek a részek találhatóak meg, ott nem jelentek meg a leggyakoribb és legerősebb összefüggést mutató csoportok között (természetesen a szemantikai viselkedésük, így a hasonlóságuk ott is fennáll). A vegyes orvosi szövegekben azonban ez a csoport a kevert domén fölött hangsúlyosabb összetartozást mutat.



2. ábra: Az orvosi korpusz leggyakoribb főneveinek hasonlósági mátrixa

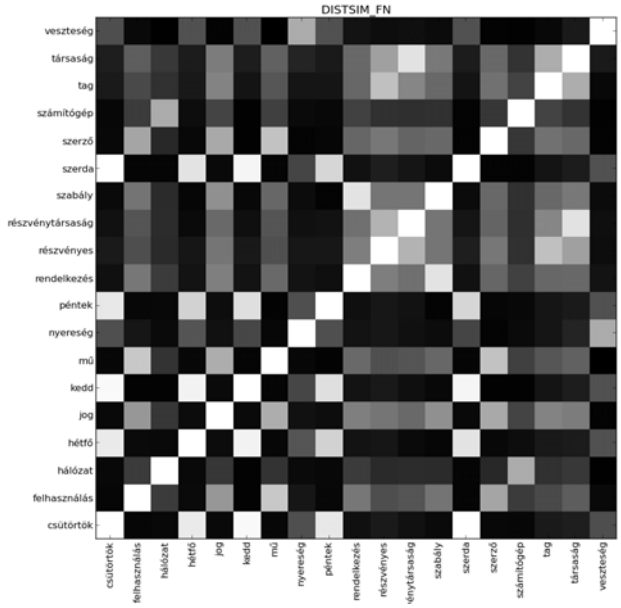
Az orvosi korpuszban vizsgált gyakori igék között olyan csoportok jöttek létre, mint a *mutat*, *igazol*, *látszik*, *ábrázolódik*. Természetesen nem csak az összetartozásnak van jelentősége (igaz ez mindhárom domén esetén mindegyik szófajra vonatkozóan), hanem az elhatárolódásnak is. Így az igék között az *ábrázolódik* és az *elhagy* jó példa arra, hogy ezek szemantikai viselkedése között nincsen hasonlóság.

A melléknevek esetén a már fent említett irányultsági csoportok emelhetőek ki, itt már mind a négy irányra vonatkozóan, illetve megjelenik a szakterületre vonatkozó melléknevek csoportja (*szakápolói*, *neurológiai*, *pszichiátriai*).

4.3. A Szeged Korpusz disztribúciós szemantikája

Az előző két doménhez képest nagy eltérést találunk az általános szövegeket tartalmazó korpuszban.

Bár a Szeged Korpusz a témák sokkal szélesebb körét öleli fel, mint az orvosi korpusz együttesen, vagy különösen mint egy adott orvosi szakterülethez tartozó szövegek, a módszer mégis kiemel a vegyes szöveghalmazból is olyan szemantikai csoportokat, amelyeken belül előforduló szavak erős tematikus összefüggést mutatnak. A főnevekre vonatkozó 3. ábrán jól látszik, hogy egy szemantikai csoportba kerültek a leggyakoribb főnevekre vizsgálva a *részvénytársaság*, *társaság*, *részvényes*, és *tag* kifejezések, illetve a *mű, szerző, felhasználás*, valamint a *jog, rendelkezés* és *szabály* szócsoportok. Nyilvánvalóan a korpuszban további, ebbe a körbe tartozó szavak is megjelennek, azonban az algoritmus igen nagy számításigénye miatt mindegyik esetben csak a leggyakoribb 100 szóra végeztük el a vizsgálatot. Természetesen olyan általánosan gyakori szavakból álló csoportok is felfedezhetőek, mint a hétköznapok nevei, azok igen erős hasonlósága alapján.



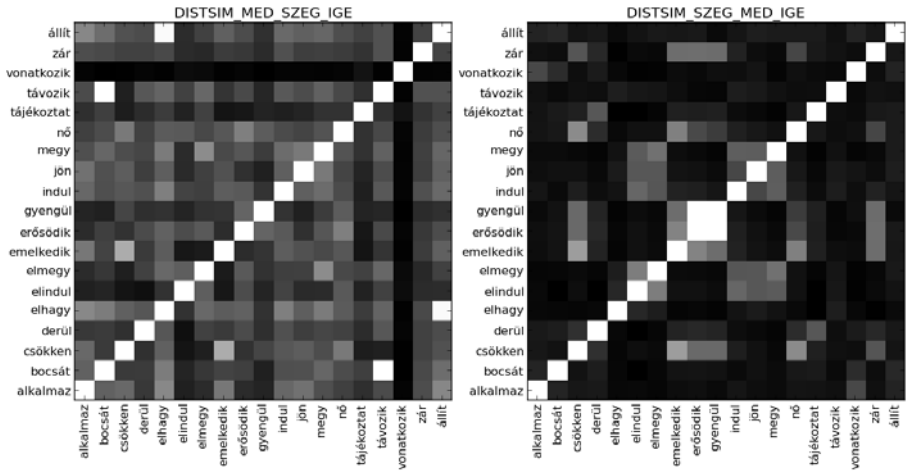
3. ábra: A Szeged Korpusz leggyakoribb főneveinek hasonlósági mátrixa

Az igékre vonatkozóan a létrejött szemantikai csoportok mellett további finomításokat tapasztalhatunk. Míg a *megy–indul*, illetve az *elmegy–elindul* szópárok külön-külön nagyon hasonlóak, addig a két páros egymáshoz való hasonlósága ennél kisebb, amit az igekötős alakok más jellegű viselkedése jól magyaráz.

A melléknevek esetén is kialakultak a fentiekhez hasonló csoportok, azonban ezek az általános mellékneveket tartalmazzák (pl. *tavalyi–idei*).

4.4. A korpuszok összehasonlítása

Amellett, hogy külön-külön megvizsgáltuk az egyes szófajok leggyakoribb példányait, összehasonlító elemzést is végeztünk. Ehhez először létrehoztuk a korpuszok leggyakoribb szavainak metszetét (szófajonként), majd az ebben a listában szereplő szavak disztribúciós hasonlóságát megmértük mindkét, az összehasonlításban részt vevő korpusz esetén. Az összehasonlítások során az általános orvosi és a szemészeti korpuszt vizsgáltuk a Szeged Korpusszal szemben. Általánosságban elmondható, hogy az általános szövegekben kiemelkedő hasonlóságot mutató szavak (például hónapnevek) hasonlósága megmaradt az orvosi szövegek esetén is, azonban a kontraszt az egyes csoportok között kisebbnek bizonyult. A másik általános jelenség, hogy az orvosi szövegekben egyes szavakhoz új jelentéscsoportok jelentek meg, melyek az általános témájú szövegekre nem jellemzőek. A főnevek között ilyen például az *időpont*, ami az orvosi korpuszokban sokkal közelebb kerül például a hét napjainak megnevezéséhez, hiszen a klinikai események során az időpontnak leginkább abban van szerepe, hogy melyik napon esedékes egy vizsgálat.

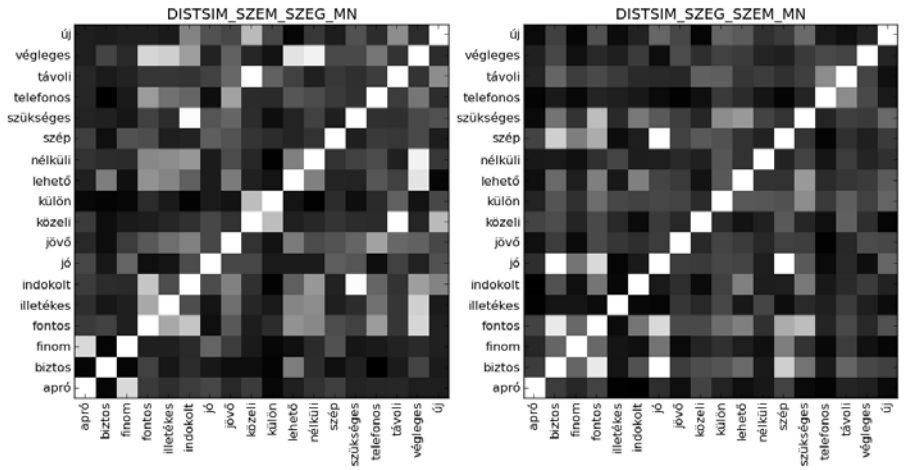


4. ábra: A leggyakoribb igék hasonlósági mátrixai az orvosi korpusz és a Szeged Korpusz alapján

Az igék esetében még jelentősebb különbségeket láthatunk, amit a 4. ábra illusztrál. A *távozik* és a *bocsát* szavak összefüggése az orvosi korpuszban nyilvánvaló (az *otthonába bocsát* kifejezés miatt), ami az ábrán is kiugró értékként jelenik meg, azonban a Szeged Korpuszban vizsgálva ugyanez a két szó egészen távoli. Szintén jól látszik, hogy míg az általános nyelvezetű szövegben az *emelkedik*, *erősödik*, *gyengül*, *csökken*, *nő* szavak egy jelentéskörbe tartoznak, addig az orvosi szövegek esetén ezeknek a viselkedése eltérő. Különbségként látszik még az *állít* és az *elhagy* szavak hasonlósága. Az orvosi korpuszban ezek nagyon hasonló viselkedésűek, mindkettő a gyógyszeres kezelésekkel kapcsolatos *(gyógyszer*

elhagyása, gyógyszer adagolásának beállítása). A Szeged Korpuszban ezek között semmilyen hasonlóság nem látszik.

A melléknevek esetén szintén látszanak olyan különbségek a szemészeti és az általános korpusz között, hogy míg az elsőben a *fontos* és az *indokolt* szavak lettek hasonlóak, amikhez viszont a *jó* egyáltalán nem kapcsolódik, addig az általános szövegekben a *fontos*, a *biztos*, és a *jó* tartoznak egy jelentéskörbe. További jelentésbeli eltolódások láthatóak az 5. ábrán.



5. ábra: A leggyakoribb melléknevek hasonlósági mátrixai a szemészeti korpusz és a Szeged Korpusz alapján

5. Konklúzió

Cikkünkben bemutattuk három korpusz összehasonlítását néhány fő statisztikai jellemzőjük alapján. Látható, hogy a klinikai dokumentumokból álló korpusz szavainak szófaji eloszlása és helyessége jelentősen eltér az általános szövegeket tartalmazó korpusztól, ezért az utóbbiakra széles körben elfogadott megállapítások, illetve az ezekre a megállapításokra alapozott alkalmazások nem feltétlenül érvényesek, nem feltétlenül alkalmazhatóak orvosi szövegek esetén. Mindenképpen szükséges tehát a klinikai szövegek feldolgozására alkalmas eszközök egyedi fejlesztése.

További vizsgálatokat végeztünk az egyes korpuszok disztribúciós szemantikájára vonatkozóan is. Ennek során szintén lelepleződtek az alapvető különbségek, melyek a különböző szövegek közötti tartalmi eltérésből adódnak. Az orvosi szövegeknél, a viszonylag szűk domén miatt, ez a módszer alkalmas lehet disztribúciós tezaurusz építésére magyar nyelvű dokumentumok esetén is, hiszen látható, hogy a kimutatható hasonlósági relációk relevánsak, valódi összefüggéseket jelenítenek meg.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014 pályázatok támogatásával készült.

Hivatkozások

1. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association **1**(2) (1994)
2. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform **35** (2008) 128–44
3. Vincze, V.: Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 182–192
4. Siklósi, B., Novák, A., Prószéky, G. Number Lecture Notes in Computer Science 7978. In: Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. Springer Berlin Heidelberg (2013) 248–259
5. Siklósi, B., Novák, A. In: Detection and Expansion of Abbreviations in Hungarian Clinical Notes. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013) 318–328
6. Orosz, G., Novák, A., Prószéky, G. In: Hybrid text segmentation for Hungarian clinical records. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013)
7. Carroll, J., Koeling, R., Puri, S.: Lexical acquisition for clinical text mining using distributional similarity. In: Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II. CICLing'12, Berlin, Heidelberg, Springer-Verlag (2012) 232–246
8. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2. COLING '98, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 768–774

Automatikus morfológiai elemzés a korai Alzheimer-kór felismerésében

Papp Petra Anna[1], Rácz Anita[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
papp.petra.anna@gmail.com
raczanita89@gmail.com
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: Cikkünkben Alzheimer-kórral diagnosztizált páciensek és egy egészséges kontrollcsoport beszédátiratának a `magyarlanc` által automatikus morfológiai elemzésnek alávetett változatait vizsgáljuk azzal a céllal, hogy a betegek nyelvhasználati sajátosságait felmérjük. Bemutatjuk a gépi elemzés kézi annotálással történő korrigálása során talált hibákat és javaslatokat teszünk a lehetséges javítási módokra. Emellett összefüggéseket keresünk a kórban szenvedők és egészséges társaik nyelvi produkciója között acélból, hogy egy, a jövőben megvalósuló rendszer keretén belül az esetlegesen érintettek beszédjük alapján felismerhetőek legyenek.

Kulcsszavak: Alzheimer-kór, beszédátirat, morfológiai elemzés

1 Bevezetés

A szóbeli megnyilatkozások alapvetően különböznek az írásbeli formáktól, és olyan sajátosságokat is felmutatnak, amelyekre az írott szövegeken tanított nyelvi elemzők nincsenek kellőképpen felkészítve. Ez hatványozottan érvényes az Alzheimer-kórban szenvedő nyelvhasználók beszédére. A betegség beszédközpontra gyakorolt hatásának következtében nem csak a beteg számára okoz nehézséget a beszéd, hanem annak mások által történő megértése is akadályokba ütközik. A kórra utaló tünetek mihamarabbi észlelése annál is inkább lényeges, mert bár az Alzheimer-kór nem gyógyítható, mégis az időben érkező segítség nagyban javíthatja az érintettek életminőségét [1].

Ennek elősegítésére távlati célunk között szerepel egy olyan automatikus rendszer kifejlesztése, amely képes a korai Alzheimer-kórra jellemző nyelvi tünetek időben történő detektálására, így a beteg időben részesülhet a megfelelő kezelésben. Természetesen ennek a rendszernek nem célja, hogy helyettesítse az orvosi diagnózis felállítását, mégis fontos kiindulópontként szolgálhat annak megállapításában, hogy az Alzheimer-gyanús páciensek beszéde alapján ténylegesen fennáll-e a kór veszélye vagy sem. A rendszer beszélt nyelvi sajátosságokra, illetve a beszédátiratok automatikus morfológiai és szintaktikai elemzésén alapuló jellemzőkre épül.

Cikkünkben a beszédátiratok automatikus morfológiai elemzésére összpontosítunk. Vizsgálataink során azonosítottuk az elemző program számára nehézséget okozó

jelenségeket, melyek egyrészről a szóbeliség sajátosságaival magyarázhatóak, másrészről pedig magának az Alzheimer-kórnak tudhatóak be. A továbbiakban bemutatjuk a jellemző hibák csoportját, javaslatokat teszünk a lehetséges javítási megoldásokra. Emellett kitérünk arra is, hogy az egyes hibakategóriák előfordulási arányai alapján kimutatható-e szignifikáns különbség az Alzheimer-kóros betegek, illetve az egészséges kontrollcsoport nyelvhasználata között.

2 A beszédátiratok

Az orvosok, beszédfeldolgozók, nyelvtechnológusok, informatikusok és pszichiáterek együttműködését kívánó feladat első lépését a nyelvi produkciók rögzítése képezte. Ennek kivitelezése orvosi felügyelet alatt történt az ún. Memóriaklinikán. A kutatás több hónapja folyamatosan zajlik az Alzheimer-kórral egyértelműen diagnosztizálható páciensek, valamint egészséges személyek körében. A vizsgáltak mindegyike azonos feladatot kapott: két rövid árnyjáték megtekintését követően össze kellett foglalniuk az első tartalmát, majd el kellett mesélniük a tegnapi napjukat. Utolsó feladatként a második videót összegezték. A páciensek teljes nyelvi produkciójáról hangfelvétel készült, melynek felhasználásához teljes beleegyezésüket adták. Az így létrejövő személyenkénti 3-3 monológot természetesen az elemzés teljes folyamán az adatvédelmi elveknek és személyiségi jogok védelmének megfelelően kezeltünk.

A következő lépésben az anonimizált hangfelvételek a Szegedi Tudományegyetem beszédfeldolgozóihoz kerültek, akik azokat elemezték és a beszédjellemzők figyelembevételével, valamint azok kiejtéshű visszaadásával elkészítették szöveges átirataikat. Ebben a lépésben különleges hangsúlyt fektettek arra, hogy minden hangot és szót annak ténylegesen elhangzó formájában és hosszúságában adjanak vissza írásban, tehát a kiejtési sajátosságok mellett a beszédátiratok lényeges információkat hordoztak mind a beszédtempóról, mind a szünetekről, hezitációkról, nyújtott vagy téves szóindításokról stb.

Ezt követően a beszédátiratok a nyelvtechnológusokhoz kerültek, akik azokat több szempont szerint is elemezték. Egyfelől a beszéd során alkalmazott szókincset összesítettük és vizsgáltuk abból a célból, hogy a gyakori szavak lexikonjának összeállításával még pontosabb képet nyerhessünk a kórral járó esetleges lexikai sajátosságokról. Ezt a kérdéskört a későbbiekben még részletesen tárgyaljuk.

Másrészről a beszédátiratokat automatikus morfológiai elemzésnek is alávetettük a `magyarlanc` szoftver [2] segítségével. A gépi elemzést kézi ellenőrzés követte, melynek során egyértelműen azonosítottuk azon problémacsoportokat, amelyek különös nehézséget jelentettek az automatikus elemző számára. E jellemzők felismerése és kategóriákba sorolása azért oly lényeges, mert lehetővé teszi a `magyarlanc` felkészítését a kiejtéshű átiratok minél pontosabb elemzésére. Az azonosított hibák kategóriáit a következő pontban mutatjuk be.

3 Jellemző hibák a vizsgált személyek nyelvhasználatában

Ahogyan arra már korábban utaltunk, a szóbeli megnyilatkozások jellemzői jelentős eltéréseket mutatnak az írásbeli formáktól. Ennek megfelelően a szóbeliség átirataiban bizonyos hibakategóriák összeállítására törekedtünk, melyeket a későbbiekben a morfológiai elemző finomításához kívánunk felhasználni. Jelen vizsgálatunk 27 Alzheimer-kórban szenvedő páciens, valamint 19 kontroll személy beszédátiratain alapul.

A vizsgált betegek jellemző tévesztései magukban foglalták a törléseket, hasonulásokat, a beszélők által létrehozott szavakat, a kettős szóindításokat, homofónokat, valamint hezitációkat. Az említett hibakategóriák nem mindegyike statisztikailag is szignifikáns indikátora azonban az Alzheimer-kórnak.

A törlések például az élőbeszéd tipikus megnyilvánulásai, és a betegek csoportjának beszédátirataiban is magasan reprezentáltak: *azé (azért), há (hát)*. A törlések bizonyos esetekben ugyanakkor homofóniához vezethetnek, azaz olyan alakokhoz, mint például *mer (mert)* vagy *mér (miért)*. Ez az eset azért tekinthető speciálisnak, mert e szavak ténylegesen nem homonimák, csupán a sztenderdtől eltérő kiejtésüknek és átiratuknak köszönhetően válnak kétértelművé, ami további nehézségeket jelent a megfelelő automatikus elemzés számára. Ezek a hibatípusok ugyanakkor nem tekinthetőek a betegségre egyértelműen utaló tévesztéseknek, hiszen mind a sztenderd, mind a nyelvjárási beszélők nyelvhasználatának egyik jellemzője lehet.

Ugyanez érvényes a hasonulásokat alkalmazó nyelvhasználókra is, hiszen nem csupán az Alzheimer-kórosok, hanem az egészséges nyelvi beszélők is egyszerűsíthetik a szavakat ejtéskönnyítés céljából, így például természetes hangtani folyamatok következtében az *egyszer* gyakorta *eccer* alakban jelenik meg, míg a *képben* olykor *kébbe* alakot ölt.

Az új szavak alkotása a produktív nyelvhasználat egyik fontos jele, mely azonban szintén mindkét vizsgált csoportnál megfigyelhető volt. Az Alzheimer-kórral diagnosztizált betegek által kreált új kifejezések megfigyeléseink szerint azonban annyiban mégis különböznek az egészséges kontrollcsoportétól, hogy ezek jelentése kevésbé kikövetkeztethető a szövegkörnyezetből (pl. *sziriátum, bügyöre*). Mivel ez a kérdéskör további és tágabb vizsgálatokat igényel, ezért a morfológiai elemző finomítása egyelőre nem terjedt ki erre a problematikára.

Ugyanez igaz a kettős szóindításokra is, mint például *fé-férfi*, *asz-asztal*. Ezen alakok bár olykor az egészséges beszélőknél is fellelhetőek, az Alzheimer-kóros pácienseknél különösen magas számban észleltük őket. Erre a jelenségre az elemző hibáinak manuális javításakor lettünk figyelmesek.

A betegek beszédében mégis a hezitációk száma és hosszúsága szolgálhat a legfontosabb tényezőként az egészséges társaiktól való megkülönböztetésben. Tény, hogy az élőbeszéd az egészséges személyek esetében is gyakorta hezitációkkal tűzdelt, mégis a betegséggel diagnosztizáltak nyelvi produkciójában található hezitációk száma szignifikánsan különbözik a kontrollcsoport adataitól.

Erre vonatkozó adataink is ezt bizonyítják. A beszédiratok összesítéséből és elemzéséből kiderült, hogy az 549 hezitációs jel a kontrollcsoport esetében csaknem fele a beteg társaik által produkáltának (1005), a különbség statisztikailag szignifikáns ($p= 0,0424$). Ezek alapján feltételezhető, hogy a hezitációk száma a kór fontos indikátora lehet és segítségünkre lehet annak felismerésében. Jelenlegi adataink szerint a diagnózis felállítása így nagyban épülhet a hezitációk gyakoriságára.

A következőkben példával is demonstráljunk az Alzheimer-kóros betegek és a normál beszélők megnyilvánulása közötti különbséget. A következő egy a kórral diagnosztizált beteg beszédátiratából származó részlet:

Baloldalt egy fie..fiatalember.. mászott le.. lepkehálóval.. amivel a lepkéket szokták megfogni. És .. avval átment a másik oldalra. És utána jött egy… nő… és a kosárba vót.. üveg.. pohárak… és akkor ittak belőle egyegy valami volt benne amit ittak. Nem tudom mi lehetett… És abból ittak és utána elment mindakettő.

Ez alapján jól megfigyelhető a kettős szóindítások és a hezitációk gyakorisága is. A szöveg jellemző eleme az *és* kötőszó, illetve annak gyakori használata. A *valami* névmás és a *nem tudom* kifejezések itt is megjelennek, ezen egységek fontosságára még kitérünk. A szövegrészlet rövidsége rámutat ugyanakkor arra, hogy komoly memóriazavarokkal küszködik a beszélő, hiszen a leglényegesebb szavakon kívül mást nemigen tudott visszaidézni.

A következő szövegrészlet egy egészséges személytől származik. Ebben is található néhány hezitálás, de lényegesen kevesebb, mint az előbbiben. Mint említettük, törlést az átlagos nyelvhasználók is elkövetnek, ezt jól reprezentálja a példa több pontja is.

Egy férfi és egy nő szerintem kirándulni mentek. És.. azt a röpködő valamit nem ismertem föl. Pillangónak kellett volna lenni, gondolom, mert olyan lepkefogóval volt a férfi, éss azt szerette volna megfogni. Közbe egy ilyen mélyedésbe lekerült. Ott virágot tépett. Odaadta a szíve hölgyének, aki erre föl a kosarából egy flakon valamit, amit nem lehet tudni milyen itóka volt, odaadta. A férfi abból ivott utána a nő is ivott belőle. És… és ennyi.

4 Jellemző hibák a morfológiai elemzésben

Az írott szövegeken tanított automatikus morfológiai elemző komoly nehézségekbe ütközött a kiejtéshű beszédátiratok elemzésekor. A gépi analízis kézi ellenőrzése során azonosított hibák alapvetően három, egymástól többé-kevésbé egyértelműen elkülöníthető kategóriába sorolhatóak. E csoportok szoros összefüggésben állnak az előző pontban említett nyelvi tévesztésekkel is, és egyértelmű elkülönítésük jelentős szerepet játszhat az elemző javításában.

Az első esetben X-es (ismeretlen) kóddal látja el az elemző azon megnyilatkozásokat, melyek morfológiai azonosításakor problémákba ütközik. Ez magában foglalja egyrészt a különböző hezitációkat jelölő formákat, másrészt pedig a törléseket, azaz a szóbeliséget és kiejtést hűen tükröző, és így a helyesírás szabályait felbontó átiratokat (pl. *há*, *azé*). A hezitációk magas aránya miatt külön figyelmet fordítottunk azon X-es kódú elemekre, amelyek hezitációra utalnak.

A hibaforrások másik nagy csoportját azon esetek alkotják, melyekben az elemző hibás (de létező) kódot rendel hozzá a szavakhoz. Kutatásaink szerint ez a jelenség a kettős szóindítások mellett főként az imént említett homofón szavaknál tapasztalható.

A harmadik osztályba végül a páciensek által újonnan képzett, és ily módon a nyelvhasználatban nem elterjedt egységeket soroltuk. Ezek esetében az elemző nem

jár el egységesen: többnyire ismeretlen szónak jelöli vagy a szóalak és toldalékai (pl. *sziriátum* latinra utaló *-um* végződése) alapján helyesen felismeri és elemzi azokat.

5 A `magyarlanc` átalakítási lehetőségei

A `magyarlanc` morfológiai és szintaktikai elemző a Szeged Korpusz [3] sztenderd nyelvhasználatú szövegein lett betanítva. A statisztikai elemzések alapján a beszédátiratokban talált ismeretlen szavak kb. 3,7%-os aránya jobban közelít a HunLearner [4] magyar nyelvtanulói korpuszban (szintén nem sztenderd magyar szövegekben) található ismeretlen szavak arányához (5,5%), mint a Szeged Korpuszéhoz (0,4%), amely számszerűsítve is alátámasztja a `magyarlanc` adaptációjának szükségességét a beszédátiratokra.

Az adaptációt elsődlegesen a morfológiai elemző szótárának kibővítésével kívánjuk megvalósítani, melyben elsődlegesen a beszédátiratokban talált hezitációkra fókuszálunk. Ez a folyamat tartalmazza azok kézi kigyűjtését a korpuszból és egyszerűsített átiratukat, melyben a hezitáció lehetséges megnyilvánulásait egységes alapokra hozzuk. Ezt a fajta normalizációt azért találtuk szükségesnek, mert az átiratokban a habozást jelölő alakok például a szókeresés időtartamára is utalnak, mint például *öööö* vagy *ööööööööö*. Célunk ezen alakok egy olyan prototipikus formával történő helyettesítése (pl. *öö*), mely segítségünkre lehet az eddig ismeretlenként kezelt egységek megfelelő felismerésében.

6 Eredmények

A jellemző hibák feltárásán és azok javítására vonatkozó javaslatainkon túl tanulmányunkban azt a célt is megfogalmaztuk, hogy különbségeket keressünk az Alzheimer-kórban szenvedő páciensek, valamint az egészséges kontrollcsoport között. Ehhez kapcsolódóan a következő táblázatban mutatjuk be a rendelkezésre álló beszédátiratok részleteit a vizsgált csoportok megoszlásában:

1. táblázat: A vizsgált csoportok beszédátiratainak összesített statisztikája

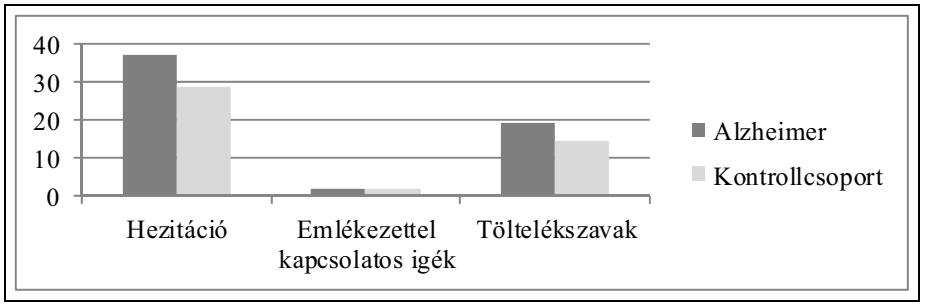
	Alzheimer	Kontroll
Mondatok száma	786	585
Átlagos mondatszám egy beszélőre	29	31
Tokenek száma	11 614	8108
Átlagos tokenszám egy beszélőre	430	426

Az adatok egy szembetűnő tényt tűnnek alátámasztani. Az Alzheimer-kórral diagnosztizált betegek nyelvi produkciójának terjedelme ugyanis nem tér el lényegesen az egészséges kontrollcsoport által átlagosan használt mondat-, illetve szószámtól. Az

Alzheimer-kóros és az egészséges emberek megnyilatkozásai közötti különbség tehát kevésbé értelmezhető pusztán mennyiségi szempontból.

Az előbb közölt két beszédátirat-részletből is kitűnik azonban, hogy jelentős minőségbeli eltérések jellemzik a két csoport átiratait.

Ahogyan azt a következő ábra is bemutatja, az Alzheimer-kóros páciensek beszédét gyakrabban kísérik töltelékszavak, a nyelvi redundancia és a hezitációk száma jelentősebb, mint a kontrollcsoportéban:



1. ábra: Alzheimer-kóros és egészséges nyelvhasználók hezitációinak, emlékezettel kapcsolatos megnyilatkozásainak és töltelékszavainak átlagos száma

Jelen statisztika három nyelvi jelenségről szolgáltat információt. Az első ilyen kategória a már korábban tárgyalt hezitációk és azok beszélőnkénti átlagos száma. A táblázatból kitűnik, hogy ez képezi a legjelentősebb minőségi különbséget a két vizsgált csoport megnyilatkozásaiban. Míg az egészséges emberek átlagosan körülbelül 29 alkalommal tartottak szünetet beszédükben, addig ez a szám az Alzheimer-kóros páciensek esetében ennél nyolc alkalommal, azaz csaknem 30 százalékkal gyakoribb volt. Ez az adat tehát egyértelműen alátámasztja azon korábban tett megállapításunk létjogosultságát, miszerint az Alzheimer-kór időben történő diagnózisában nagy szerepet játszhat a hezitációk gyakorisága.

A második, emlékezettel kapcsolatos megnyilatkozások csoportjába tartoznak az olyan kifejezések, mint például az *elfelejtettem* vagy a *nem tudom* (illetve ennek egyszerűsített alakja, a *nem tom*). Bár nem mutat szignifikáns különbséget, de valamivel jellemzőbb a betegekre, hogy a különböző kérdésekre *nem tudom* vagy *nem emlékszem* jellegű válaszokat adnak.

A töltelékszavak kategóriájába többek között a következő szavakat soroltuk: *ilyen, olyan*, *izé*, *és aztán*, *és akkor*, illetve a határozatlan névmásokat, úgymint *valamilyen*, *valahogy*, *valamerre*. Ezeket a szavakat nem bontottuk további csoportokba, mert ezek együttesen mutatnak rá, hogy az Alzheimer-kóros beszélők gyakran helyettesítenek szavakat határozatlan névmásokkal vagy az *izé* szócskával. Melléknevek helyett pedig előszeretettel használnak parafrázisokat. Ennek megfelelően nem ritkák az *egy ilyen bagolyszerűség* vagy az *olyan délelőtt volt* körülíró, bizonytalanságra utaló kifejezések (vö. [5]), melyek jelenléte a harmadik pontban közölt Alzheimer-kóros beteg beszédében is megtalálhatóak. Az eredmények statisztikailag szignifikánsak ($p = 0{,}0316$), így egyfajta indikátora lehet a kórnak a töltelékszavak gyakori használata is.

7 Az eredmények felhasználása

Amint azt az előbbiekben bemutattuk, a hezitációk, töltelékszavak előfordulási arányai szignifikáns eltéréseket mutatnak az Alzheimer-kórban szenvedők és az egészséges emberek esetében, így a fenti jellemzők jól hasznosíthatók a korai Alzheimer-kór diagnosztizálásában. Mindemellett ezek az eredmények a jövőben más nem sztenderd szövegek, például az internetes nyelvhasználat elemzésében is segítségünkre lehetnek.

A diagnózis felállítását megkönnyítendő olyan alapszókincs összeállításán dolgozunk, amely tartalmazza az egyes árnyjátékok tartalmának bemutatásakor használandó szavak listáját. Ebben az előforduló szavak gyakorisági statisztikáira, illetve a nyelvészek által összeállított listára építünk. Utóbbi szemléltetésére itt is közlünk egy részletet, amely azt mutatja be, hogy az egyik lejátszott filmhez kapcsolódó központi jelentőségű névszókkal általában milyen cselekvések társíthatóak a bemutatott kontextusnak megfelelően:

Névszó	**Cselekvés**
Férfi	ül; hív/int a pincérnek; kettőt mutat
Pincér/felszolgáló	Megérkezik/bejön/jön; tisztítja/törli/söpri az asztalt;
Gyűrű	Férfi előveszi, majd kinyitja fényesíti, nézegeti
Pincér	Hoz / leteszi a tálcán 2 poharat+1 üveget
Nő (esernyővel)	Bejön/megérkezik; kezet csókol neki a férfi; nő összecsukja/becsukja az ernyőjét, leteszi a székre
Pincér	Kihúzza a széket; igazgatja a széket
Férfi	Int/szól a pincérnek, hogy töltsön innivalót
Pincér	Lenyomja a nő fejét; nyakába önti/leönti a nőt
Férfi	Felpattan; előrántja a botot/sétapálcát
Pincér	Elbújik a nő széke mögé; vívnak/kardoznak az esernyővel (pincér) és bottal (férfi)

Ez a lexikai oldalról való megközelítés azért is lesz központi jelentőségű, mert így lehetőségünk nyílik az Alzheimer-kórosok nyelvi produkciójának minőségbeli különbségeinek felmérésére.

Tervezzük emellett a rendelkezésre álló szótárunk további bővítését, és például további töltelékszavak kigyűjtését, rendszerezését. A morfológiai hibák kigyűjtésére és rendszerezésére is hangsúlyt szeretnénk a későbbiekben fektetni. Ennek keretein belül a szintaktikai hibákat is elemzés alá vonjuk a Szeged Dependency Treebank [6] vonzatkereteinek segítségével.

Hosszabb távú terveink közé tartozik továbbá, hogy gépi tanulási eszközökkel a beszélőket beszédátirataik alapján elkülönítsük annak tekintetében, hogy fennáll-e náluk az Alzheimer-kór veszélye vagy sem.

8 Összegzés

Tanulmányunkban 27 Alzheimer-kórral diagnosztizált páciens és 19 egészséges személy nyelvi produkciójáról készült beszédátiratot vizsgáltunk acélból, hogy a nyelvi megnyilatkozásaikban alapvető különbségeket definiáljunk. Elemzéseink kiindulópontját a `magyarlanc` automatikus morfológiai elemző képezte, mely számára a szóbeli formák és különösen az alzheimeresek által produkált nyelvi alakok komoly kihívást jelentettek.

Ahogyan azt a statisztikák kimutatták, jelentős minőségbeli eltérések észlelhetőek az egészséges és a beteg személyek nyelvi produkciójában, melyek főként a hezitációk és a töltelékszavak gyakoriságát foglalják magukban. Célunk, hogy ezen aspektusok figyelembevételével tovább finomítsuk az elemzőt. Meggyőződésünk, hogy ez nagyban hozzájárulhat ahhoz, hogy az esetlegesen érintettek esetében beszédük alapján mielőbb felismerhessük a kórt és így megfelelő kezelésben részesülhessenek.

Köszönetnyilvánítás

Jelen kutatást a Telemedicina fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Hoffmann, I., Németh, D., Dye, Ch. D., Pákáski, M., Irinyi, T.; Kálmán, J.: Temporal parameters of spontanteous speech in Alzheimer's disease. In: International Journal of Speech-Language Pathology (2010) 12(1) 29–34
2. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013)
3. 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
4. Vincze V., Zsibrita J., Durst P., Szabó M. K.: HunLearner: a magyar nyelv nyelvtanulói korpusza. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 97–105

5. Vincze, V.: Bizonytalanságot jelölő kifejezések azonosítása magyar nyelvű szövegekben. In: Tanács A., Varga V., Vincze V. (szerk.): X. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2014)
6. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Málta (2010)

A magyar Braille-rövidírás megújítása félautomatikus módszerrel

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat A dolgozatban az új magyar Braille-rövidírást illetve létrehozásának módját mutatjuk be. A félautomatikus eljárás két részből áll: egy automatikus, korpuszvezérelt elven működő algoritmus határozza meg a legalkalmasabb rövidítendő elemeket; ezt követi a manuális véglegesítő lépés a kényelmes használhatóság szempontjainak figyelembevételével. A létrejött 33 elemű szabályrendszer könnyen megtanulható, jól olvasható, jól felismerhető. Rövidítési képessége 13,3%, mely 3,4%-kal növeli meg a ma használatos kis rövidírás (9,9%) hatékonyságát. Az új rövidírás alkalmas a vakok általi tesztelésre és majdani használatra.

Kulcsszavak: Braille-írás, Braille-rövidírás, korpuszvezérelt, rövidítés, rövidítési képesség, szabály, gyakoriság, használhatóság

1. Célkitűzés

A vakok által világszerte használt, tapintáson alapuló Braille-írásnak számos nyelvre létezik ún. Braille-*rövidírás* változata (angol: [1]; német: [2,3]). Ezek az általános Braille-írást nyelvspecifikus rövidítési, tömörítési szabályokkal egészítik ki. Rövidírás használatával gyorsul az írás-jegyzetelés és az olvasás folyamata. Napjainkban, a speciális Braille-nyomtatók egyre szélesebb körű elterjedésével az is fontos, hogy a rövidírással írt szöveg kinyomtatva jelentősen kisebb terjedelmű. 2012-2013-ban valósult meg a projekt a Magyar Vakok és Gyengénlátók Országos Szövetsége és az MTA Nyelvtudományi Intézet együttműködésében, melynek keretében a magyar Braille-rövidírást a mai nyelvhasználatot is figyelembe vevő új rövidítésekkel bővítjük, azzal a céllal, hogy a rövidítési képessége a korábbi nagyjából 10%-ról jelentős mértékben, akár 15-20% közelébe növekedjen [4].

2. A magyar Braille-írás

A Braille-karakterek (ún. Braille-cellák) két oszlopban elrendezett 3-3, azaz összesen hat kidomborodó pontból állnak. Az egyes pontokra a következő elrendezésben számokkal hivatkozunk: $\begin{smallmatrix}1&4\\2&5\\3&6\end{smallmatrix}$. A kidomborodó és ki nem domborodó pontok mintázataiból összesen $2^6 = 64$ féle különböző karakter áll elő. A tapintható írásrendszerek történeti bemutatásáról l. [5]-t.

2.1. A magyar teljesírás

A magyar Braille-karaktereket az 1. táblázatban láthatjuk a Görgényi Miklóstól átvett elrendezésben [6, 1. fejezet]. Az első sorban a felső 4 pontot elfoglaló jelek vannak, a következő 3 sor ehhez teszi hozzá rendre a 3-as, a 3-as/6-os, illetve a 6-os pontot. Az 5. sor az első sor jeleinek *csúsztatott*, azaz egy ponttal lefelé mozdított megfelelőit tartalmazza. Az utolsó három oszlopban vannak azok a jelek, melyekben nem szerepel az 1-es/2-es pont. A bal és jobb oszlopban is, valamint a felső és alsó sorban is pontot tartalmazó ún. *erős* (angolban: *strong* [1, 28. oldal]) jeleket szürke háttérrel jelöltük. Az $5 \cdot 13 = 65$ pozíción szépen ábrázolódik a 64 jel, az üres jel kétszer fordul elő (a 11. oszlopban). Az angol ábécé betűi nagyjából az első három sor első tíz oszlopában találhatók.

1. táblázat. A magyar Braille-ábécé.

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
⠁ [a]	⠃ [b]	⠉ [c]	⠙ [d]	⠑ [e]	⠋ [f]	⠛ [g]	⠓ [h]	⠊ [i]	⠚ [j]	⠀ [␣]	⠡ [á]	⠈ [@]
⠅ [k]	⠇ [l]	⠍ [m]	⠝ [n]	⠕ [o]	⠏ [p]	⠪ [ö]	⠗ [r]	⠎ [s]	⠞ [t]	⠲ [.]	⠌ [í]	⠵ [Z]
⠥ [u]	⠧ [v]	⠭ [x]	⠽ [y]	⠵ [)]	⠟ [q]	⠼ [#]	⠳ [ü]	⠦ [(]	⠾ [ű]	⠤ [-]	⠬ [ú]	⠙ [D]
⠮ [é]	⠵ [z]	⠠ [C]	⠛ [G]	⠎ [S]	⠝ [N]	⠪ [ő]	⠞ [T]	⠬ [ó]	⠺ [w]	⠠ [V]	⠁ [A]	⠇ [L]
⠂ [,]	⠆ [;]	⠒ [:]	⠓ [H]	⠦ [?]	⠖ [!]	⠶ [=]	⠴ [”]	⠔ [*]	⠠ [Q]	⠀ [␣]	⠗ [R]	⠋ [F]

A Braille-ábécé nagybetűket nem tartalmaz, a táblázatban szereplő nagybetűk speciális jelentéssel bírnak. A tényleges nagybetűket két Braille-karakterrel írjuk le: a nagybetűjelölő ⠠ [A] után írjuk a megfelelő betűt, pl. *Ő* = ⠠⠪. A táblázatban tehát a speciális karakterek – amilyen a nagybetűjelölő is – átírására egyezményes nagybetűs jeleket használunk. Az eredeti táblázatot kicsit módosítva a magyar teljesírás újabb szabályainak megfelelően W helyett @-t, pont helyett H-t, X helyett pontot írtam. A magyar ékezetes betűk és kettősbetűk – legtöbbször tükrözéssel kialakított – önálló Braille-jellel rendelkeznek, utóbbiak a következő bevett jelölésekkel: cs=C, gy=G, ly=L, ny=N, sz=S, ty=T, zs=Z. További karakterek: az ⠈ [R] ékezetjelölő; a ⠨ [V] védő karakter, ha ez szerepel egy karakter(sorozat) előtt, akkor azt nem rövidítésként, hanem literálisan, karakterenként kell olvasni; a ⠙ [D] a számjelölő, mely az *a*-tól *j*-ig terjedő betűkből számjegyet képez (az 5 Braille-ben ⠙⠑ [De]); a ⠓ [H] és az ⠋ [F] jelhez pedig nincs jelentés rendelve, ezek ún. *szabad gyök*ök. A táblázatban látható karakterekkel lehet az egyes Braille-cellákat sima szövegre átírni, ezt az átírást a dolgozatban használni is fogjuk: a *Magyarország* szó átírása például *AmaGarorSág*.

2.2. A magyar „kis” rövidírás

Amint már utaltunk rá, a Braille-rövidírások az írott anyag tömörítésére szolgálnak, csökkentve a helyigényt és gyorsítva az írás-olvasást. A rövidírás rövidítési

szabályok gyűjteménye. Ma hazánkban sztenderd módon az ún. „kis” rövidírást használják, ami a korábbi jóval bonyolultabb nagy rövidírás [7] könnyen megjegyezhető szabályaiból áll. A kis rövidírás rendszerét a 2. táblázatban foglaltuk össze [6] 7. fejezete alapján. Feltüntettük a mért rövidítésiképesség-értékeket is.

2. táblázat. A kis rövidírás szabályai.

rövidítéscsoport	rövidítési képesség
1. Nagybetűjel ⁝ törlése.	+2,3%
2. Vessző utáni szóköz törlése.	+1,4%
3. A határozott névelők rövidítése: r(⁘[az])=⁝[.], r(⁘*[a]*)=⁖*[,]*), és az utánuk lévő szóköz törlése.	+1,9%
4. Az alábbi 44 szabály alkalmazása.	+4,3%

7 szóvégi rövidítés		16 egyjelű szórövidítés				21 kétjelű szórövidítés					
-ban/-ben	b	Cak	C	meL	m	aNNi	ai	mind	md	rövid	rd
-ból/-ből	b.	de	d	nem	n	boldog	bg	mint	mt	forr	rr
-hoz/-hez/-höz	h.	és	é	óta	ó	eNNi	ei	orSág	og	Sabad	Sd
ként	k.	hoG	h	pedig	p	gond	gd	olvas	os	tanáC	tC
-ról/-ről	r.	is	i	tehát	t	függ	gg	öSSe	öe	teljes	ts
-tól/-től	t.	íG	í	után	u	Gors	Gs	pont	pt	világ	vg
-val/-vel	v	kell	k	úG	ú	keres	ks	pénz	pz	volt	vt
		leS	l	van	v						
+1,4%		+1,8%				+1,1%					

A kis rövidírás rövidítési képesség mindösszesen: 9,9%

A *-val/-vel* ragot hasonulás esetén is *v*-vel rövidítjük. A pontot tartalmazó szóvégi rövidítéseknek bevezetés alatt áll egy újabb formája, mely a rag első és utolsó betűjéből áll. Az egyjelű szórövidítéseket önálló szóként és összetételben, a kétjelűeket ezen kívül bármilyen toldalékolt forma esetén is alkalmazzuk. Előfordul, hogy egy rövidítésként is értelmezhető karaktersort literálisan akarunk értelmeztetni, ilyenkor a már említett ⁝[V] védőjellel prefixáljuk a karaktersort. A ⁝⁘⁘⁝[Serb] tehát a *szer* főnév *-ban/-ben*-ragos alakja, a ⁝⁘⁘⁝⁝[SerVb] viszont a *szerb* népnév.

Látjuk, hogy az első 3 csoportban lévő „trükkös” (információvesztő és szóközlenyelő) szabályok nagyon jelentős rövidítési képességgel bírnak. A negyedik csoportban lévő szabályoknál az egy szabályra eső rövidítési képesség folyamatosan csökken, rendre: 0,2%, 0,12%, a kétjelű szórövidítéseknél – melyek sok esetben viszonylag ritka szavakat rövidítenek – pedig csak 0,05%.

3. Rövidítési képesség vs. használhatóság

Célunk tehát a magyar rövidírás rövidítési képességének jelentős növelése, de nem csak ez a szempont vezérli a szabályrendszer kialakítását. Ugyanilyen fontos

az is, hogy a végső rendszer kényelmesen használható legyen. A munkálatok során a következő használhatósági követelmények körvonalazódtak:

1. az új szabályrendszer az ismert kis rövidírást egészítse ki;
2. *jó olvashatóság*: a rövidítések emlékeztessenek az eredetire;
3. *jó felismerhetőség*: tapintás útján könnyen felismerhető jelek alkalmazása;
4. *könnyű megtanulhatóság*: kevés, egyszerű szabály.

A nagy rövidítési képesség és kényelmes használhatóság egymás ellen ható követelmények, itt egy körültekintően kidolgozott kompromisszumra van szükség annak érdekében, hogy a potenciális felhasználók elfogadják és szívesen alkalmazzák az új rövidírást. A fenti feltételeknek kevéssé megfelelő nagy rövidírás éppen bonyolultsága miatt nem terjedt el korábban. Jelen dolgozatban bemutatjuk a kialakított kompromisszumos javaslatot, mely törekszik a maximális rövidítési képesség elérésére, miközben megfelel a használhatósági feltételeknek is.

4. Alapötlet: korpuszvezéreltség

A rövidírás-rendszerek kifejlesztése sok esetben nagy időigényű feladat, az egységes angol rövidírás kialakítása majdnem két évtizedet vett igénybe [8].

Jelen munkálat alapötlete azon a felismerésen alapul, hogy az ideális rövidítési szabályok a magyar nyelv rendelkezésre álló korpuszgyakorisági adatai alapján, korpuszvezérelt módon, automatikusan meghatározhatók, és ezek alapján a lehető legnagyobb rövidítési képességgel bíró új magyar rövidírás záros határidőn belül elkészíthető. A háttérben az az egyszerű gondolat van, hogy nyilván a lehető leggyakoribb elemeket (betűsorozatokat) érdemes a lehető legrövidebbre rövidíteni, ekkor nyerjük összességében a legtöbbet. A rövidség szempontjából tehát nem volt ideális választás annak idején a ritka *ty* kettősbetű egyjelű rövidítése: r(⠽⠞[ty])=⠾[T] a nála akár 20× gyakoribb kétkarakteres elemek (pl.: *et*) helyett. A fenti gondolat kiegészül azzal, hogy „mohó" eljárást követünk, azaz mindig azt az aktuális új szabályt választjuk, ami éppen a legnagyobb rövidítést eredményezi.

Cél volt a lehető legkisebb emberi beavatkozás, de nyilvánvalóvá vált, hogy a használhatósági feltételeknek való megfelelés nehezen automatizálható. A teljes folyamat tehát nem megy automatikusan: szükséges a szakértői közreműködés a szabályok kézi véglegesítése során. Leegyszerűsítve mondhatjuk, hogy automatikusan áll elő az, hogy *mit* rövidítünk, és manuálisan, hogy *mire*.

A kutatás során a fenti ötlet szerint jártunk el, mert így minden említett követelménynek meg tudtunk felelni, és formailag is olyan szabályokat tudtunk alkotni, melyek hasonlóan a kis rövidírásban használt szabályokhoz. Alább néhány alternatív megközelítést említek. Szóba kerülhet (1) a gyorsírás vizsgálata; (2) az sms és/vagy twitter korpuszok vagy (3) rövidítéstárak tanulmányozása; vagy annak direkt felmérése, hogy (4) a fiatal vakok hogyan rövidítenek. Nem járható út (5) a teljes magánhangzó-elhagyás ötlete a nehéz olvashatóság, (6) prefixfák használata pedig az írás nehezítettsége miatt. Érdemes lehet megvizsgálni (7) az

általános tömörítő algoritmusok architektúráját is. Egy ilyen módszert alkalmazott Arató András kandidátusi értekezésében [9, 7. fejezet], melynek keretében a rövidírás automatikus kialakíthatóságát is érinti.

5. A rövidítendő elemeket megadó automatikus eljárás

Az alapelv tehát az, hogy a lehető leggyakoribb és leghosszabb elemeket próbáljuk a lehető legrövidebbre rövidíteni. Ezt úgy valósítjuk meg, hogy meghatározzuk az adott helyzetben éppen legnagyobb rövidítési képességet adó szabályt, majd az így kialakult új helyzetben ismét az aktuális legnagyobb rövidítési képességet adó szabályt, és így tovább. A rövidítési képességet – jele: rk() – az 1. ábrán látható módon számítjuk.

$$\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] \cdot \mathrm{fq}(w)$$

1. ábra. A rövidítési képesség számítása. w – az eredeti rövidítendő karaktersorozat, $r(w)$ – a rövidítés, $l()$ – a hossz (karakterszám), fq() – a gyakoriság (millió szóra eső előfordulási szám). Megjegyzés: az empirikus úton meghatározott, azaz korpuszon mért rövidítési képességre is ugyanúgy az rk() jelölést alkalmazzuk, a szövegösszefüggésből mindig világos, hogy a kettő közül melyikre gondolunk.

A *sem* elem gyakorisága (3302) például kicsivel több, mint a *Serint* elem gyakorisága (2515). Ha egy karakterre rövidítünk – például s és S a két rövidítésjelölt –, a rövidítési képesség a következő: $\mathrm{rk}(\mathrm{sem}) = (3-1) * 3302 = 6604$, valamint $\mathrm{rk}(\mathrm{Serint}) = (6-1) * 2515 = 12575$. Nyilvánvalóan érdemes a (ritkább) *Serint*-et rövidíteni, mivel így az elért rövidülés majdnem kétszeres. De még akkor is ugyanez lenne a jó döntés, ha a *sem* egyjelű és a *Serint* kétjelű rövidítése közül kellene választanunk: $\mathrm{rk}(\mathrm{Serint}) = (6-2) * 2515 = 10060$.

Az algoritmus a következő lépésekből áll.

1. Korpusz alapján elkészítjük a szavak gyakorisági listáját.
2. Számba vesszük az összes elvben rövidíthető nyelvi elemet: a szavak gyakorisági listájából származtatjuk a karakter-n-gramok gyakorisági listáját, minden karakter-n-gramot külön-külön számolva.
3. A rendelkezésre álló rövidítésjelek hosszának ismeretében kiszámoljuk az elemek rövidítési képességét (1. ábra).
4. Rendezzük a gyakorisági listát rövidítési képesség szerint.
5. A rendezett lista *első* bejegyzése adja a maximális rövidítési képességet, vesszük ezt az elemet, és hozzárendelünk egy megfelelő rövidítést.
6. Az így kialakított rövidítő szabályt *alkalmazzuk* a szavak gyakorisági listájára.

7. Újból a 2. pontra lépünk, amíg el nem érünk egy elfogadható összesített rövidítési képességet, vagy a szabályrendszer mérete át nem lép egy adott küszöböt.

Ez az algoritmus a korábban javasolt algoritmus [4] továbbfejlesztett változata, mely az éppen létrehozott szabálynak a szavak gyakorisági listájára való alkalmazása által jól kezeli a rövidítendő elemek esetleges átfedésének a rövidítésiképesség-értékekre gyakorolt módosító hatását. Azért szükséges mindig az eredeti, szavakat tartalmazó gyakorisági listára alkalmazni a szabályt, mert az n-gramok adott esetben egy rövidítendő elemnek csak egy részletét tartalmazzák az n-gram elején vagy végén.

Pontosan azért van szükség a rövidítési képességek újraszámolására, és a lista újrarendezésére, mert az éppen aktuálisan létrehozott szabály alkalmazása befolyásolja (csökkenti) bizonyos elemek hosszát, így az azok esetleges továbbrövidítése során elérhető rövidülés változik. Az újrarendezés során létrejövő sorrendváltozást, helycserét a 2. ábrán szemléltetjük.

w	$\mathrm{fq}(w) \cdot l(w) - l(r(w))$	$\mathrm{rk}(w)$
Ser	10901	$\times 1 = 10901$
Serint	2515	$\times 4 = 10060$
maGar	3326	$\times 3 = 9978$
...		

(a) A gyakorisági lista eleje.

w	$\mathrm{fq}(w) \cdot l(w) - l(r(w))$	$\mathrm{rk}(w)$
maGar	3326	$\times 3 = 9978$
...		
Srint	2515	$\times 3 = 7545$
...		
Sr	10901	$\times 0 = 0$

(b) Sorrend a *Ser* → *Sr* rövidítés után.

2. ábra. Sorrendváltozás a futás során.

A 2(a) ábrán karakter-n-gramok gyakorisági listájának eleje látható az algoritmus futásának egy pontján. Tegyük fel, hogy két karakterre rövidítünk. Az algoritmus 5. pontja szerint a rövidítendő elem a *Ser* lesz, a hozzá társított rövidítés legyen a *Sr*. Ennek az új szabálynak a szógyakorisági listára történő alkalmazása után (algoritmus 6., 7. és 2. pont) áll elő a 2(b) ábrán látható helyzet. Mivel a *Serint*-ből az új szabály az egy karakterrel rövidebb *Srint*-et hozza létre, ez az elem (jóval) lejjebb kerül a listán, és a következő rövidítendő elem a *maGar* lesz. A *Ser*-ből lett *Sr* kétjelű rövidítéssel nyilván nem rövidíthető tovább. Kiemelendő, hogy kialakulhat olyan helyzet, hogy az algoritmus további futása során a *Srint* végül mégiscsak a lista elejére kerül és (tovább)rövidül.

A következő példa is az algoritmus működését világítja meg. Egy karakterre rövidítve a legnagyobb rk()-gel rendelkező magyar karaktersorozat az *et*: ha *y*-nal rövidítjük, akkor a (főként igevégződésként) nagyon gyakori *ett* elem *yt*-ként jelenik meg. Ha azonban úgy döntünk, hogy csak kétjelű rövidítéseket alkalmazunk, akkor az *et* nem rövidül ($\mathrm{rk}() = 0$), viszont az algoritmus futása során hamar a lista elejére kerül az *ett*.

A rövidítendő elemek esetleges *átfedése* miatt nem lehet azt megtenni, hogy a szabályokat egymástól függetlenül alakítjuk ki. Az *et* egyjelű rövidítése önmagá-

ban rk() = 0, 77%-os, a *te* elemé pedig 0, 57%-os rövidülést jelent. A gyakorlatban egymás után alkalmazva a két szabályt már csak 0, 77% + 0, 24% az eredmény, mivel az első szabály nagyon sokszor az *ete* elem első két karakterét rövidíti, elvéve ezzel a lehetőséget a második szabály elől. Az épp kialakított új szabály teljes szógyakorisági listára való alkalmazása (algoritmus 6. pont) oldja meg ezt a problémát. Így mivel minden ponton világos, hogy adott szabálynak mennyi a tényleges rk()-e, könnyen kiválasztható a legjobb szabály, a szabályok egymással összefüggésben, egymásra épülve jönnek létre.

Olyan eset azonban előfordulhat, hogy egy szabály bal oldala egy rövidítést teljes egészében tartalmaz, ezt nevezzük *továbbrövidítés*nek. Ha például r(*ek*)=*!*, és a lista elejére kerül a (már csak 4 karakteres) *Ger!* elem, akkor ezt rövidíthetjük *Gk*-val. A továbbrövidítést kifejtve két független szabályt kapunk – r(*Gerek*)=*Gk* és r(*ek*)=*!* –, melyeket a tényleges rövidítési folyamat során ebben a sorrendben kell alkalmaznunk. Pontosan ezen a módon jön létre a r(*Serint*)=*St* – r(*Ser*)=*Sr* szabálypár is (vö: 2(b) ábra).

Fontos, hogy a továbbrövidítés során csak *teljes* rövidítéseket rövidítsünk tovább: r(*ett*)=*eT* esetén ne akarjuk például a *melleT* szóban lévő *elle* elemet rövidíteni, mely egy rövidítéssel részben fed át. Ez ahhoz vezetne, hogy nem tudunk egyértelmű, független szabályokat kialakítani a fenti módon, és több menetben kellene kifejteni a rövidítéseket, ami nagyon megnehezítené a rövidírás olvasását. Továbbrövidítéskor tehát teljes egészében tartalmaznia kell az új szabálybaloldalnak a korábbi rövidítést, azaz a rövidítéseket egy egységként kell kezelnünk. Ezt technikailag úgy oldottuk meg, hogy a (többkarakteres) rövidítéseket megjelöltük egy speciális kezdő- (B) és végjellel (E). A lényeges pont az, hogy az n-gram gyakorisági lista származtatásakor (algoritmus 2. pont) a B..E közötti szakaszon nem vágunk, ezt a szakaszt „egy karakternek" tekintjük. A B*Sr*E*int* elemből képzett n-gramok így a következők lesznek: B*Sr*E, B*Sr*E*i*, B*Sr*E*in*, B*Sr*E*int*, *i*, *in*, *int*, *n*, *nt*, *t*. Az n-gramok hossz-számításakor természetesen a két speciális jelet figyelmen kívül kell hagyni.

Felmerülhet az olvasóban, hogy a rk() számítására bemutatott képlet hiányos. Abban az esetben, ha a rövidítés (jelentős számban) előfordul magyar nyelvű szövegben, azaz a rövidítéssel formailag megegyező eredeti szövegelemek elé a 2.2. részben írtak szerint védőjelet kellene tenni, hogy ne rövidítésként értelmeződjenek. Emiatt a rövidítési képesség csökkenne, a képlet kiegészülne az alább látható utolsó taggal:

$$\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] \cdot \mathrm{fq}(w) - \mathrm{fq}(r(w))$$

Ezt az utolsó tagot azonban – két okból – elhanyagoljuk. Egyrészt mert végül kizárólag kétjelű rövidítéseket tartalmaz az új szabályrendszer és ezekhez minden esetben találtunk olyan rövidítést, ami egyébként magyar szövegben nem vagy csak nagyon ritkán fordul elő, így védési igénye minimális; másrészt azért, hogy a „mit rövidítsünk" és a „mire rövidítsünk" kérdését valóban szétválaszthassuk egymástól.

Az algoritmusról szóló eszmefuttatás végén megjegyezzük, hogy a bemutatott „mohó" megközelítésről – miszerint ha egyszer kitaláltunk egy szabályt,

akkor azon többet nem változtatunk, sőt le is futtatjuk a teljes szógyakorisági listán, mielőtt továbblépnénk – nem bizonyított, hogy ténylegesen a maximális rövidítési képességű szabályrendszert eredményezi. Azonban mivel komolyan figyelembe kell vennünk a használhatósági szempontokat, és ennek következtében számos, a rk()-t befolyásoló manuális döntést hozunk, esetlegesen elfogadhatjuk a szuboptimális megoldást is kiindulópontként.

6. A korpusz és a rendszer futtatása

Eredetileg a Magyar Nemzeti Szövegtár [10] gyakorisági listájából terveztünk kiindulni. Péter Zsigmond (l. a Köszönetnyilvánítást) javaslatára, hogy a szöveganyagot jobban közelítsük a „vakos" nyelvezethez, végül jelentős mennyiségű ilyen szöveget is hozzávettünk. A *Vakok Világa* folyóirat 31 számának 180000 szónyi anyagát kombináltuk a Szövegtárral 4:1 arányú súlyozással a Szövegtár javára.

A futtatás során legelső lépésként alkalmazzuk a kis rövidírás (2.2. rész) szabályait. Ezeket a rövidítéseket olyan védelemmel látjuk el, ami biztosítja, hogy az eredeti, ismert formájukban megmaradjanak, ne rövidüljenek tovább. Maga az algoritmus tehát már eleve a kis rövidírással rövidített anyag alapján számított gyakorisági adatokat kapja meg.

Egy 60 szabályból álló rendszer előállításakor a futási idő nagyjából 10 perc. Ezt az elfogadható teljesítményt úgy érjük el, hogy a szógyakorisági listának csak az első 50000 bejegyzését vesszük. Ez a szelet a korpusz anyagának 85%-át tartalmazza, így nem torzítja jelentősen az n-gram gyakorisági adatokat, ugyanakkor a futási időt két nagyságrenddel (kb. századára) csökkenti.

7. A rövidítésjelek manuális kiválasztása a használhatósági megfontolások alapján

A használhatósági feltételeknek nagyon nehéz lenne teljesen automatizált módon megfelelni [4]. Szükséges ezért az automatikusan kialakított szabályrendszer interaktív módosítása, manuális véglegesítése szakértők bevonásával a használhatóság maximalizálása érdekében. Szigorúan automatikus úton történik tehát a fenti algoritmussal (5. rész) az épp aktuális (következő) legjobban rövidíthető elem meghatározása, illetve egy ajánlott, jó olvashatóságú rövidítést is automatikusan megad hozzá a rendszer. Ez manuálisan felülbírálható az alábbiak szerint.

7.1. Használhatósági követelmények

A 3. részben említett használhatósági követelmények közül a 2-4. ponttal foglalkozunk most részletesen.

A *jó olvashatóság* azt jelenti, hogy a rövidítések emlékeztessék az olvasót a rövidített szóra. Gyorsolvasáskor gyakran csak a szó első egy-két vagy utolsó

egy-két betűjét olvassuk el, fontos, hogy ezek a betűk az olvasó eszébe juttassák az egész szót. Tapasztalat szerint a jól olvasható rövidítés pozíciótól függetlenül mindig azonos jelentésű, a szó kezdő és záró betűjéből, illetve a szót alkotó jellegzetes mássalhangzóból áll. Ideális eset, mikor teljes szót/szóalakot rövidítünk (nem szórészletet), és a rövidítés a kezdő és a záró mássalhangzóból áll, ahogy ez a kis rövidírásban sok helyen látható: r(*mint*)=*mt*, r(*rövid*)=*rd*. Kiegészítő lehetőség, hogy adott esetben az is megfelelő, ha a jel *kinézete* emlékeztet arra a dologra, amire referál. Agglutináló nyelv lévén magyarban – néhány esettől (pl.: r(*hoG*)=*h*) eltekintve – nem tehetjük meg azt, hogy kizárólag teljes szavakat rövidítünk, mert ez nagyon alacsony rk()-t eredményezne. Ehelyett morfémákat (töveket és toldalékokat) rövidítünk, sőt bizonyos esetekben akár nagyon gyakori szótagokat, és ezeket egymás után illesztve kapjuk meg a szóalakokat. Az itt körvonalazódó elv úgy is megfogalmazható, hogy *„értelmest értelmesre”* rövidítsünk. Azaz lehetőleg értelmezhető elem legyen a rövidítendő, a használt rövidítésből pedig könnyen kikövetkeztethető legyen az eredeti elem.

A *jó felismerhetőség* követelményének akkor felelünk meg, ha tapintás útján könnyen azonosítható jeleket alkalmazunk a rövidítésekben. A 209. oldalon található 1. táblázatban szürkével megjelölt erős jelek felelnek meg ennek a követelménynek. Törekszünk rá, hogy minden rövidítés tartalmazzon erős jelet.

A *könnyű megtanulhatóság* azt jelenti, hogy a szabályok egyszerűek és jellegükben hasonlóak a kis rövidírásban meglévőkhöz, valamint, hogy kevés új szabályt hozunk létre. Az egyszerűség érdekében környezetfüggetlen szabályokat alkalmazunk, az új szabályok számát alacsonyan tartjuk. Ezt minden további nélkül megtehetjük, mert az architektúra lehetővé teszi, hogy a szabályrendszert a jövőben könnyen bővíthessük, ahogy erről a következő részben szót ejtünk.

7.2. A rövidítésjelek kiválasztása

A fenti megfontolások alapján az itt részletezendő kézi módosításokat végezzük az algoritmus eredményeként adódó rövidítendő elem – ajánlott rövidítés párokon. Lényegében minden egyes párról egyedileg döntünk, és utána futtatjuk tovább az algoritmust. Más szóval egyesével vesszük hozzá az új szabályokat a már meglévő szabályrendszerhez. Ez vonja magával azt a lehetőséget, hogy a most kialakított, kevés szabályt tartalmazó javaslat a jövőben bármikor ezen a módon tovább bővíthető igény szerint.

Sok esetben *felülbíráljuk* a rendszer által ajánlott rövidítést, amit a már felhasznált rövidítésjelek ismeretében a rövidítendő elem karaktereiből állít össze heurisztikák alapján. A javaslatban csak kétjelű rövidítéseket használunk. A legtöbb esetben könnyű kiválasztani egy olyan ritka kétjelű rövidítést, ami megfelelően illeszkedik a rövidítendőhöz, pl.: r([maGar])=[mG]. A rövidítésjel gyakoriságát a védési igény minimalizálása érdekében minden esetben ellenőrizzük és egy előre meghatározott küszöb (800/millió szó) alatt tartjuk. Korábban rövidített (szóvégi) elem hangrendi párjához törekszünk ugyanazt a rövidítést rendelni. Ilyen a javaslatban a *-ság/-ség* és a *-nak/-nek*.

Dönthetünk úgy, hogy az adott elemet csak bizonyos *pozícióban* (csak szó elején vagy végén) rövidítjük. Általában azért tesszük ezt, mert lényegében

csak abban a pozícióban fordul elő az adott rövidítendő. Törekszünk rá, hogy az előfordulási arány (például r([meg])=[mg] szó elején) lehetőleg 90% fölötti legyen, hogy ne sérüljön az adott szabály, mint aktuálisan legjobb szabály létjogosultsága. Ilyenkor lehetőségünk van arra, hogy nem általában ritka, hanem csak az adott pozícióban ritka, azaz *komplementer eloszlású* elemet válasszunk rövidítésnek, amint ez a r([lehet])=[lt] esetében meg is történt: a *lehet* elemet csak szó eleji helyzetben rövidítjük, a hozzá társított *lt* rövidítésjel nagyon gyakori ugyan, de lényegében sosem fordul elő szó elején.

Szükség esetén azt is kiköthetjük, hogy az adott elemet – tudva, hogy ezzel veszítünk a rövidítési képességből – *kizárjuk* a rövidíthető elemek köréből használhatósági megfontolások miatt. A javaslat készítése során a következő elemeket zártuk ki: *ala*, *ált*, *áro*, *eGe*, *eke*, *eket*, *ele*, *ere*, *erül*, *ete*, *hat*, *kez*, *kor*, *lat*, *leg*, *let*, *tal*, *tás*, *tele*, *ter*, *tés*, *val*. A puszta betűsorozatok mellett értelmes elemeket is látunk a listán. Ezek főként azért maradtak ki, mert túl ritkák a „kívánt" pozícióban: a *kor* szó végi és a *leg* szó eleji gyakorisága egyaránt 50% alatt marad.

7.3. Elvetett ötletek

A munka során számos ötletet, mely tovább növelte volna a rövidítési képességet, a használhatósági követelmények miatt elvetettünk. Ezek legtöbbször a 7.1. részben említett „értelmest értelmesre" elvet szegik meg, viszont rövidítési képességük jelentős (lenne). Érdemesnek tartjuk, hogy ezekről is említést tegyünk.

Az 5. részben írtak szerint magyarban a legnagyobb rk()-gel rendelkező karaktersorozat az *et*. Értelmetlen karaktersorozatokat azonban a nehéz olvashatóság miatt nem rövidítünk.

Annak ellenére, hogy a kis rövidírásban vannak nagyon hatékony szóközt elnyelő szabályok – a névelőket és a vesszőt érintő szabályokról van szó – ilyen szabályt sem alkalmazunk. A leggyakoribb szóvégi karakter (t) és az azt követő szóköz rövidítése kiemelkedően hatékony szabály: rk([t␣], [T])=1,2%. Az efféle szabályok nyilván rontják az olvashatóságot, mivel több szót egy hosszú egységgé vonnak össze, mégis esetleg érdemes volna megfontolni például a jelenleg *h*-ként rövidített *hoG* kettősponttal való rövidítését mindkét szóköz eltüntetésével: r([␣hoG␣])=[:]; vagy a határozatlan névelő szóközelnyelő rövidítését: r([eG␣])=[;]. E két szabály együttes rövidítési képessége 0,4%.

Mivel egyjelű rövidítéssel nem lehet megfelelni a jó olvashatóság fent leírt követelményének, egyjelű rövidítést egyáltalán nem alkalmazunk. Ugyanakkor a leghatékonyabb szabályok éppen a nagyon gyakori betűkapcsolatok egy karakterre való rövidítései lennének, ezek a szabályok kiemelten értékesek a rövidítési képesség szempontjából. Negyven szabállyal, mely *négy* darab ilyen rövidítést tartalmaz (*et*, *el*, *en*, *er*), közel 17%-os összesített rövidítési képesség érhető el, ami nagyon nagy mértékben meghaladja javaslatunk hatékonyságát. Felmerülhet például a következő elemek egyjelű rövidítése: *el*, *tt*, *meg*, *Ser*, *ás/és*, *eG*, az alábbi nagyon ritka jelek – közülük is főként az erős jelek – bevetésével: [H], [F], [@], [q], [=], [*], [T], [w]. Megjegyzendő, hogy az angol és a német

rövidírásban is alkalmaznak ilyen szabályokat, például r(⠊⠞[it])=⠭[x] az angolban (csak önálló szóként) [1, 103. oldal], r(⠍⠍[mm])=⠭[x] a németben [3]. Hasonló indokok miatt prefixet/posztfixet tartalmazó rövidítést sem használunk.

Ahogy említettük, az egyszerűség érdekében nem használunk környezetfüggő szabályokat. Ez abban nyilvánul meg, hogy a rövidítéseket a rövidítendő környezetétől függetlenül minden esetben alkalmazzuk, valamint, hogy a rövidítések szintén környezetüktől függetlenül mindig ugyanazt jelentik. Ez a döntés az algoritmus egyszerűsítését hozta magával. Amint láttuk, azt megengedjük, hogy bizonyos szabályok csak szó elején/végén legyenek alkalmazhatók, olyan azonban már nem fordul elő, hogy egy rövidítésjelnek két különböző jelentése legyen két különböző pozícióban. A német rövidírásban erre is van példa, érvényes például a következő szabály: r(⠊⠍⠍⠑⠗[immer])=⠭[x] [3], függetlenül attól, hogy az x a (az önálló szóként elő nem forduló) *mm* betűkapcsolat rövidítésére is használatos.

8. A javaslat

A magyar Braille-rövidírás megújítására vonatkozó javaslat a 3. táblázatban látható. A 33 szabály az 5. részben ismertetett algoritmus lefuttatásával, a kimenet 7. rész szerinti manuális finomhangolásával jött létre, megfelelő kompromisszumot alakítva ki a minél nagyobb rövidítési képesség és a használhatóság szempontjai között.

9. A javaslat kiértékelése

9.1. A rövidítési képesség mérésének módszertana

A dolgozatban korábban említett empirikus méréseket, és a javaslat teljesítményének kiértékelését is az itt ismertetendő módon végeztük. A 4. táblázatban látható három tesztfájllal dolgoztunk.

Minden szöveget először a 2.1. részben ismertetett „egy hang egy karakter" formára hoztunk. Megállapítottuk az adott tesztfájl karakterszámát, alkalmaztuk először a kis rövidírás szabályrendszerét, majd az épp mérendő szabályrendszert, meghatároztuk, hogy hány százalékkal csökkent a karakterszám, majd a három tesztfájlra kapott értéket átlagoltuk. A szóközök és az újsor-karakterek is beleszámítottak a karakterszámba. A 2.2. részben írtaktól kis mértékben eltérve a kis rövidírás egyjelű szórövidítéseit két kivétellel kizárólag önálló szóként alkalmaztuk, a *Cak*-ot, az *után*-t és a kétjelű szórövidítéseket viszont minden előfordulásukban. A kis rövidírás esetén nem alkalmaztuk a védőjelet, mivel azt tapasztaltuk, hogy lényegében nincs használatban; az új szabályok esetén viszont minden esetben alkalmaztuk, a kapott értékek tehát az új szabályok védési igényével csökkentett értékek. A mérés során minden rövidítést karakterszintű védelemmel láttunk el, megakadályozva ezzel, hogy rövidítés részletét véletlenül továbbrövidítsük, kivéve persze az egyetlen tényleges továbbrövidítést, a r(*Serint*)=*St* – r(*Ser*)=*Sr* esetét.

3. táblázat. A magyar Braille-rövidírás megújítására vonatkozó javaslat avagy az új magyar Braille-rövidírás. A szabályok előállítási sorrendben vannak feltüntetve. Jelöljük, ha a rövidítést a jel kinézete miatt választottuk, valamint ha csak bizonyos pozícióban rövidítjük az adott elemet. A komplementer eloszlás fogalmát (8., 21. és 32. szabály) a 7.2. részben vezetjük be. A csúsztatott jel fogalma a 2.1. részben található.

	rövidítendő	rövidítés	megjegyzés
1.	 [meg]	[mg]	szó elején (98%)
2.	 [ett]	[eT]	= ott
3.	 [Ser]	[Sr]	
4.	 [ott]	 [oT]	= ett 
5.	 [maGar]	 [mG]	
6.	 [jelen]	[jn]	
7.	 [ség]	 [sg]	= ság
8.	 [lehet]	 [lt]	szó elején (97%) *lt* komplementer: gyakori, de nem szó elején
9.	 [vezet]	[vz]	
10.	 [nek]	[nx]	= nak ; szó végén (83%)
11.	 [köz]	[kz]	
12.	 [nak]	 [nx]	= nek  ; szó végén (96%)
13.	 [fel]	 [fl]	szó elején (86%)
14.	[Serint]	[St]	szó elején (95%)
15.	 [rend]	 [rH]	csúsztatott *d*  *rd* foglalt a kis rövidírásban (*rövid*), *rn* túl gyakori
16.	 [ember]	 [wr]	jel kinézete
17.	 [kormáN]	 [kN]	
18.	 [ellen]	[oó]	jel kinézete
19.	[elnök]	[eö]	
20.	[elő]	[eő]	
21.	[tart]	[tt]	szó elején (78%) *tt* komplementer: gyakori, nem szó elején
22.	[áll]	[Ll]	jel kinézete (egyéb ötlet: [áy])
23.	[mond]	[mH]	csúsztatott *d* *md* foglalt a kis rövidírásban (*mind*)
24.	[támogat]	[tg]	
25.	[ért]	[éT]	
26.	[ság]	[sg]	= ség
27.	[következ]	[kv]	
28.	[akkor]	[ao]	
29.	[budapest]	[bp]	
30.	[kerül]	[eü]	
31.	[történ]	[öé]	(egyéb ötlet: [Tn])
32.	[több]	[tb]	szó elején (95%) *tb* komplementer: gyakori, de nem szó elején (egyéb ötlet: [t=])
33.	[kapColat]	[kC]	

4. táblázat. A rövidítési képesség méréséhez használt tesztfájlok.

megnevezés	méret
egy zenei híreket tartalmazó fájl az MVGYOSZ-ből	11000 szó
a *Vakok Világa* 31 számának anyaga	180000 szó
Mikszáth Kálmán: *Szent Péter esernyője* c. regénye	53000 szó

9.2. A kiértékelés eredménye

A 8. részben ismertetett javaslatunk kiértékelésének eredménye az 5. táblázatban látható. A javaslat kevés szabályt tartalmaz, könnyen megtanulható. A rövidített szöveg olvashatósága kiváló, felismerhetősége is megfelelő. A rendszer rövidítési képessége kielégítő: a kis rövidírás által képviselt rövidítési képességet harmadával megnöveltük.

A 13,3%-os eredmény az angol Braille-rövidírás közel 20%-os rövidítési képességével [11] összevetve nem tűnik soknak. Érdemes ugyanakkor látni, hogy az angol rendszer majd 200 szabályt tartalmaz, és számos, a 7.3. részben leírt eljárást kiterjedten alkalmaz. Nehezebben tanulható, olvashatósága jelentősen rosszabb. Javaslatunkban az egyes szabályok átlagos rövidítési képessége 0,1%. Ezt is érdemes összevetni a 7.3. részben említett példák rövidítési képességével.

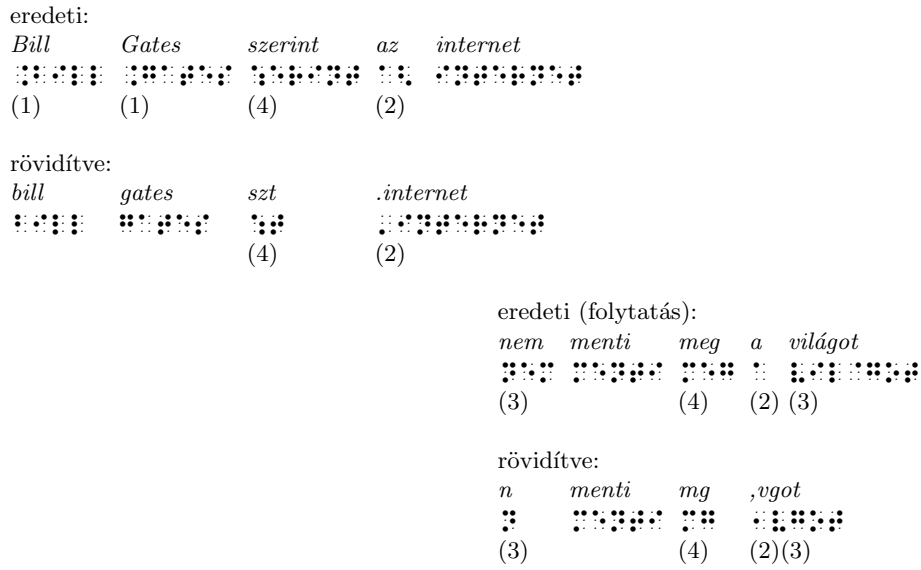
5. táblázat. A javaslat 33 szabályának összesített rövidítési képessége.

tesztanyag	kis rövidírás rk()	új rövidírás rk()	Δrk()	Δrk() %
zenei hírek	9,5%	12,5%	+3,0%	
Vakok Világa	9,5%	13,9%	+4,4%	
Szent Péter...	10,7%	13,5%	+2,8%	
átlag	**9,9%**	**13,3%**	**+3,4%**	**+34%**

Az új rövidítendő elemek ábécérendes listája a következő: *akkor*, *áll*, *budapest*, *ellen*, *elnök*, *elő*, *ember*, *ért*, *ett*, *fel(-)*, *jelen*, *kapColat*, *kerül*, *kormáN*, *következ*, *köz*, *lehet(-)*, *maGar*, *meg(-)*, *mond*, *-nak/-nek*, *ott*, *rend*, *ság/ség*, *Ser*, *Serint(-)*, *támogat*, *tart(-)*, *több(-)*, *történ*, *vezet*

9.3. Példák

A 3. ábrán egy példamondaton mutatjuk be az új rövidírás működését, a 6. táblázatban pedig jellegzetes új rövidítéseket tartalmazó szavak láthatók.



3. ábra. Az új rövidírás működése egy példán. A kis rövidírás (1) elhagyja a nagybetűjelet; (2) összevonja és rövidíti a határozott névelőket: (3) szórövidítéseket alkalmaz (*nem*, *világ*). Ez után következnek (4) az új rövidítő szabályok (*Serint*, *meg*). A mondat hossza 55 karakterről 40 karaterre csökkent, ami 27,3%-os rövidülést jelent.

6. táblázat. Jellegzetes új rövidítéseket tartalmazó szavak.

rövidítendő	rövidítés	Δhossz		
gyakran előforduló rövidített szavak				
[között]	[között]			
[eGSer]	[eGSr]			
[lehetett]	[lTeT]			
nagy arányban rövidülő szavak				
[kapColatban]	[kCb]	-8	72%	
[következő]	[kvő]	-6	66%	
[maGarorSág]	[mGog]	-6	60%	
három rövidítést tartalmazó, sok karakterrel rövidülő szavak				
[kapColatrendSer]	[kCrHSr]	-9	60%	3új
[jelentőségteljes]	[jntősgts]	-8	50%	1r+2új
[boldogságban]	[bgsgb]	-7	58%	2r+1új

10. Konklúzió

Az eredeti alapötlet bevált, a közel maximális rövidítési képességgel bíró, ugyanakkor kényelmesen használható új magyar rövidírás szabálykészlete félautomatikus módon, a rövidítendők korpuszvezérelt meghatározásával és a rövidítések kézi finomításával előállítható.

Az összesített rövidítési képesség 13,3%. Ez jelentős – több mint 30 százalékos – növekedés a kis rövidítés 9,9%-os hatékonyságához képest, a használhatósági követelmények miatt azonban ez jóval alacsonyabb, mint a lehetséges elvi maximum. A módszerből adódóan a szabályrendszer a jövőben könnyen bővíthető. Jóval több és/vagy nehezebben olvasható szabállyal természetesen megközelíthető akár a 20% is. Esetünkben az volt a koncepció, hogy az új magyar rövidírás bevezetésének megkönnyítése érdekében a változtatás, bővítés mértékével nagyon óvatosak voltunk, és a kompromisszumos javaslat kialakítása során nagyobb hangsúlyt fektettünk a használhatóságra, mint a rövidítési képesség minden áron való növelésére. A rövidítendő elemeket meghatározó korpuszvezérelt algoritmusnak köszönhetően az adott feltételek mellett az objektíve legjobb rendszert hoztuk létre.

Ha összevetjük a kis rövidírásban lévő kétjelű rövidítések (210. oldal: 0,05%), illetve az új kétjelű szabályok egy szabályra eső rövidítési képességét (220. oldal: 0,1%), akkor azt látjuk, hogy a korpuszvezérelt módon létrehozott rendszer még úgy is *kétszeres* teljesítményre képes az intuíció, illetve hagyomány talaján álló rendszerrel szemben, hogy már eleve jelentősen rövidített szövegen kell dolgoznia.

Konklúzióként levonhatjuk tehát – és ez érvényes lehet a különféle annotációs vagy ontológiaépítési feladatoktól, a szótárkészítésen át akár egyes elméleti nyelvészeti kérdésekre is –, hogy ha valamit meglévő (gyakorisági) adatokból automatikusan származtatni tudunk, akkor nem érdemes intuitív megközelítést alkalmazni. Vagy másképp fogalmazva: érdemes az intuíciót bizonyos adatvezérelt módszerekkel legalábbis kordában tartani.

A rendszer alkalmas a bevezetést előkészítő közvetlen vakok általi tesztelésre, bízom benne, hogy találkozhatunk majd vele Braille-nyomtatványokban vagy akár a közterek Braille-feliratain.

Köszönetnyilvánítás

A projekt munkálatai kapcsán nagy köszönettel tartozom a Magyar Vakok és Gyengénlátók Országos Szövetsége Braille-bizottsága részéről *Péter Zsigmond*-nak, aki a Braille-írás és a Braille-rövidírások szakértő ismerőjeként a munka számos pontján volt segítségemre: megismertetett a Braille-írással, ellátott szakirodalommal, bevezetett a használhatósági megfontolások rejtelmeibe, a javaslat kialakítása során pedig közösen hozhattuk meg a rendszert érintő konkrét használhatósági döntéseket.

Hivatkozások

1. Simpson, Ch., ed.: The Rules of Unified English Braille. Version I. Round Table on Information Access for People with Print Disabilities Inc., Australia (2010)
2. Freud, E.: Leitfaden der deutschen Blindenkurzschrift: Teil 2. Verlag der Deutschen Blindenstudienanstalt, Marburg (1973)
3. `fakoo.de`: Einführung in die deutsche Braille-Kurzschrift `http://www.fakoo.de/kurzbraille.html`
4. Sass, B.: Az új magyar Braille-rövidírás korpuszvezérelt kialakításának lehetőségei. In: IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2013), SZTE, Szeged (2012) 348–350
5. Flamich, M., Hoffmann, R.: A tapintható írásrendszerek történeti áttekintése. Iskolakultúra **20**(1) (2010) 3–17
6. Görgényi, M., ed.: A magyar pontírás. Teljesírás. Pécs (1998)
7. Görgényi, M., ed.: A magyar pontírás. Rövidírás. MVGYOSZ, Budapest (2001)
8. Bogart, D.: Unifying the English Braille Code. Journal of Visual Impairment & Blindness **103**(10) (2009) 581–583
9. Arató, A.: A BraiLab beszélő számítógépcsalád. Kandidátusi értekezés. (1992)
10. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
11. Durre, I.K.: How much space does Grade 2 Braille really save? Journal of Visual Impairment & Blindness **90**(3) (1996) 247–251

VI. Információkinyerés és -visszakeresés

Gazdasági hírek tartalmának feldolgozása banki előrejelző rendszer támogatásához

Tarczali Tünde, Skrop Adrienn, Mokcsay Ádám

Pannon Egyetem, Rendszer- és Számítástudományi Tanszék
8200 Veszprém, Egyetem u. 10.
{skrop,tarczali}@dcs.uni-pannon.hu
adam@mokcsay.hu

Kivonat: Kutatásunk célja egy olyan „early warning" mechanizmus és alkalmazás kifejlesztése, amely a weben megjelenő „szoft" információk feldolgozásán alapulva pénzügyi intézetek számára kockázat előrejelző szolgáltatást nyújt. A rendszer feladata a vizsgálandó alanyokkal kapcsolatos hírek, úgynevezett szoft információk keresése a weben, a talált hírek vizsgálata szövegbányászati eszközökkel, jellemzőik azonosítása és ezek alapján előre meghatározott kockázati kategóriákba sorolása. Cikkünkben ismertetjük a tervezett rendszer felépítését és az elkészült modulok működését.

1 Bevezetés

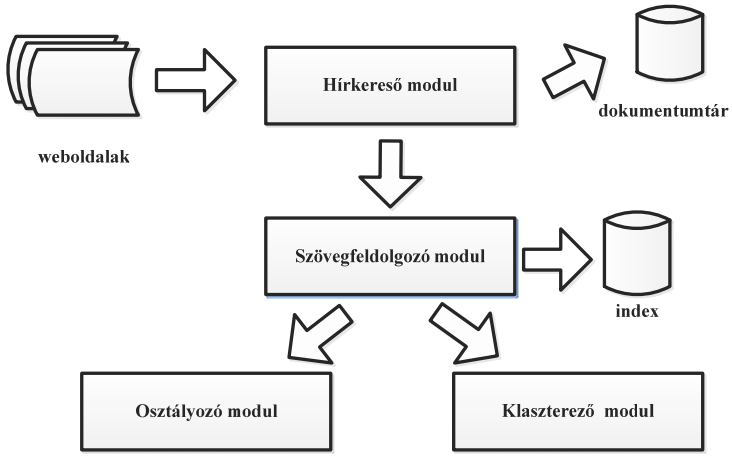
A hírelemzés szöveges hírek különböző kvalitatív és kvantitatív tulajdonságainak mérésével, elemzésével foglalkozik. Ilyen tulajdonságok például a szentiment, a relevancia és az újdonság. A hírelemzés magában foglalja mindazon technikákat és módszereket, melyek segítségével nyilvános információforrások feldolgozhatók, osztályozhatók [3]. A hírelemzés egyik fontos területe a gazdasági hírek elemzése, amely elsősorban azzal foglalkozik, hogy különböző gazdasági hírekre mikor és miként kell reagálnia a piacnak ahhoz, hogy a profitot növelni tudják.

A 2008 óta tartó és a pénzügyi szektort az ügyfeleik helyzetén keresztül is érintő pénzügyi-gazdasági válság középpontba helyezte a hitelkockázat minél hatékonyabb kezelésére és esetlegesen a kockázati tényezők előremutatására irányuló alkalmazások kifejlesztését. Jelen kutatás célja egy olyan automatizált kockázat előrejelző (early warning) módszer kifejlesztése, amely múltbéli információkból építkezve próbálja időben felismerni és jelezni az ügyfelek nem teljesítési kockázatát. A rendszer sajátossága, hogy a bankokban szokásos belső minősítésen alapuló módszer helyett az ügyfelek fizetésképtelenségre vonatkoztatott kockázatát a weben róluk megjelenő szoft információk szemantikai elemzésével jelzi előre.

2 A rendszer felépítése és működése

A tervezett rendszer felépítését az 1. ábra szemlélteti. A Hírkereső modul feladata a figyelendő alanyokra – ügyfelekre – vonatkozó, időzített keresések futtatása a weben.

A Hírkereső modul két funkciót lát el: egyrészt a múltbéli céges információk alapján mintákat gyűjt az osztályozáshoz használandó tanító minták meghatározásához, másrészt jelzi, ha egy ügyféllel kapcsolatban új hír jelent meg a weben. A weboldalak feldolgozását a Szövegfeldolgozó modul végzi. A modul feladata az internetes hírek előfeldolgozása, vektortér modellbeli reprezentálása [6], korpusz előállítása és a híreknek a tartalmazott szavak alapján történő kategorizálásának támogatása. A szemantikailag hasonló dokumentumok klaszterezését a Klaszterező modul végzi. A klaszterezés az AI^{2}R adaptív klaszterező eljárás segítségével történik [1]. A Hírkereső modul által szolgáltatott új hírek kockázati kategóriákba sorolása az Osztályozó modul feladata. Az osztályozásra naiv vektortér alapú módszert alkalmazunk [2], így a hasonlóság mértékének változtatása kevésbé számításigényes.



1. ábra. A rendszer felépítése.

2.1 Hírkereső modul

A Hírkereső modul feladata releváns, nem strukturált szoft információk keresése a weben. A modul megvalósítása hagyományos kulcsszavas metakeresővel történt. A metakereső olyan, webszervereken keresztül elérhető szoftver, mely egy adott kérdést elküld több webkeresőnek, összegyűjti és – valamilyen eljárással – egyesíti az eredményeket. A metakereső legfőbb előnye, hogy több kereső érhető el egyetlen, egyszerű interfésszel.

A megvalósított metakereső a Google és a Bing találati listáját használja fel, kezdőképernyőjét a 2. ábra mutatja. A metakereső egy weblapon keresztül érhető el, amelyet PHP motor generál, ezzel biztosítva annak dinamikus mivoltát, hiszen a működés során adatbázissal dolgozik a rendszer. Az adatbázist MySQL program kezeli, a rendszer pedig egy Linux alapú szerveren helyezkedik el. A felhasználó több paramétert képes megadni egy kereső kifejezés felvételénél, amelyet a program a beépít az egyes keresők felé intézett kérésbe. Ezekkel a paraméterekkel a keresés időzítése állítható be. Lehetőség van kereső-kifejezések importálására is, ebben az esetben egy XML kiterjesztésű fájlt vár a rendszer bemenetként.

2. ábra. Hírkereső modul kezdőoldal

A metakereső két alapvető tulajdonsága, hogy a keresés előre meghatározott kulcsszavak alapján történik, valamint a metakereső által visszaadott találati lista elemzése, a releváns oldalak végső ellenőrzése szakértő által történik. A Hírkereső találati listáját a 3. ábra szemlélteti.

3. ábra. Hírkereső modul találati lista

Minden előre definiált kereső kérdéshez meghatározásra kerül egy találati lista. A rendszer feladata, hogy a találati listát a beállított időzítésnek megfelelően frissítse és jelezze új, potenciálisan releváns találatok megjelenítését. A program lehetőséget biztosít az eredmények exportálásra, amely egy XML kiterjesztésű fájlt eredményez. A Kezelő menüpont segítségével a korábbi beállításokat módosíthatjuk.

2.2 Szövegfeldolgozó modul

A modul feladata az interneten fellelhető információk feldolgozása és gazdasági felszámolásra utaló releváns szavak kiemelése. Bemenetként a modulban megadhatóak hírekre mutató internetes linkek, vagy a Hírkereső modul által kimenetként szolgáltatott XML kiterjesztésű fájl, amely linkgyűjteményeket tartalmaz, akár meghatározott csoportokat is alkotva. Ennek segítségével egyszerre több, a szakértő által kiválasztott cikk együttes vizsgálatára nyílik lehetőség. A beolvasás lehetőségeit szemlélteti a 4. ábra.

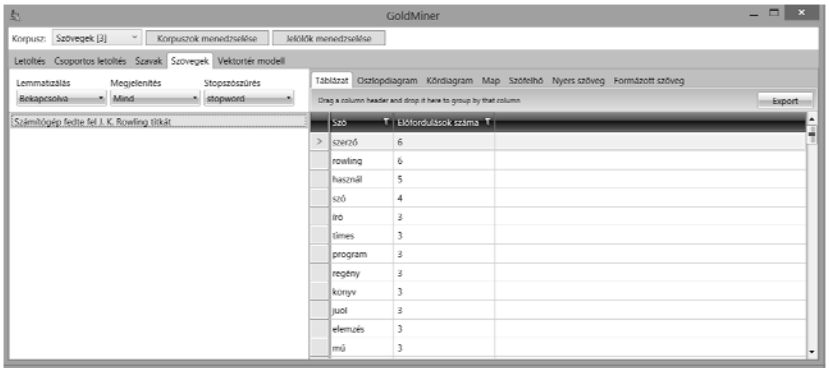
4. ábra. A hírek letöltése link megadásával

Az ábrán látható módon a cikkre mutató link megadásával a szoftver az internetről letölti a cikket és ezután történik meg annak feldolgozása. A hírek letöltésére automatizált letöltőket építettünk be a szövegelemző szoftverbe. Nem volt célunk saját letöltő készítése, hiszen a projekt céljának eléréséhez megfelelőek voltak a beépített automatikus letöltők. Egy linken található cikk betöltése mellett – a munka megkönnyítésére – lehetőség van több cikk egyidejű letöltésére is. Ennek megvalósítására egy XML file-t hoztunk létre, amely a következő formátumban tartalmazza a cikkek elérhetőségét:

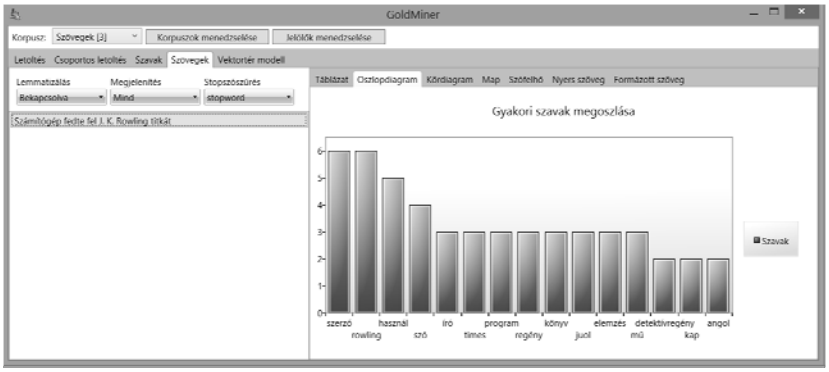
```
<?xml version="1.0" encoding="ISO-8859-2"?>
<download>

<corpus name="Első">
 <article href="http://link.url" />
 <article href="http://link.url2" />
 <article href="http://link.url3" />
</corpus>

<corpus name="Második">
 <article href="http://link.url4" />
 <article href="http://link.url5" />
 <article href="http://link.url6" />
</corpus>

<download>
```

A szövegek mondatokra, szavakra történő tagolásával (tokenizálással), valamint a stopszavak szűrésével végrehajthatóak olyan vizsgálatok, amely alapján a cikkekre vagy cikkgyűjteményekre jellemző szavakat, szóösszetételeket kaphatunk meg. A program beépített stopszótárral rendelkezik. A program első indításakor az adatbázis feltöltődik a stopszavak listájával. Ezekre a szavakra a program „Stopszó" címkét aggat. A stopszavak megadására külön listában van lehetőség, így a felhasználó maga is meghatározhatja ezeket. A projektünk témája indokolja, hogy jelen esetben arra keressünk választ, hogy az egyes cikkekben milyen szavak utalhatnak a vállalatok csőd közeli voltára. A programban lehetőség van a felhasználó által karban tartott jelölők tárolására, amelyekhez tetszőleges számú és nevű címke hozható létre. Ezen címkék hozzáadása történhet egy olyan szövegfájl alapján is, mely tartalmazza a jelölni kívánt szavakat.

Mivel a program a gazdaság képviselőinek készült, ezért szükség volt a statisztikai adatok grafikus megjelenítésére, amely segíti a szakértői értékelést. Erre mutatnak példát az 5-8. ábrák.



5. ábra. Táblázatos vizualizáció bekapcsolt lemmatizálás mellett

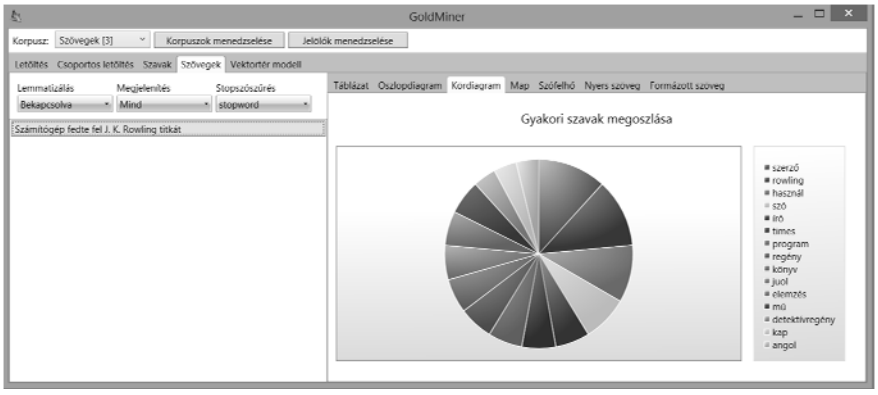


6. ábra. Cikkek vizuális elemzése oszlopdiagramon

Az egyenként történő statisztikai feldolgozás a tokenizálás után történhet a lemmák vizsgálatával, illetve anélkül. Itt egy beépülő modul segítségével vizsgáljuk a szavak

szótöveinek előfordulási gyakoriságát. A további feldolgozást segítendő lehetőség van a táblázat exportálására .xlsx formátumban.

A cikkekben található szavak előfordulási gyakoriságának elemzésére illetve a kimutatások elkészítéséhez a felület biztosít egy, az Excel programból jól ismert oszlopdiagramot, amely a vízszintes tengelyen tartalmazza a gyakran előforduló szavakat. Az oszlopok magassága a szó előfordulási gyakoriságával arányos.



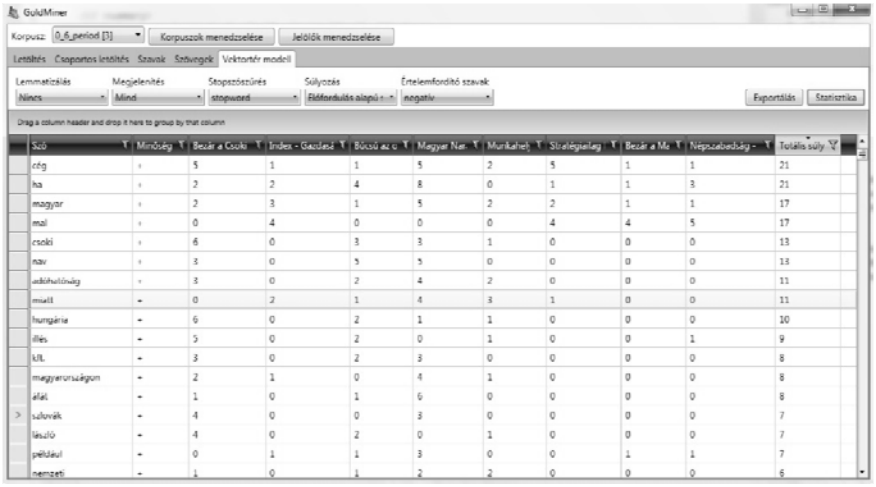
7. ábra. Cikkek vizuális elemzése kördiagramon

Az oszlopdiagram mellett az elemzés könnyítését szolgálja a kördiagrammal történő megjelenítés is. Itt a körcikkek vastagsága arányos a szavak előfordulási gyakoriságával.

8. ábra. Cikkek vizuális elemzése szófelhővel

Egy érdekes vizuális megjelenítést célzó ábra a szófelhő. Az interneten a cikkek megjelölésére gyakran használt eszköz a címkézés. A címkék előfordulásának gyakoriságát illetve a cikkek olvasásának gyakoriságát gyakran mutatják címkefelhővel. Ezt a megjelenítési módszert alkalmaztuk a szavak gyakoriságának bemutatására. A nagyobb betűvel megjelenő szavak jelentik a szövegben gyakran előforduló szavakat.

A program a szövegeket szöveggyűjteményekben, korpuszokban tárolja. A szoftver a karbantartott korpuszokból képes vektortér modell előállítására. Ennek szükségességét az adja, hogy a cikkek elemzése cégekhez és a tanító fázisban a csőd közeli állapothoz viszonyított időszakokra vonatkoztatva történik. Az alábbi kép mutatja a program működésének azt a fázisát, ahol egy cikkcsoportra vizsgáljuk a szavak előfordulását.



9. ábra. Vektortér modell kialakítása a kiválasztott cikkekre

A vektortér modellben [5] mindazon lehetőségek megvannak, amelyek az egyes cikkek elemzésénél is segítségünkre lehetnek. A kanonikus alak megtalálására alkalmazható eljárás például a szavak csonkolása. Ekkor szótőként általában nem a szótári szóalakot kapjuk, ám a legtöbb esetben ez is kellően pontos. Léteznek egyéb szótár alapú algoritmusok is. Ilyen algoritmus pl. a Porter féle algoritmus, Lovins-tövező, vagy a Snowball alapú magyar tövező. A szótövezést a Hunstem program végzi [4]. A program felismeri a szavak töveit, ezzel lehetővé téve a szótő szerinti csoportosítást és a generált vektortér modell dimenziószámának redukálását. A szavak szótövezése mellett megvalósításra kerültek olyan súlyozások, amelyek a különböző vizsgálatokat segítik. A következő súlyozási módszereket [6] implementáltuk:

- bináris
- előfordulás alapú
- logaritmikus
- gyakoriság alapú
- TF-IDF

Ezeken kívül lehetőség van az értelemfordító szavak vizsgálatára is. Két szó távolságban vesszük figyelembe azt, hogy a cikkekben megjelenő értelemfordító szavak negatív értelmet adnak egyes kifejezéseknek.

2.3 Klaszterező modul

A szemantikailag hasonló dokumentumok klaszterezését a Klaszterező modul végzi. Klaszterezés során a dokumentumokat – általában – diszjunkt halmazokba csoportosítjuk. Minden klaszter – bizonyos értelemben – hasonló dokumentumokból áll. A modul célja az azonos kockázati kategóriát képviselő hírek egy csoportba sorolása.

A különböző klaszterezési technikák közül a gazdasági területet igényeihez leginkább illeszkedő módszert kellett meghatározni. Az a fontos igény került figyelembevételre, hogy ne csak az azonos kifejezéseket tartalmazó cikkek, hanem egy cikkhez szemantikailag hasonló tartalmúak is egy klaszterbe kerüljenek. Ez az elvárás indokolta, hogy az interakciós információ-visszakereső I^2R (Interaction Information Retrieval) technikát választottuk.

Az I^2R matematikai modellje a mesterséges neuronhálózat alapvető állapotegyenletén alapszik. Eszerint a dokumentumok azonosíthatóak egy neuronhálózattal, ahol az egyes dokumentumok egy-egy neuronnak felelnek meg, amelyek képesek különböző szintű aktivitást produkálni. Egy új dokumentum szintén egy neuronnak felel meg, amely beépül a hálózatba – mint egy új objektum – és így a hálózat részlegesen megváltozik: új kapcsolatok alakulnak ki az új és az eredeti objektumok között, továbbá az eredeti hálózatban kialakult kapcsolatok egy része módosulhat. Ez a hatás indítja el a klaszterezési folyamatot.

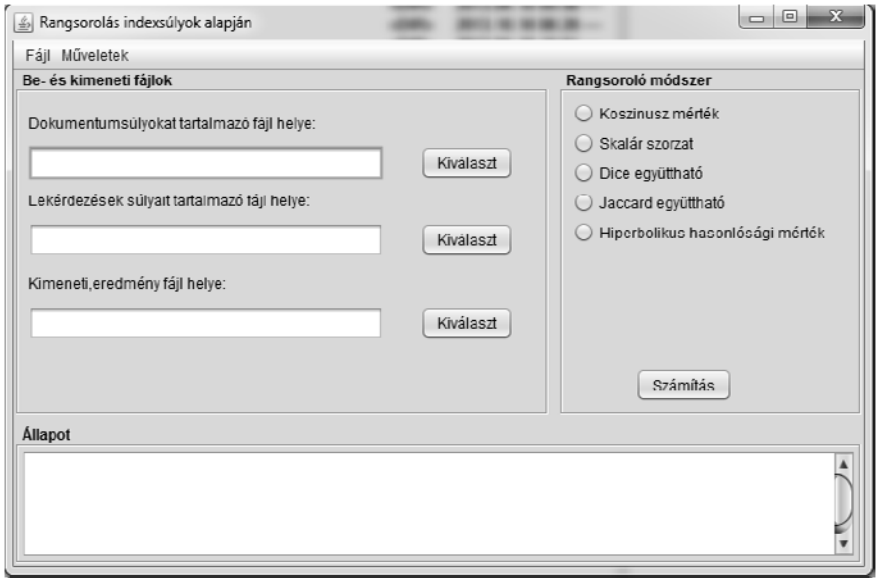
2.4 Osztályozó modul

A Hírkereső modul által szolgáltatott új hírek kockázati kategóriákba sorolása az Osztályozó modul feladata. A Klaszterező modul által meghatározott csoportok nem jellemezhetők a hagyományos értelemben vett címkékkel, hanem kockázati kategóriákat jelölnek, ezért az osztályozásra naiv vektortér alapú módszert alkalmazunk.

Mind a klaszterekben szereplő cikkeket, mind az új híreket a szentiment elemzés során definiált vektortérbeli vektorokként ábrázoljuk. Az új hírek klaszterbe sorolása a vektortérben használt hasonlósági mérték segítségével történik. A módszer azon alapul, hogy az új hírt reprezentáló vektor és egy klaszterbeli vektor elég közel vannak-e egymáshoz. A vektorok hasonlóságát különböző hasonlósági mértékek segítségével lehet mérni.

A vektortér modellt hagyományosan euklideszi térben definiálják. Az Osztályozó modulban implementálásra kerültek az euklideszi tér szokásos hasonlósági mértékei, mint a belső szorzat, a koszinusz mérték, a Dice együttható és a Jaccard együttható. A hagyományos modell mellett implementálásra került a hiperbolikus információ-visszakereső modell is, melynek lényege, hogy a benne alkalmazott hasonlósági

mérték a Cayley-Klein hiperbolikus távolságból származik. Gyakorlati tesztsorozatok segítségével fogjuk meghatározni, hogy melyik módszer alkalmas gazdasági hírek osztályozására. Az osztályozó modult a 10. ábra szemlélteti.



10. ábra. Osztályozó modul

3 A kutatás eredményei

Kutatás-fejlesztési feladatunk célja az interneten elérhető gazdasági tartalmú információk, hírek megkeresése és feldolgozása, a releváns tartalom kinyerése és a cikkek osztályozása. A kutatás első lépéseként meghatározásra kerültek azok a jellemzően szöveges információk, amelyek valamely negatív esemény bekövetkezését jelezhetik. A múltbéli céges információk elemzésére a kutatáshoz rendelkezésre áll a Dun&Bradstreet teljes magyar sokaságra vonatkozó minta adatbázisa. Szakértői segítséggel kiválasztásra kerültek azok cégek, illetve ezután azok a rájuk vonatkozó cikkek és időszakok, amelyek elemzése a készített alkalmazással folyamatosan történik. A meghatározott információk alapján lefolytattuk azokat az internetes kereséseket, amelyek alapján a cikkek szakértők általi szűrésével előállt az a releváns információkat tartalmazó cikkhalmaz, amelynek feldolgozásával a csőd előrejelzése támogatható. Ezen adatok alapján webes kereséssel felállítjuk azon tanító halmazokat, amelyek alkalmazásával a megjelenő cikkekről eldönthető, szolgáltatnak-e információkat a cégek pénzügyi helyzetével kapcsolatban.

Köszönetnyilvánítás

A publikáció az Európai Unió, Magyarország és az Európai Szociális Alap társfinanszírozása által biztosított forrásból a TÁMOP-4.2.2.C-11/1/KONV-2012-0004 azonosítójú „Nemzeti kutatóközpont fejlett infokommunikációs technológiák kidolgozására és piaci bevezetésére” című projekt támogatásával jött létre.

A kutatás a GOP-1.1.1-11-2011-0045 azonosítójú EWS – Adat- és folyamatbányászati algoritmusokon alapuló automatizált kockázat előrejelző rendszer prototípusának fejlesztése pénzügyi intézetek számára című projekt támogatásával valósult meg. A cikk tartalma kizárólag a szerzők felelőssége, és nem feltétlenül tükrözi a támogatók álláspontját.

Hivatkozások

1. Dominich, S.: Connectionist interaction information retrieval. Information processing & management. Vol. 39(2) (2003) 167–193.
2. Góth, J., Skrop, A.:Varying retrieval categoricity using hyperbolic geometry. Information Retrieval. Vol. 8(2) (2005) 265–283
3. Mitra, G., Mitra, L.: The Handbook of News Analytics in Finance. John Wiley & Sons (2011)
4. Németh, L.: A Szószablya fejlesztés. 5th Hungarian Linux Conference (2003)
5. Subecz, Z.: Információkinyerés természetes nyelvű szövegekből. Szolnoki Tudományos Közlemények XV., Szolnok (2011)
6. Tikk, D. (szerk.): Szövegbányászat. Az informatika alkalmazásai sorozat. ISBN 978-963-9664-45-6. (2007)

Igei események detektálása és osztályozása magyar nyelvű szövegekben

Subecz Zoltán, Nagyné Csák Éva

Szolnoki Főiskola
5000 Szolnok, Tiszaligeti sétány 14.
{subecz,csak}@szolf.hu

Kivonat: Jelen tanulmányunkban bemutatjuk megközelítésünket, amely igei eseményeket képes detektálni és osztályozni magyar szövegeken. Első lépésben azonosítottuk a többszavas főnévi+igei kifejezéseket. Majd detektáltuk az eseményeket, a detektált eseményeket pedig osztályokba soroltuk. A feladatok mindegyikéhez gazdag jellemzőkészleten alapuló bináris osztályozót használtunk. Az osztályozót kiegészítettük szabályalapú módszerekkel is. Módszerünket a Szeged Korpusz öt különböző doménjén is megvizsgáltuk, és hasonlósági gráfok segítségével elemeztük a részkorpuszok kapcsolatát.

1 Bevezetés

Munkánkban természetes szövegekben előforduló **események detektálásával és osztályozásával** foglalkoztunk. Az események **detektálásának** a feladata az esemény-előfordulások azonosítása a szövegekben, az **osztályozással** pedig a megtalált eseményeket előre meghatározott kategóriákba rendeljük. Esemény-előfordulásnak tekintünk minden olyan kifejezést, ami olyan eseményt vagy állapotot jelöl, amit egy adott időponthoz, vagy intervallumhoz tudunk kapcsolni.

Noha az igéken kívül lehetnek események más szófajú szavak is (pl. főnevek, igenevek stb.), a szövegekben a *legtöbb esemény igékhez kapcsolódik*, ezért jelen munkánkban az **igei eseményekkel** foglalkoztunk. Az igék közül azonban *nem mindegyik* tekinthető eseményjelölőnek (például: van, volt, lesz, marad, segédigék), így ezek kiszűrésére külön figyelmet kell fordítani. Vannak olyan események, amelyeket két szóval fejezünk ki (pl. döntést hoz), ezek szintén külön kezelést igényelnek. Több munka is foglalkozott már részletesen a *többszavas igei kifejezésekkel* [8, 6, 7], ezek eredményeit felhasználtuk.

A feladat a szövegekben megtalálható események detektálása és osztályozása. Munkánkban elsősorban az **igei egy- és többszavas eseményekkel** foglalkozunk. A rendszer bemenete egy tokenszinten címkézett tanító korpusz. A feladatot három részre osztottuk. A szövegekben először a **több szavas főnév + igei kifejezéseket** válogattuk ki, majd a maradék igékből **detektáltuk az eseményeket**. A megtalált eseményeket ez után **osztályoztuk**. A feladat megoldásához *statisztikai és szabály alapú módszereket* is alkalmaztunk.

2 Kapcsolódó munkák

Sok kutatás foglalkozik az események detektálásával. A legtöbb munkában csak adott eseményekkel foglalkoznak (például üzleti), vagy még azon belül is csak kiemelt eseményekkel (például cégfelvásárlás). Jelen munkánkban **minden igei esemény** detektálásával és osztályozásával foglalkoztunk. Néhány kutatás foglalkozott angol nyelvre igei események detektálásával és osztályozásával. A legtöbb munkában az igei mellett más szófajhoz tartozó eseményeket is megvizsgáltak.

Bethard [1] statisztikai jellemzők alapján detektált eseményeket. Figyelembe vett többszavas kifejezéseket is. A következő jellemzőcsoportokat használta fel az osztályozónál: az adott szó, trigramok a szó elején, végén, morfológiai jellemzők, szófaj, szintaktikai jellemzők, időbeliség kifejezése, tagadási jellemző, WordNet hiperním jellemző. Nem csak a vizsgált szóra, hanem a környező néhány szóra is kigyűjtötték ezeket a jellemzőket. Detektálásra a modell 88,3-os F-mértéket ért el, osztályozásra 70,7-ot.

Llorens és társai [3] CRF modellt alkalmazott szemantikai szabályok felismerésével események detektálásához és osztályozásához. Morfológiai, szintaktikai, szemantikai jellemzőket használtak fel az osztályozáshoz. Egyes jellemzőket nem csak az adott szóhoz, hanem néhány szavas környezetükhöz is kigyűjtöttek. Detektálásra a modell 91,33-os F-mértéket ért el, osztályozásra 73,51-ot.

Marsic [4] csak igei eseményekkel foglalkozott, azok detektálásával és osztályozásával. Statisztikai módszereket használt a feladathoz. Morfológiai és szintaktikai jellemzőket használt fel. Detektálásra a modell 86,49-os F-mértéket ért el.

Bittar [2] francia nyelvű szövegekhez végzett eseménydetektálást. Detektálásra a modell 88,8-os F-mértéket ért el.

Az általunk megvalósított megközelítés *gépi tanuló módszer* alapján detektálja és osztályozza az eseményeket, amit *szabály alapú módszerrel* is kiegészítettünk. A feladathoz *gazdag jellemzőteret* használtunk fel. A detektálás előtt kinyertük a *többszavas kifejezéseket*. A **detektálásnál** 93,85-os, két **osztályozásnál** pedig 85,93 és 66,06-os F-mértéket értünk el.

3 A Korpusz, programok

Alkalmazásunkban a **Szeged Korpusz** egy olyan változatát használtuk fel, amelyikben annotálva vannak a többszavas kifejezések [8]. A korpusznak egy részét használtuk fel, ami *5010 mondatot* tartalmaz a következő területekről: *üzleti rövidhírek, szépirodalom, jogi szövegek, újsághírek, fogalmazás*. Tanításhoz véletlenszerűen kiválasztottuk a korpusz 90%-át, kiértékelésre pedig a maradék 10%-ot.

A detektálásához is ezt az 5010 mondatot használtuk fel. A mondatokat nyelvész segítségével *annotáltuk* a detektáláshoz és az osztályozáshoz is. Az annotátorok közötti egyetértés a detektálásnál 87%-os volt, az osztályozásnál 81%.

Az osztályozáshoz a *Weka*[1] *programcsomagnak* a C4.5 döntési fa algoritmust implementáló J48 tanuló algoritmust alkalmaztuk. A magyar nyelvű szövegek feldolgozásához a *Magyarlanc 2.0* [9] csomagot használtuk.

4 Többszavas kifejezések detektálása

A faladat első részeként detektáltuk a szövegekben a *többszavas kifejezéseket*. Az alkalmazásunkban felhasználtuk az [6]-os publikációban bemutatott alkalmazás elveit. Erről a modulról részletesen írtunk az [7]-es publikációban is, így itt most csak a lényegét foglaljuk össze.

Az 5010 mondatot tartalmazó korpuszunk 100291 db tokent, és ezen belül 542 többszavas kifejezést tartalmazott.

Az alkalmazásban a következő *alapjellemzőket* használtuk fel az osztályozásnál: felszíni jellemzők, lexikai jellemzők, morfológiai jellemzők, szintaktikai jellemzők. Az 5010 mondatos korpuszon az alkalmazásunk ezekkel a jellemzőkkel a következő eredményeket érte el: Pontosság=90,48 Fedés=41,30 *F-mérték=56,72*

A mérésünket még kiegészítettük két jellemzővel. Az első esetben *frekvenciainformációkat* vettünk fel. Minden főnév + ige többszavas kifejezéshez (lemma + abszolút lemma párhoz) meghatároztuk, hogy milyen arányban volt a tanító korpuszon esemény. A tanításnál és a kiértékelésnél felhasználtuk ezt az arányt is, mint jellemzőt. Ezzel a kiegészítéssel a következő eredményt értük el: Pontosság=96,43 Fedés=58,70 *F-mérték=72,97*. Ez a jellemző jelentősen javította az eredményt.

A másik esetben ezt még kiegészítettük a következő jellemzővel. Az alapjellemzők között a lexikai jellemzőknél volt egy *lista*, amiben tárolásra kerültek gyakori többszavas kifejezések [6]. Ezt a listát kiegészítettük a tanító korpuszból vett újabb többszavas kifejezésekkel. Az újabb jellemző akkor kapott igaz értéket, ha a többszavas kifejezés-jelölt szerepelt ebben a listában (szótárillesztés). Ezzel a kiegészítéssel a következő eredményt értük el: Pontosság=93,18 Fedés=89,13 *F-mérték=91,11*. Ez a jellemző is jelentősen javította az eredményt.

5 Igei események detektálása

Ebben a modulban az igei és főnévi igenévi eseményeket *detektáltuk*. A feladatot bináris *osztályozásra* vezettük vissza, amit *szabály alapú módszerrel* is kiegészítettünk. Ehhez a modulhoz külön osztályozót készítettünk. Az osztályozásnál eseményjelölteknek az igéket és a főnévi igeneveket válogattuk ki.

Az 5010 mondatunk 9445 igét tartalmazott, amiből 5487 volt eseményt jelölő.

[1] Weka [2013] Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

5.1 Jellemzőkészlet

Az eseményjelöltekhez a következő jellemzőket gyűjtöttük ki:

- **Felszíni jellemzők-1: Bigramok, trigramok, fourgramok**: A vizsgált szavak elején és végén lévő 2-es, 3-as, 4-es betűcsoportok. A jellemzők közé felvettük, hogy egy adott szó milyen betűcsoporttal kezdődik és végződik.
- **Felszíni jellemzők-1**: Szóhossz lemmahossz, valamint a szó sorszáma a mondaton belül.
- **Lexikai jellemzők**: Az adott szó létige, vagy segédige-e? Egy-egy listába kigyűjtöttük a létigéket és a segédigéket. Jellemzőként megadtuk, hogy az adott szó szerepel-e valamelyik listában.

Mivel egy szónak az eseményjellegét meghatározhatja az is hogy előtte, vagy utána áll-e létige vagy segédige, ezért ezt a négy bináris jellemzőt is felvettük.

- **Morfológiai jellemzők**: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. Jellemzőként definiáltuk az eseményjelöltek MSD-kódját felhasználva a következő morfológiai jegyeket: típus (SubPos), mód (Mood), eset (Cas), idő (Tense), személy (PerP), szám (Num), határozottság (Def).

Jellemzőként felvettük még az igekötőt és az adott szó, valamint az előtte és az utána álló szó szófaját.

- **Szintaktikai jellemzők**: Megadtuk, hogy az adott eseményjelölthöz milyen szintaktikai kapcsolattal tartoznak szavak. (például alany, tárgy, …). Kiemelt figyelmet szenteltünk ezek közül a PRED kapcsolatnak, mert nem esemény igéknél gyakran ilyen kapcsolata van az igének. Ezért ezt is definiáltuk a jellemzők között.
- **Szemantikai jellemzők**: Ehhez a Magyar WordNet-et [5] használtuk fel. Először olyan osztályozót készítettünk, amelyikbe jellemzőként felvettük, hogy a vizsgált szónak mik a *hipernimái*. A tanítással az osztályozó kiválogatta a döntési fába azokat a *synseteket*, amelyek alá jellemzően események tartoznak. Ezeket a kiválogatott synseteket használtuk fel a fő feladathoz. Egy listában felvettük ezeket, majd jellemzőként megadtuk, hogy az adott eseményjelölt szerepel-e valamelyik ilyen synset *hiponimái* között.

Ha csak a WordNet jellemzőt alkalmaztuk önállóan, akkor bár nem a legjobb, de 91,4-es F-mértéket értünk el.

A gépi tanuló módszerünket kiegészítettük szabály alapú módszerrel is. A jogi korpuszon sok olyan kifejezés volt, amelyekben az ige más szövegekben általában eseményt jelöl, de ebben a szövegkörnyezetben nem. Például: *A törvény **kimondja**, hogy…* Az okirat **meghatározza**, hogy… Ezekhez az esetekhez definiáltunk szabályokat. Például: ha alany=törvény és ige=kimondja, akkor kimondja ≠ esemény.

A kiértékelés során a pontosság (P), fedés (R) és F-mérték (F) metrikákat használtuk. Több fajta mérést is végeztünk. Megvizsgáltuk az alkalmazást az öt korpuszon együtt *tízszeres keresztvalidációval*, valamint külön-külön a *részkorpuszokon* is teszteltük a működését. Porlasztásos méréssel vizsgáltuk meg az egyes *jellemző csoportok*

jelentőségét az adott feladathoz. *Domainek közötti keresztméréseket* is alkalmaztunk, mely során a forráskorpuszon tanított modellt értékeltük ki a célkorpuszon.

Két baseline megoldást vizsgáltunk. Az egyikben minden igét és főnévi igenevet eseménynek tekintettünk. A másikban csak azokat az igéket és főnévi igeneveket tekintettük eseménynek, amelyek nem létigék és nem segédigék.

5.2 Eredmények – Detektálás

Az első **baseline** megoldásunk 80,92-es F-mértéket ért el, a másik pedig 85,32-öt.

Teljes jellemzőkészlettel véletlen felosztással a következő eredményeket értük el: Pontosság=93,68, Fedés=94,63, F-mérték=94,15. Tízszeres **keresztvalidációval** 93,85-ös F-mértéket kaptunk. Ha csak az igéket vizsgáljuk, akkor 93,05-ös F-mértéket értünk el. Ha *elhagytuk a szabály alapú módszert*, akkor csak 93,72-es F-mértéket értünk el.

Megvizsgáltuk, hogy az egyes **jellemzőcsoportok** hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztünk, amelynek az eredményei az 1. táblázatban találhatóak. Ekkor a teljes jellemzőkészletből elhagytuk az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottunk. Az eredmények alapján a leghasznosabbnak a szemantikai, a lexikai és a szintaktikai jellemzők bizonyultak.

1. táblázat: Az egyes jellemzőosztályok

Jellemző	Pontosság	Fedés	F-mérték	Eltérés
Felszíni jellemzők-1: Bi-,tri-, fourgramok	92,44	95,79	94,08	-0,07
Egyéb felszíni jellemzők	92,47	96,23	94,31	+0,16
Lexikai jellemzők	91,73	94,92	93,30	-0,85
Morfológiai jellemzők	92,71	95,94	94,29	+0,14
Szintaktikai jellemzők	92,54	95,36	93,92	-0,23
Szemantikai jellemzők	91,54	94,19	92,85	-1,3

Kiegészítő mérések a Felszíni jellemzők-1 nélküli esetre.

A jellemzőkészletet kiegészítettük **szózsák** jellemzőkkel. Először felvettük a jellemzők közé az adott eseményjelölt szintaktikai alárendeltjeinek lemmájának halmazát. Ezzel 93,49-es F-mértéket értünk el. Utána ehhez hasonlóan a jellemzőkészletet az eseményjelölthöz tartozó szavak és a kapcsolat típusa halmazzal bővítettük. Ezzel 93,81-es F-mértéket értünk el. Látjuk, hogy ezek a kiegészítések nem javítottak a vizsgált eredményen.

Következő mérésként **frekvenciainformációkat** vettünk fel. A tanító halmaz alapján kigyűjtöttük, hogy az egyes igék lemmája milyen arányban esemény. Ezt az arányt is felvettük a jellemzők közé. Ez javított az eredményen: 95,82-es F-mértéket kaptunk. Mivel az igekötő is megváltoztathatja egy ige eseményjellegét, ezért a kö-

vetkező esetben nem a lemmához, hanem az igekötő+lemma párhoz vettünk fel az előzőhöz hasonló arányt. Ezzel még jobb eredményt értünk el: F-mérték= F:95,95.

Korpuszonként is megvizsgáltuk az alkalmazás működését. Ennek eredményei a 2. táblázatban láthatóak. Legjobban az Üzleti rövidhírek és az Újsághírek doménen teljesített a modell, leggyengébben pedig a Jogi doménen.

2. táblázat: Eredmények az egyes részkorpuszokon

Korpusz	**Pontos-ság**	**Fedés**	**F-mérték**
Fogalmazás	94,84	98,00	*96,39*
Jogi	92,11	86,42	*89,17*
Szépirodalom	96,03	96,03	*96,03*
Üzleti rövidhírek	97,14	97,84	*97,49*
Újsághírek	96,73	98,01	*97,37*

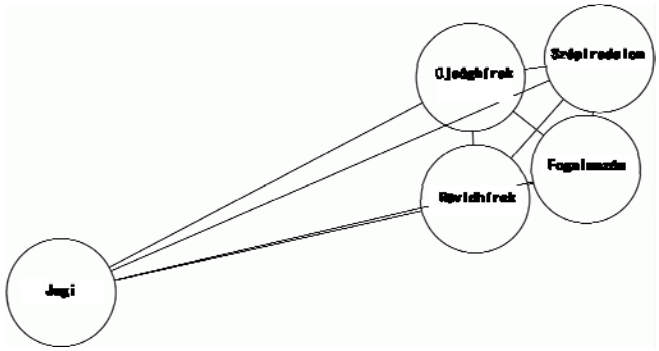
A **domainek közötti keresztméréseknél** a forráskorpuszon tanított modellt értékeltük ki a célkorpuszon. Ennek eredményét a 3. táblázatban láthatjuk. A fogalmazás korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,42-es F-mértéket elérve. A jogi korpuszon szintén az újsághírek doménen tanított modell teljesített a legjobban 83,11-es F-mértéket elérve. A szépirodalom korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,73-es F-mértéket elérve. Az üzleti rövidhírek korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,71-es F-mértéket elérve. Az újsághírek korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,80-es F-mértéket elérve.

3. táblázat: Keresztmérések eredményei az egyes részkorpuszokon

Korpusz	**Pontos-ság**	**Fedés**	**F-mérték**
Fogalmazás	97,49	98,91	*98,20*
Jogi	68,56	99,09	*81,04*
Szépirodalom	92,04	97,58	*94,73*
Üzleti rövidhírek	92,73	98,04	*95,32*
Újsághírek	91,26	98,62	*94,80*
Jogi	95,75	93,88	*94,81*
Fogalmazás	81,70	72,67	*76,92*
Szépirodalom	88,19	72,34	*79,48*
Üzleti rövidhírek	94,72	76,47	*84,62*
Újsághírek	90,74	71,90	*80,23*
Szépirodalom	97,52	98,24	*97,88*
Fogalmazás	94,68	95,64	*95,16*
Jogi	67,05	97,79	*79,56*
Üzleti rövidhírek	92,91	96,16	*94,51*
Újsághírek	91,38	96,22	*93,74*
Üzleti rövidhírek	98,00	99,09	*98,54*
Fogalmazás	92,63	95,86	*94,21*

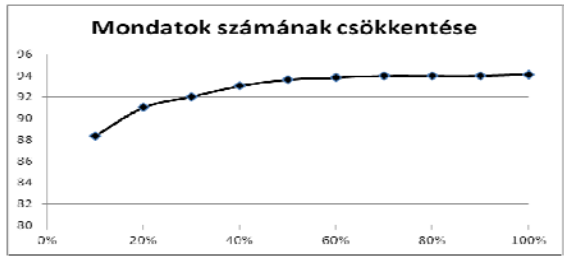
Jogi	69,83	96,74	*81,11*
Szépirodalom	91,26	96,48	*93,79*
Újsághírek	91,29	95,86	*93,52*
Újsághírek	95,06	99,27	*97,12*
Fogalmazás	93,33	97,60	*95,42*
Jogi	72,13	98,05	*83,11*
Szépirodalom	90,83	98,09	*94,32*
Üzleti rövidhírek	93,48	98,04	*95,71*

A keresztmérések eredményei alapján az egyes **domének közti hasonlóságokat** megjelenítettük egy *irányítatlan súlyozott gráf* segítségével. (1. ábra) Az ábrán látható, hogy a jogi korpusz a legkevésbé hasonló a többihez e szempontok alapján.



1. ábra: Doménhasonlósági gráf a keresztmérések eredményei alapján

A következő mérésben **csökkentettük a mondatok** számát. Az F-mértékekre kapott eredmény a 2. ábrán látható. A mondatok számát csökkentve romlik az eredmény.



2. ábra: Mondatok számának csökkentése

6 Igei események osztályozása

Az igei események detektálása után **osztályoztuk** azokat. Az osztályozást több szempont szerint is elvégeztük. Az igék alapkategóriáit vizsgáltuk meg: cselekvés, törté-

nés, létezés, állapot. Ezek közül az eseményeknél a **cselekvés és a történésnek** van fő szerepe, így ezt a két kategóriát emeltük ki. Az 5010 mondaton belül 3905 cselekvés és 1582 történés típusú esemény volt.

Ugyanazt a **jellemzőkészletet** használtuk fel, mint a detektálásnál. Mind a két osztályozásnál, a szemantikai jellemzőknél, a **WordNet**et felhasználva először osztályozóval olyan *synseteket* kerestünk, amelyek *hiponimái* között jellemzően az adott osztály szavai szerepelnek. Ezeket a *synseteket* egy listában felvéve jellemzőként, definiáltuk, hogy az adott szó szerepel e valamelyik ilyen synset *hiponimái* között.

Ha csak a WordNet jellemzőt alkalmaztuk önállóan, a cselekvés vizsgálatnál 87,26, a történés vizsgálatnál 73,31-es F-értéket értünk el.

Mind a két vizsgálathoz készítettünk 1-1 *baseline* megoldást.

6.1 Eredmények – Osztályozás

A **baseline** modellünk minden eseményt cselekvésnek tekintett. Ezzel 78,70-as F-mértéket ért el.

Teljes jellemzőkészlettel véletlen felosztással a következő eredményt értük el az F-mértékre: Cselekvés: 85,93; Történés: 66,06

Tízszeres **keresztvalidációval** kaptuk: Cselekvés: 84,9; Történés: 65,34

Megvizsgáltuk, hogy az egyes **jellemzőcsoportok** hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztünk, amelynek az eredményei az 4. táblázatban találhatóak, osztályozásonként külön sorban, a következő sorrendben: Cselekvés (Cs); Történés (T). Ekkor a teljes jellemzőkészletből elhagytuk az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottunk. A cselekvés és a történés osztályoknál a szemantikai jellemzők voltak a legmeghatározóbbak.

4. táblázat: Az egyes jellemzőosztályok

Jellemző	Pontosság	Fedés	F-mérték	Eltérés
Felszíni jellemzők-1:	Cs: 82,64	89,49	85,93	*0*
Bi-,tri-, fourgramok	T: 74,48	59,34	66,06	*0*
Egyéb felszíni jellemzők	Cs: 80,67	85,91	83,21	*-2,72*
	T: 78,40	53,85	63,84	*-2,22*
Lexikai jellemzők	Cs: 81,61	88,37	84,85	*-1,08*
	T: 69,80	57,14	62,84	*-3,22*
Morfológiai jellemzők	Cs: 81,01	89,71	85,14	*-0,79*
	T: 87,85	51,65	65,05	*-1,01*
Szintaktikai jellemzők	Cs: 82,39	87,92	85,06	*-0,87*
	T: 77,78	61,54	68,71	*+2,65*
Szemantikai jellemzők	Cs: 78,71	81,88	80,26	*-5,67*
	T: 62,58	53,30	57,57	*-8,49*

Kiegészítő mérések a Felszíni jellemzők-1 nélküli esetre.
Itt is elvégeztük azokat a kiegészítő méréseket, mint a detektálásnál.

A jellemzőkészletet kiegészítettük **szózsák** jellemzőkkel. Először felvettük a jellemzők közé az adott szóhoz szintaktikailag tartozó szavak lemmájának halmazát. Ezzel a következő F-mértékeket értük el:

Cselekvés: 83,35; Történés: 66,07

Utána ehhez hasonlóan a jellemzőkészletet a vizsgált szóhoz tartozó szavak és a kapcsolat típusa halmazzal bővítettük. Ezzel a következő eredményeket értük el:

Cselekvés: 85,12; Történés: 66,67

Ez az utóbbi kiegészítés javított az eredményeken.

Következő mérésként **frekvenciainformációkat** vettünk fel. A tanító halmaz alapján kigyűjtöttük, hogy az egyes igék lemmája milyen arányban tarozik a vizsgált osztályba. Ezt az arányt is felvettük a jellemzők közé. Ez mindegyik osztálynál javított az eredményeken. A következő F-mértékeket kaptuk:

Cselekvés: 86,98; Történés: 76,70

Itt is megvizsgáltuk, hogy ha nem csak a szavakhoz tároljuk el az arányt, hanem az igekötő+lemma párhoz is, akkor az hogyan befolyásolja az eredményt. Ez volt ahol javított az F-mértéken:

Cselekvés: 88,20; Történés: 75,00

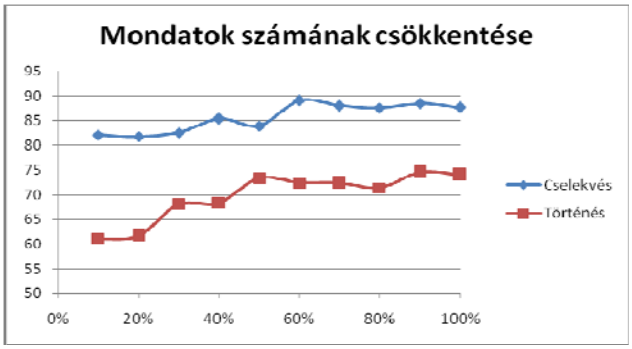
Korpuszonként is megvizsgáltuk az alkalmazás működését. Ennek eredményei az 5. táblázatban láthatóak. A cselekvéseket osztályozó modell a jogi korpuszon a történéseket osztályozó az üzleti rövidhírek korpuszon teljesített a legjobban.

5. táblázat: Eredmények az egyes részkorpuszokon

Korpusz	**Pontosság**	**Fedés**	**F-mérték**
Fogalmazás	Cs: 83,81 T: 57,14	85,44 32,43	*84,62* *41,38*
Jogi	Cs: 90,57 T: 77,78	87,27 87,50	*88,89* *82,35*
Szépirodalom	Cs: 79,44 T: 68,75	85,86 64,71	*82,52* *66,67*
Üzleti rövidhírek	Cs: 91,43 T: 88,33	85,33 94,64	*88,28* *91,38*
Újsághírek	Cs: 83,33 T: 70,00	82,52 61,76	*82,93* *65,63*

Domainek közötti keresztméréseket itt is végeztünk. A forráskorpuszon tanított modellt értékeltük ki a célkorpuszon. Legjobb eredményt a cselekvések osztályozásánál értük el, a szépirodalom doménen tanított modellel a fogalmazás korpuszon 85,5-os F-mértékkel. A leggyengébb eredményt pedig a történések osztályozásánál a fogalmazás doménen tanított modellel a szépirodalom korpuszon 53,91-os F-mértékkel.

A következő mérésben csökkentettük a mondatok számát. Az F-mértékekre kapott eredmények a 3. ábrán láthatóak. A **mondatok számát csökkentve** mindkét osztályozásnál romlottak az eredmények.



3. ábra: Mondatok számának csökkentése – osztályozás

7 Összegzés

Munkánkban bemutattunk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben igei eseményeket azonosítani és azokat osztályozni. A problémát három lépésben oldottuk meg. Először detektáltuk a többszavas főnévi+igei kifejezéseket. Majd detektáltuk az igei és főnévi igenévi eseményeket, és osztályoztuk azokat. Módszerünket a Szeged Korpusz öt doménjén próbáltuk ki tízszeres keresztvalidációval. A modellünk jellemzőkészletét teszteltük porlasztásos módszerrel. A módszert teszteltük a doméneken egyesével, majd tanítva az egyiken a többi doménen pedig kiértékelve. Az egyes domének közötti hasonlóság kifejezésére hasonlósági gráfokat is megadtunk. A detektálásra az alapjellemzőkkel és 93,85-os F-mértéket, a két szempont szerinti osztályba sorolásra pedig 85,93 és 66,06-os F-mértéket értünk el. Kiegészítő mérésekkel javítottuk ezeket az értékeket. Ezek jó eredménynek számítanak a bemutatott előző munkákkal összehasonlítva.

Hivatkozások

1. Bethard, S.J.: Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach PhD thesis, University of Colorado (2002)
2. Bittar, A.: Annotation of Events and Temporal Expressions in French Texts, ACL-IJCNLP '09 Proceedings of the Third Linguistic Annotation Workshop (2009) 48–51
3. Llorens, H., Saquete, E., Navarro-Colorado, B.: TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics (2010) 725–733
4. Marsic, G.: Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations. PhD thesis, University of Wolverhampton (2011)
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), Szeged, University of Szeged (2008) 311–320
6. Nagy T. I., Vincze V., Zsibrita J.: Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 47–58
7. Subecz Z., Nagyné Csák É.: Események detektálása természetes nyelvű szövegekben. Matematikát, fizikát és informatikát oktatók XXXVII. országos konferenciája, Miskolc (2013) 201–208
8. Vincze V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2009) 390–393
9. Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 368–374

Felszíni szintaktikai elemzés és a jóindulatú interpretáció elve információ-visszakeresésben

Gyarmathy Zsófia, Simonyi András, Szőts Miklós

Alkalmazott Logikai Laboratórium, Budapest
e-mail:{szots,simonyi}@all.hu, gyzsof@gmail.com

Kivonat Tanulmányunkban egy új szintaktikai elemzési megközelítésre teszünk javaslatot, amely egy, a szemantikai predikátum-argumentum viszonyokra építő, orvosi ajánlásokban történő információ-visszakeresést megvalósító rendszerben kerül alkalmazásra. Szakítva a hagyományos, mélyelemzési módszerekkel, egy fedésorientált, „jóindulatú interpretáció-val" kiegészített felszíni szintaktikai elemzést javaslunk. Másszóval nem kívánjuk meg a szintaktikai reprezentáció helyességét, azaz nem törekszünk a pontosságra, hanem sokkal inkább csak a fedésre. A keresés pontosságát ehelyett szemantikai információk és a keresőkifejezés segítségével javítjuk. Bemutatjuk, hogy ezáltal csak kevés esetben lesz rosszabb a pontosság, míg számos jelenség (például kontrolligék, koordinációk, szabad határozók) esetében komoly előnyt jelent a javasolt fedésorientált megközelítés.

Kulcsszavak: felszíni szintaktikai elemzés, szemantikus információ-visszakeresés, frame-szemantika, argumentumstruktúra

1. Felszíni szemantikai elemzés egy IR-rendszerben

A cukorbetegség hosszútávú kezelését támogató informatikai platform kifejlesztésére létrejött európai REACTION projekt[1] keretén belül egy orvosi ajánlásokban *információ-visszakeresést* (l. [5]) megvalósító rendszert építünk ki, amely számba veszi a predikátumokat és a hozzájuk tartozó argumentumokat. Ezáltal sokkal eredményesebb lehet a keresés (l. pl. [13]), javítva nem csupán a pontosságot (mivel a megfelelő argumentumrelációjú találatok magasabbra lesznek rendezve, azaz megkülönböztethetjük pl. a *Péter kedvenc tanára Mari* és a *Mari kedvenc tanára Péter* mondatokat), hanem a fedést is (mivel megtaláljuk a hasonló eseményeket kifejező mondatokat, valamint a hasonló argumentumstruktúrával rendelkező mondatokat is).

A szemantikai predikátum-argumentum struktúra azonosítása, amit felszíni szemantikai elemzésnek hívnak (l. pl. [9]), legalább az alábbi lépésekből áll (l. pl. [8]):

1. A predikátumok és az argumentumok határának beazonosítása, és az argumentumok predikátumokhoz kapcsolása. Ez alapvetően egy *szintaktikai lépés*

[1] Remote Accessibility to Diabetes Management and Therapy in Operational healthcare Networks, http://www.reaction-project.eu.

a folyamatban, és a korábbi megközelítések a szövegek *teljes* szintaktikai elemzését feltételezték a feladathoz. Még ha természetesen nem is kizárólag mélyelemzés adta szintaktikai jegyeket használtak fel a klasszifikációs algoritmusokban (hanem például olyanokat is, mint a POS-tag, azaz a nyelvtani kategória), minden elemzés hivatkozik az argumentumnak a teljes szintaktikai fában elfoglalt helyére is (erre szolgál például [3] esetében a „Parse Tree Path" jegy, [9] esetében a „Path" jegy stb.).

Ezek az elemzések – az elméleti nyelvészeti hagyományoknak megfelelően – egyértelmű és helytálló predikátum-dependens, illetve fej-argumentum viszonyokat feltételeznek, azaz a folyamat minden szintjén a legpontosabb reprezentációt kívánják meg. Mi ezzel a hagyománnyal kívánunk a REACTION projekt keretén belül szakítani, és a célszövegek esetében csupán *felszíni* szintaktikai elemzést javaslunk, amelynek az esetében *nem* tételezzük fel annak helyességét, azaz a folyamat ezen a pontján pusztán a maximális fedést kívánjuk meg, a *pontosságot nem* (l. további fejezetek).

2. Szükség van még természetesen a predikátum szemantikai típusának beazonosítására. Ez egy jelentés-egyértelműsítési lépés, amely esetünkben – mivel a FrameNet keretet használjuk – *frame-azonosítást* jelent. Erre a fázisra is számos megoldási javaslat van már az irodalomban (például [2]); mi azonban – egyelőre, a maximális fedést biztosítandó[2] – minden olyan frame-et megengedünk egy predikátum esetében, amelynél az fel van sorolva, az ismeretlen predikátumok esetében pedig úgy járunk el, hogy az igékre egy *alapértelmezett*, nagyon általános frame-et veszünk fel, míg főnevek és melléknevek esetén egyelőre nem tekintjük predikátumnak a frame-hez nem rendelt szavakat.
3. Az argumentumok felcímkézése megfelelő szemantikus szerepekkel (*semantic role labelling*), amely lépés természetesen erősen függ a használt szemantikai szerepektől. Általános, a nyelvészetből is jól ismert tematikus szerepekre, amilyeneket például a VerbNet lexikon ([4]) használ, könnyebb statisztikai tanuló algoritmust adni, mivel minden predikátum ugyanazt a szemantikai szerephalmazt használja, s így viszonylag nagy a mintamennyiség. Mi azonban – a rendszer egyéb előnyei miatt, l. [10,12] – a FrameNet [1] frame-szemantikai megközelítését alkalmazzuk, amelyben a szemantikai szerepek frame-specifikusak. Ez a minta szegénysége miatt[3] megnehezíti a statisztikai tanulást.[4]

[2] Jelenleg ugyanis a rendszer többi részének teljesítményét szeretnénk tesztelni, márpedig egy újabb független paraméter nagyon elbonyolítaná a mérést.

[3] Sőt, számolni kell „hibás" mintával is, mivel egyazon szerepnév más-más szerepet takarhat. Noha a legtöbbször az azonos nevű szerepek valójában hasonló általános szereprelációt tükröznek (l. [6]) – például a Goal szerep mint cél szinte minden esetben tekinthető ugyanazon általános szerepnek –, sok esetben mást takar az ugyanazon elnevezésű szerep – például a Patient a tematikus szerepek esetén is megszokott jelentése mellett az orvoslásbeli pácienst is jelölheti egyes egészségügyi frame-ekben.

[4] A probléma természetesen azokkal a predikátum-argumentum párokkal van, amelyek esetén a FrameNet szótár hiányos, és nem specifikálja, hogy az adott szintaktikai vonzat a predikátum milyen szemantikai argumentumának felel meg.

Ennek ellenére már számos FrameNet-alapú statisztikai SLR-algoritmust javasoltak (pl. [14,7]) legtöbbször az alapcikknek számító [3]-ra építve, különböző szintaktikai és szemantikai jegyeket felhasználva különböző klasszifikációs eljárásokkal; illetve vannak olyan javaslatok is, amelyek más erőforrások klasszifikációs eljárásainak kimenetét aknázzák ki a FrameNet típusú szerepek klasszifikációjához (pl. [10]).
Mi azonban egyelőre – egyszerűsítési okokból – a szótárban nem specifikált vonzatok esetében oly módon járunk el, hogy először az adott predikátum más vonzatkereteiben nézzük meg, milyen szemantikai szerepet kap az adott vonzattípus (tipikusan valamilyen prepozíciós bővítmény), majd második lépésben az azonos frame-hez tartozó, hasonló predikátumok vonzatkereteit nézzük át e célból, végül, amennyiben itt sem találtuk meg ezt a vonzattípust, alapértelmezett eseteket alkalmazunk. Természetesen az egyre tágabb körben talált mintához egyre kisebb megbízhatósági valószínűséget allokálunk.

A felszíni szemantikai elemzés a kétezres években lezajlott kiterjedt kutatások (pl. [3]) ellenére továbbra sem kielégítően megoldott, ezért, szemben a teljes szintaktikai elemzést feltételező gyakorlattal, megkíséreljük pusztán felszíni szintaktikai elemzéssel megközelíteni a felszíni szemantikai elemzés feladatát. Mivel ez a fent leírt rendszer első fázisát érinti, ezért alapvetően meghatározza a további lépések sikerességét is.

2. Felszíni szintaktikai elemzés

A projekt során a korábbi, MaSzeKer projektben[5] szabadalmi igénypontokra kifejlesztett elemzőrendszert ([12]) alakítjuk át a megváltozott feladatnak megfelelően. A szabadalmi igénypontok szintaxisa *kötöttebb* volt (például nem tartalmaz felszólító módú mondatokat), viszont egy igényponton belül szemantikailag nagyon közel álló entitásokról tartalmazott szemantikailag hasonló állításokat (például egyes kémai anyagok összetevőiről, jellemzőiről). A szabadalmi igénypontokhoz tehát elengedhetetlen a mély szintaktikai elemzés, hogy egészen pontosan beazonosíthassuk az egyes kifejezések közötti kapcsolatot a szemantikai reprezentáció kiépítéséhez.

Ezzel szemben a REACTION projektbeli cukorbetegséggel kapcsolatos ajánlások sokkal közelebb állnak a természetes nyelvhez, mint a kötöttebb szabadalmi szövegek, például vannak bennük „phrasal verb"-ök (pl. *carry out*), birtokos szerkezetek, folyamatos igeidő, névmások és mondatkezdő prepozíciós frázisok, ezen felül pedig messze nagyobb a bennük előforduló nyelvtani szerkezetek és a megfogalmazás változatossága. Emiatt a MaSzeKer-beli szövegekre kifejlesztett dedikált szintaktikai elemző nem tud velük megbírkózni. Mi több, bármilyen mélyelemzés sikertelenségre van ítélve, ha a célszöveg a ***blood ketone monitoring*** *with increased healthcare professional support is preferable to urine ketone monitoring* ***in young adults with type 1 diabetes***, míg a keresőkifejezés a „blood ketone monitoring of adults with type 1 diabetes".

[5] Modell Alapú Szemantikus Kereső Rendszer, TECH_08_A2/2-2008-0092.

Ezért inkább amellett döntöttünk, hogy feladjuk a szövegek teljes szintaktikai elemzését, és ehelyett egyfajta felszíni szintaktikai elemzést végzünk. A felszíni szintaktikai elemzésnek is több lépése van hagyományosan:

1. *POS-tagging*, azaz a szavak nyelvtani kategóriájának megállapítása. Ezen a ponton még nem térünk el a mélyelemzésektől.
2. *Chunking*, azaz az összetevők határainak kijelölése. A mi esetünkben ez alapvetően a MaSzeKer-ben kifejlesztett MagNP-kijelölő modult takarja.
3. *Relációfeltárás*, azaz az összetevők közötti szintaktikai viszonyok megállapítása. A jelen tanulmányban ennek a fázisnak egy újfajta, fedésorientált megközelítését mutatjuk be.

Ezen a területen is a statisztikai tanulóalgoritmusok, azon belül is a kevert tanulóalgoritmusok (ensemble learning) alkalmazása a jellemző [11]. Mi ezzel szemben i) a MaSzeKer elemzőrendszerbe jobban illeszkedő szabályalapú megközelítést alkalmazunk, és ii) ahogy fentebb említettük, az elemzésben nagyobb hangsúlyt fektetünk a fedésre, mint a pontosságra, azaz megengedünk „hibás" predikátum-bővítmény kapcsolatokat is a kialakuló szintaktikai reprezentációban. Így például a *treatment [of a patient] [with diabetes]* esetében a *diabetes* főnévi frázist egyaránt felvesszük a *treatment* és a *patient* bővítményeként, miközben csupán az utóbbi elemzés a helyes. Mivel azonban a keresőkifejezésben minden valószínűség szerint nem fogunk *treatment* és *diabetes* között olyan kapcsolatot találni, amely a *with*-es vonzatnak (eszközhatározó) felel meg, így ezt a hibás elemzést a keresés során nem fogjuk felhasználni.[6]

A mi rendszerünk továbbá hibrid rendszer, amennyiben a *főnévi frázisok szintjééig* – az általunk használt terminológiában a MagNP-k[7] szintjééig – *mélyelemzést* végzünk a szövegeken.[8] Mivel a MaSzeKer során egy jól működő modult fejlesztettünk ki a MagNP-k kijelölésére és szintaktikai elemzésére, ezt egy az egyben át tudjuk venni a REACTION projektbeli ajánlások elemzésére. Ami megoldandó, az a MagNP-k és a predikátumok közti (szintaktikai, szemantikai) viszonyok feltárása. Ez tehát lényegében az egyetlen modul, amit a MaSzeKer projekt során kialakított szintaktikai parszerben meg kell változtatni az ajánlásokbeli keresés céljából.

Ezen a ponton pedig összefonódik és egymást meghatározza a rendszerben a szintaxis, a szemantika és a keresés. Ugyanis a MagNP-k és a predikátumok közötti viszonyok megállapításában sokkal megengedőbbek vagyunk, mint egy mélyelemzés, azaz nagyobb a kötési lehetőség, és megengedjük, hogy egy MagNP több predikátum bővítménye is legyen (akár egyazon nyelvtani funkcióban is),

[6] Természetesen egyes esetekben ez a keresésvezérelte „szűrési" eljárás nem fogja eredményesen elkülöníteni a helyes és a helytelen kapcsolatokat, elfogadván helyteleneket is, azonban ezek aránya a szövegtípustól függ: A REACTION-beli szövegek (cukorbetegségekkel kapcsolatos ajánlások) jellegükből adódóan alkalmasak erre a megközelítésre.

[7] Egy MagNP egy minden utómódosítójától megfosztott főnévi frázis.

[8] Erre a célra újraírószabályokat alkalmazunk, azaz frázisstruktúra-nyelvtant használunk.

valamint hogy egy predikátumhoz több, ugyanolyan nyelvtani funkciójú bővítmény (például tárgy) kapcsolódjon. A több kötési lehetőség közül pedig azokat tartjuk majd meg, amelyek a *keresés*, illetve a szemantika[9] szempontjából a *legideálisabbak*: ezt nevezzük a *jóindulatú interpretáció elvé*nek.

3. A jóindulatú interpretáció elve

A jóindulatú interpretáció elvének működését a következő absztrakt példa illusztrálja. Tegyük fel, hogy a keresőkifejezésre felépített szemantikai gráfban megtalálhatók az A, B és C csomópontok, és a következő élek:

- $A \xrightarrow{arg1} B$
- $A \xrightarrow{arg2} C$

Továbbá tegyük azt is fel, hogy a (felszíni szintaktikai elemzéssel elemzett) illesztendő szövegre felépített szemantikus gráfban megtalálhatók az A, B és D csomópontok, és a következő élek:

- $A \xrightarrow{arg1} B$
- $A \xrightarrow{arg3} B$
- $A \xrightarrow{arg2} D$
- $B \xrightarrow{arg2} D$

Ekkor azt fogjuk „jóindulatúan" feltételezni, hogy az $A \xrightarrow{arg1} B$ él illeszkedik, tehát az illesztendő szöveg részleges találat a keresőkifejezésre. Ez akkor is fenntartható, amennyiben például a $A \xrightarrow{arg3} B$ és a $B \xrightarrow{arg2} D$ élek „hibásan" kerültek be a szemantikus gráfba, a pontosságot figyelmen kívül hagyó felszíni szintaktikai elemzés révén.

Innentől kezdve gyakorlati kérdés, hogy mennyire megszorított, illetve szabad szintaktikai kötési lehetőségek bizonyulnak a keresés pontossága *és* fedése szempontjából legideálisabbnak (megkeresve a legjobb „trade-off"-ot a két mérték között). Elképzelhető – a szövegek jellegétől függően –, hogy egy „anything goes", azaz megkötések nélküli fej-dependens összekapcsolás működik a legjobban, amennyiben megfelelő szemantikai eszközökkel (például szelekciós restrikciókkal) kordában tudjuk tartani az elemzések elburjánzását. Ehhez azonban szemantikai információval gazdagon feltöltött lexikonra van szükség, amely például specifikálja az egyes predikátumok megfelelő argumentumainak a szemantikai típusát (azaz a predikátum szelekciós restrikcióit). Noha a lexikális erőforrások fedése és információgazdagsága terén jelentős előrelépések történtek az elmúlt évtized során is, efféle szemantikai információ meglétére még kevéssé támaszkodhatunk a legtöbb szótári tétel esetén (l. 9. lábjegyzet).

Mi, részben ezért is, első körben egy megszorítottabb megközelítést választottunk, és megfogalmaztunk egy *véges szabályrendszert* arra vonatkozóan, hogy

[9] A lexikonban rendelkezésre álló szemantikai információ (elsősorban az egyes argumentumokra vonatkozó szelekciós megszorítások) jelenleg még elég korlátozott, ezért megszorító hatása egyelőre lényegében elhanyagolható.

az egyes esetekben milyen főnévi frázisokat milyen fejekhez köthetünk, és milyen mondattani szereppel. Ezáltal pusztán a *legreálisabb* elemzési lehetőségeket tartjuk meg (így továbbra is különbséget tudunk tenni a *Péter kedvenc tanára Mari* és a *Mari kedvenc tanára Péter* között szintaktikai szinten is), ám eközben teret hagyunk a jóindulatú interpretáció elvének, ami összességében számos előnnyel járhat a mélyelemzéses megközelítésekhez hasonlítva, ahogy lentebb érvelni fogunk. Ez a szabályrendszer azonban nem a mélyelemzéseknél megszokott formátumú (például újraírószabály) és pontosságú. Olyan típusú szabályok ezek, mint például „egy nem prepozíciós MagNP, ha követi a igét, akkor lehet a direkt és indirekt tárgya annak".[10]

A mély elemzés és az itt alkalmazott felszíni közötti alapvető különség abban áll, hogy az utóbbi „megengedőbb", ennek folytán *több lesz az „igaz pozitív"* találat, mert a kereső megtalál olyan ajánlásokat, amelyeket a mélyelemzés nem, vagy csak nagyon alacsonyra rendelt résztalálatként. Jelentősen javul tehát a rendszer *fedése*. Viszont éppen ezért *több lesz a „téves pozitív"* találat is, mert olyat szövegrészeket is találatnak vesz (egy predikátumhoz kapcsolva nem egybe tartozóakat), amelyek valójában nem azok. Ez csökkenti a rendszer *pontosságát*.

Reményeink szerint azonban az ajánlások esetében ez a pontosságcsökkenés alacsony lesz. Ha például *blood pressure*, *patient* és *high* szerepel egy ajánlásban, igen kicsi (persze nem nulla) a valószínűsége, hogy egy magas páciens vérnyomásáról van szó (*the blood pressure of a patient who is high*), tehát valószínű, hogy a *high* a *blood pressure*-re vonatkozik (*the blood pressure of the patient is high*); ugyanígy feltételezhetően minden egy mondaton belüli információ egyetlen páciensre vonatkozik. A felszíni elemzés tehát azért működhet a REACTION-beli ajánlásokon, mert az ajánlások jellemzően *rövidek*, így emiatt és a szövegtípus sajátossága miatt kicsi az esélye, hogy a keresésben szereplő főnévi és egyéb frázisok „rekombinálása" a szövegen belül sokszor hozna be téves pozitívot.

4. A felszíni elemzés előnyei

Az itt felvázolt, jóindulatú interpretáción alapuló, fedésorientált felszíni elemzés számos esetben lehetővé teszi a keresés jobb fedését, illetve kiválthat bonyolultabb dedikált szintaktikai modulokat. Fentebb már a „blood ketone monitoring" példáján bemutattuk, hogy a természetes nyelvben általánosságban is igen sokféle megfogalmazása lehet egyazon gondolatnak, ilyen esetekben pedig bármiféle mélyelemzés kudarcra van ítélve. Egy másik jó példája az itt felvázolt megközelítés előnyének ilyen szempontból a következő célszövegbeli részlet:

(1) *Cataract extraction should not be delayed [in patients with diabetes].*

[10] Mi a MaSzeKer-beli elemzőhöz hasonlóan egy dependencianyelvtant használunk a MagNP-k feletti szinten, ez a választás azonban az itt tárgyaltak szempontjából kevéssé releváns. Egy frázisstruktúra-nyelvtan például azonban alapjaiban összeegyeztethetetlennek tűnik az itteni koncepcióval, már pusztán amiatt, mert egy összetevőnek több szülőnódusa is kellene, hogy lehessen, valamint mert nem folytonos al-fákra is szükség lenne.

Az általunk alkalmazott szabályrendszer jelenleg felveszi az *extraction* fejhez a *patients in*-prepozíciós dependenst, hiszen teljesen reális lehet egy „*cataract extraction in patients with diabetes*" keresőkifejezés, amelyre helyesen, magasra értékelt találatként kapnánk meg a fenti részletet.

Egyes esetekben a mélyelemzésre felépített szemantikai reprezentáció is módosítható, kiegészíthető lehet megfelelő reasoninggel, azonban egy ilyen szintű reasoning modul komoly kihívásokat jelent, és igen kétséges, hogy az ehhez szükséges tudásbázis rendelkezésre áll-e vagy kiépíthető-e reális időkereteken belül.

Azonban ezen általános kérdéskör mellett vannak egyes specifikus jelenségek is, amelyeknek kezelése sokszor külön, dedikált modult igényelne, azonban egy a javasolthoz hasonló megközelítés mellett erre nem lenne szükség. Az alább részletesebben is bemutatott ilyen jelenségek a következők:

1. ECM/raising/controll igék,
2. koordináció,
3. szabad határozók.

A fent bemutatott felszíni elemzési módszerrel az angol *raising, control*, illetve *ECM igék* (azaz lényegében a megosztott argumentumok) esetében nincs szükség külön minimodulra a célból, hogy a főige alanya, illetve tárgya a beágyazott mondat igéjének is alanya legyen, és ezáltal a megfelelő élek megjelenjenek a szemantikus reprezentációban is. Különösen problémásak ezen igetípusok, ha nem is beágyazott mondatbővítményük van, mivel ekkor nem tudnánk általános szabályt alkalmazni. A következő mondat illusztrálja ezt az esetet:

(2) *Intensive management plus pharmacological therapies should be offered [to patients with diabetes].*

Ebben az esetben az „intensive management for patients with diabetes" keresésre az *offer* jellegű igék külön kezelése nélkül csak résztalálatot kapnánk, miközben valójában teljes találat. A fent vázolt jóindulatú megközelítésben azonban a „patients with diabetes" az „intensive management" vonzata is lenne, így magasabb találati értéket kapna a célszöveg erre a keresésre.

Egy másik nyelvi jelenség, amelynek esetében a javasolt elemzési módszer kiválthat egy külön, dedikált modult, a *koordináció*. A természetes nyelvekben igen szerteágazó az ellipszis, és az egyes összetevők koordinálása, ezek azonban – a triviálisabb esetektől eltekintve – komoly kihívást jelentenek a gépi szintaktikai elemzéseknek. Íme egy nem triviális koordinációt tartalmazó példa:

(3) *Sulphonylureas should be considered as first line oral agents in patients who are not overweight, who are intolerant of, or have contraindications to, metformin.*

Ha a keresőkifejezésünk „medications for patients allergic to metformin", a fenti célszöveget mélyelemzés esetén szinte kizárt, hogy megtaláljuk (legfeljebb olyan résztalálatként, ami nagyjából egy kiterjesztett kulcsszavas keresésnek felel

meg). Egy jóindulatú felszíni megközelítéssel kicsivel túlmehetünk ezen, mivel a „metformin" dependense lehet az „intolerant" fejnek (többek között például a „have" és a „contraindications" fejek mellett). Innentől pedig feltételezve, hogy helyes a frame-szemantikai osztályunk, és az „allergic" és az „intolerant" azonos frame-be tartozik, máris sikeresen nagyobb súlyt kap találatként az ajánlás.

Hasonló módon tudunk megküzdeni a *szabad határozók* problémájával. Ezeknek a disztribúciós lehetőségei még a kötöttebb szórendű angol nyelvben is igen szélesek, ami mélyelemzés esetén megnehezíti a megfelelő fejhez kötésüket. Mi több, amint az ismert „*see [a man] [with a telescope]*" példa is mutatja, valódi szerkezeti többértelműség is fennállhat, ami lehetetlenné teszi, hogy az egyetlen pontos reprezentációt megcélzó mélyelemzés *minden* esetben sikeres legyen. Az itt javasolt keretben azonban megengedjük, hogy egyazon prepozíciós bővítmény több fejhez is kapcsolódjon, azaz ilyen esetben az „a telescope" összetevő mind a „see"-nek, mind a „man"-nek dependense lesz, tehát egy esetben sem vesztünk találatot.

5. A felszíni elemzés veszélyei és feltételei

A jelen rendszerben a legalapvetőbb problémát természetesen a téves pozitív találatok jelentik. Bár – amint említettük – a cukorbetegséggel kapcsolatos ajánlások rövidek, és emellett sem jellemző rájuk, hogy több, szemantikailag hasonló állítást tartalmaznának, ettől azért egyes esetekben előfordulhat. Ez, sarkítva, azonban valamilyen szinten kikerülhetetlen: ha a keresőkifejezésben az áll, hogy teleszkópos embert nézünk, a célszövegben pedig „*see a man with a telescope*", akkor hiába értelmezendő a célszövegben úgy, hogy teleszkóppal nézzük az illetőt (ez kiderülhet egy hosszas szövegkörnyezetből impliciten), ez a rész óhatatlanul illeszkedni fog a keresőkifejezésre. Azaz mindig lesznek „kezelhetetlen" esetek, a cél csupán ezek számának minimalizálása, aminek eszköze alapvetően egy olyan szabályrendszer megfogalmazása, amely elég restriktív ahhoz, hogy a keresési pontosság elfogadható legyen, míg a fedést lényegében nem rontja.

Van azonban két specifikus nyelvi jelenség, amelynek kezelése elengedhetetlen egy jól működő fedésorientált felszíni elemzéshez. Az egyik a *zárójelben* álló összetevők problémája. Egy példa:

(4) *Obese adults with type 2 diabetes should be offered individualised interventions to encourage weight loss (including lifestyle, pharmacological or surgical interventions) in order to improve metabolic control.*

Ebben a példában problémát okozhat, hogy például a „weight loss" a szabályok alapján (ha a rendszer nem „látja" a zárójelet mint határt) a zárójeles részt kezdő „including"-nak lesz az alanya, hibásan.

A legegyszerűbb megoldás, hogy zárójelen belüli szöveget a szöveg többi részétől *elkülönülten* kell leelemezni szintaktikailag. Az elkülönült szintaktikai elemzés csak ritkább esetekben nem működik, például akkor, ha egy főnévhez tartozó előmódosító kerül zárójelbe, például „*(oral) medications*". A nyereség azonban sokkal nagyobb, mint a veszteség, és később természetesen dedikált modul is kidolgozható a zárójeles kifejezések hatékonyabb kezelésére és kiaknázására.

Problémát okoznak a keresés során a zárójelek mellett még a *többszavas kifejezések* (multi-word expression, MWE) is, mint a *for example*, *in the case of*, *in addition to*. Egyrészt ezeket mint dependenseket és/vagy fejeket hibásan fogja kötni az elemző: például az *in the case of* esetén a *case* valamilyen fej(ek)nek az *in*-es dependense lesz hibásan, míg őhozzá mint fejhez *of*-os dependensként lesz kötve az őt követő főnévi frázis – hibásan. Ezek a szintaktikai reprezentációból azután bekerülnek a szemantikai reprezentációba, így helytelen illesztések történhetnek. Másrészt ezek a kifejezések megakadályozhatják a szintaktikai szabályok helyes alkalmazódását, és így a dependensek helyes kötését: például ha egy szabály a fej és a dependens közötti prepozíciókra hivatkozik, a „*for example*"-beli *for* illeszkedni fog a szabálymintára, pedig a „*for example*" összetett kifejezés egy határozó. Az előfeldolgozás során tehát mindenképpen érdemes a többszavas kifejezéseket kijelölni egy külön modulban.

Végül felmerült olyan probléma, amely kevésbé a szintaktikai, sokkal inkább a *szemantikai* reprezentációt érinti. A fedésorientált elemzés miatt esetünkben a szintaktikai gráfok igen nagyok lehetnek: több élt tartalmaznak, mint egy pontos, „helyes" elemzés, sőt, akár csomópontból is több kerülhet be, mivel argumentummal rendelkező predikátum is több lesz potenciálisan egy ilyen megközelítésben (ez a névszói predikátumokban jelent számszerű növekedést). Azonban elképzelhető, hogy a keresés szempontjából kevéssé releváns csomópontok és élek illeszkedése fog magasra értékelni valójában nem releváns találatokat. Egy példa:

(5) a. Keresőkifejezés: *Elderly patient with diabetes. The patient has mobility problems.*
b. Célszöveg: *All people with diabetes, and people without diabetes with a GFR less than 60 ml/min/ 1.73 m2, should* **have** *their urinary albumin/protein excretion quantified. The first abnormal result should be confirmed on an early morning sample (if not previously obtained).*

Ebben az esetben a „have" mint fej (ráadásul nem is megfelelő értelmű) előfordulása a célszövegben magas relevanciát nyújt az irreleváns célszövegnek. A legmegfelelőbb megoldásnak erre a problémára a kulcsszavas keresés újszerű felhasználása lenne: a keresőkifejezésben a felhasználó által megadott kulcsszavak jelölnék ki a szemantikai gráf lényeges csomópontjait, és az ebből kiinduló élek illeszkedése súlyozottan számítana be a relevanciaszámításba. Márpedig a példabeli keresőkifejezésben a „have" egyértelműen nem lenne kulcsszó, így illeszkedése sem hozna be magas relevanciaszámmal irreleváns találatokat.

Úgy tűnhet, hogy az itt leírt problémák és megoldhatóságuk feltételei súlyos ellenérvet jelentenek a fedésorientált felszíni elemzéssel szemben. Mindezen feltételek fennállása azonban ugyanúgy szükséges egy *mélyelemző* parszert használó keresőrendszerben is, hiszen egy mélyelemző ugyanúgy hibás fej-dependens viszonyt fog feltételezni az „*in the case of*" esetén, ugyanúgy problémái lehetnek a zárójeles kifejezésekkel (ezzel problémával ugyanis a mélyelemzést használó MaSzeKer projekt során is találkoztunk), és ugyanúgy magasra értékelhet egy célszöveget a kevéssé kulcsfontosságú frame-ek és argumentumok illeszkedése. A különbség csupán annyi, hogy a legutolsó probléma a fedésorientált felszíni

elemzés esetén hatványozottan jelentkezik, mivel abban az esetben sokkal több él kerül be a szintaktikai, és ezáltal a szemantikai reprezentációba is.

Az itt felvázolt, jóindulatú interpretációval párosított felszíni szintaktikai elemzés módszere egyértelműen olyan esetekben használható sikerrel, ahol i) a fedés sokkal alapvetőbb fontosságú, mint a pontosság, és ii) a célszövegek megfelelő jellegűek, azaz egységenként relatíve rövidek, és nem tartalmaznak nagyon hasonló jellegű állításokat. Mind a szabadalmak, mind a cukorbetegséggel kapcsolatos ajánlások közötti keresés megfelel az i) pontnak, azonban míg az utóbbi a ii)-at is teljesíti, ennek a feltételnek a szabadalmi szövegek nem tesznek eleget. A szabadalmi szövegekre megfelelő kötöttebb mélyelemző ezzel szemben a cukorbetegséggel kapcsolatos ajánlásokon bukik el azoknak sokkal szabadabb nyelvtani szerkezetei miatt. Fontos tehát a keresési rendszer egyes moduljait minden esetben a tárgynak megfelelően megválasztani.

Hivatkozások

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL'98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998) 86–90
2. Das, D., Schneider, N., Chen, D., Smith, N.A.: Probabilistic frame-semantic parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010) 948–956
3. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics, **28**(3) (2002) 245–288
4. Kipper, K., Dang, H.T., Palmer, M.: Class based construction of a verb lexicon. In: AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin TX (2000)
5. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
6. Matsubayashi, Y., Okazaki, N., Tsujii, J.: A comparative study on generalization of semantic roles in FrameNet. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (2009) 19–27
7. Moldovan, D., Girju, R., Oltenau, M., Fortu, O.: SVM classification of Framenet semantic roles. In: SENSEVAL-3 (2004)
8. Palmer, M., Gildea, D., Xue, N.: Semantic Role Labeling. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
9. Pradhan, S., Ward, W., Hacioglu, K., Martin, J., Jurafsky, D.: Shallow semantic parsing using support vector machines. In: Proceedings of HLT/NAACL (2004) 233–240
10. Shi, L., Mihalcea, R.: Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: Computational Linguistics and Intelligent Text Processing (2005) 100–111
11. Stav, A.: Shallow parsing. Seminar in Natural Language Processing and Computational Linguistics (2006)

12. Szőts, M., Gyarmathy, Zs., Simonyi, A.: Frame-szemantikára alapozott információvisszakereső rendszer. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 275–288
13. Szpektor, I., Dagan, I.: Augmenting WordNet-based inference with argument mapping. In: Proceedings of the 2009 Workshop on Applied Textual Inference (2009) 27–35
14. Thompson, C.A., Levy, R., Manning, C.D.: A generative model for semantic role labelling. In: Senseval-3 (2003) 397–408

Az Európai Médiafigyelő (EMM) magyar változata

Pajzs Júlia

MTA Nyelvtudományi Intézet Nyelvtechnológiai Kutatócsoport
1394 Budapest Pf. 360
pajzs.julia@nytud.mta.hu

Kivonat: A Közös Kutatóközpont – Europa (European Joint Research Centre) által fejlesztett európai médiafigyelő (http://emm.newsbrief.eu) világszerte több ezer hírportálról automatikusan gyűjti, és különféle kategóriákba sorolja a híreket, a nap 24 órájában, 10 percenként frissítve, nyelvtechnológia eszköztár használatával. Az MTA Nyelvtudományi Intézet Nyelvtechnológiai Kutatócsoportja együttműködési megállapodás keretében a szolgáltatás magyar nyelvű működését tette lehetővé. A magyar tulajdonneveknek az EMM rendszeren belüli felismerése és a toldalékolt változatok kezelése volt az elsődleges feladat. A nemzetközi jelentőségű híreket valamennyi feldolgozott nyelvi változatukban elérhetjük.

1 Bevezetés

Az Europe Media Monitor (EMM) teljesen automatikus médiafigyelő rendszer lehetővé teszi, hogy a felhasználók naprakészen tájékozódjanak az on-line média őket érdeklő tartalmairól. Több tucatnyi különböző nyelvből nyelvtechnológiai eszközök segítségével összegyűjti a híreket, és részleges elemzést, információkivonatolást hajt végre rajtuk. Mivel a rendszer megalkotói sok nyelv hatékony feldolgozását tűzték ki célul, nem kívántak az egyes nyelvek morfológiai, szintaktikai, szemantikai elemzésére az egyes nyelvekre kifejlesztett eszközöket alkalmazni. Némelyik korábban feldolgozott nyelv esetén járható út volt a várható toldalékolt alakok listaszerű feldolgozása [5], ez a magyar szövegekre nem lett volna reális célkitűzés [4]. A magyar modul illesztéséhez számos segédanyagot készítettünk és adtunk át, valamint az eredmény tesztelésében nyújtottunk segítséget. A cikkben az EMM rövid általános ismertetésén kívül az átadott anyagokat és az eredményeket ismertetem.

1.1 Információkinyerés az EMM-ben

A hírek klaszterekbe sorolása 10 percenként frissül. Ha egy hír jelentős részben azonos egy néhány órán belül korábban megfigyelt hírcsoport elemeivel, ennek a klaszternek tartalmához adódik. (Részletesebben kifejtve lásd [2, 3]).

Előre elkészített többnyelvű kategóriadefiníciókat tartalmazó állományok segítségével a klasztereket automatikusan témakörökbe sorolja a rendszer (természeti csapások, terrorizmus stb.).

Különféle segédállományok (ismert személynevek, titulusok, foglalkozások, népnevek stb. listák) felhasználásával igyekszik automatikusan felismerni a szövegben előforduló személyneveket. A napi hírösszesítőben (NewsExplorer) feltünteti a leggyakrabban szereplő személyneveket, ezek kapcsolatrendszerét más személyekkel, egyéb fontos neveket (pl. intézmények).

A hírek címében felismert földrajzi nevek alapján a világtérképen is elhelyezi a hírklasztereket.

A több nyelven megjelent azonos híreket mindegyik, az EMM által feldolgozott nyelven megtekinthetjük.

2 A magyar források

2.1 Javaslat a feldolgozandó hírportálok bővítésére

Már korábban is figyeltek néhány magyar nyelvű portált. Ezek bővítésére tettem javaslatot. Azt tartottam szem előtt, hogy a politikailag különböző térfélen állók képviselve legyenek. Fontos kiegészítés volt a határon túli magyar nyelvű portálok hozzáadása a rendszerhez, amely így jelenleg 66 magyar nyelvű portált kezel (http://emm.newsbrief.eu/NewsBrief/sourceslist/hu/list.html).

2.2 A személynevek felismerését segítő anyagok

- Az aktuálisan érvényes magyar keresztnév listák (3215 elem).
- Titulusok listája (úr, asszony, hölgy) (kb. 400 elem).
- Fontos beosztások listája (pl. miniszterelnök) (kb. 650 elem).
- Foglalkozásnevek, kategóriákba sorolva (pl. kutatóorvos → HEALTH, RESEARCH).
- Népnevek (francia, gall) (kb. 730 elem).

E listák segítségével igyekeznek felismerni a *Lech Kaczynski lengyel elnök, Németh Lászlóné miniszter asszony* jellegű szerkezeteket.

A listákon a toldalékolható szavak végén szerepel a *„%”* karakter, jokerként (annak jelzésére, hogy a megadott szavakat egyéb karakterek követhetik). A változó tőalakokat is feltüntettem. Átadtam az egyszerű névszói toldalékok listáját is.

2.3 Idézetek felismerését segítő igék

Az alábbi igelistát adtam át, amelyek MNSZ-beli gyakoriságuk csökkenő sorrendjében szerepelnek: mond, jelent , elmond, ír, beszél, szól, közöl, kérdez, megállapít, jelez, kijelent, válaszol, hangsúlyoz, nyilatkozik, bejelent, megerősít, megjegyez, idéz, fogalmaz, beszámol, magyaráz, elárul, fenyeget, rámutat, tisztáz, felidéz, méltat, összegez, faggat, fenyegetőzik, nehezményez, deklarál, elbeszél, panaszol, tudakol.

Minden igét egyes és többes szám harmadik személyű, jelen és múlt idejű változatában adtam meg, az elváló és hátravetett igekötős igéknél ezeket a változatokat is feltüntettem.

2.4 Földrajzi nevek

A földrajzi neveket több nemzetközi adatbázist felhasználva dolgozták fel. A különböző listákból származó adatokat igyekeztek egyértelműsíteni [1]. A keletkezett adatbázis az egyes nevek földrajzi koordinátáit tartalmazza, valamint egy kódot, amely arra utal, mennyire nagy jelentőségű az adott helynév (ország, főváros, nagyváros stb.) Ezt az adatbázist kellett kiegészítenem részben magyar nevekkel, részben nemzetközi nevek magyar változataival (Wien→Bécs, Beijing→Peking). Valamennyi már korábban is meglévő nevet, szükség esetén ki kellett egészíteni olyan alakváltozattal, amely magyar toldalékok előtt állhat (Prága→Prágá%).

A feldolgozott magyar hírek címében augusztusban talált földrajzi nevek listáját ellenőriztem. A vizsgált nevek 40%-a toldalékolt formában fordult elő a szövegben. Az előfordult toldalékolt nevek 24%-a nemzetközi név volt, így beigazolódott, hogy nem csupán a toldalékolt magyar helynevek korrekt felismerése fontos. A jelenlegi megoldás elfogadható: minden legalább 5 karakter hosszú név végén ott van a „*%*" karakter, ami jelzi, hogy bármely karaktersorozat követheti a nevet. Ebből adódnak ugyan félreelemzések (pl. a *Gabonatermesztés*ről szóló cikket *Gabon* államnál helyezi el a térképen), a félreelemzések száma azonban a vizsgált egy hónapnyi mintában 4% alatt maradt. A félreértéseknek, félreelemzéseknek természetesen más forrása is van: nem egy keresztnév földrajzi név is egyúttal, de maguk a földrajzi nevek is sokszor utalnak különböző helyekre. A többértelműségeket és félreértéseket folyamatosan gyűjtjük, javítjuk.

2.5 Kategóriadefiníciók

A tematikus keresés lehetővé tételéhez többnyelvű kategóriadefiníciós állományokat használnak. Az egyes kategóriák definíciós állománya több részből áll össze az alábbiakban láthatunk erre példát. Az első részben azok a szavak szerepelnek, amelyek tipikusan előfordulhatnak az adott témájú hírekben. A szavak után látható szám az adott szó súlyára utal, minél nagyobb a szám, annál jellemzőbb a szó az adott kategóriára. A nagy negatív súllyal jelölt *film* és *könyv* szavak azt jelzik, hogy ha ezek a szavak is előfordulnak az adott hírben, ne tekintse a kategóriába tartozónak, hiszen akkor feltehetőleg egy ilyen témájú film vagy könyvismertetést tartalmaz a hír. Az állományok második részében szókombinációk megadására van lehetőség, bizonyos kombinációk ki is zárhatók (pl. ha a *bomba* szó közelében *foci, futball*, vagy *meccs* szerepel, ne sorolja a hírt a Terrorizmus kategóriába). A „*%"* karakter ebben a példában is a joker karaktert jelöli.

Az *Embercsempészet* témakör kategóriadefiníciós állománya

```
Alert definition
Alert ID: HumanTraffic

Description: Human Traffic
Patterns

Pattern Weight
emberkeresked% 20
ember%csempész% 20
illegális%bevánd% 20

film -999
könyv -999

A combination of at least one of
Proximity: 20
emberkeresked%
ember%csempész%
nő%keresk%
szex%rabszolg%
rabszolg%
kényszermunk%

and at least one of
szervez%+bűnöz%
prostitu%
áldozat%
gyermek%
csecsemő%
bevándorl%
```

3 Kiértékelés

3.1 Toldalékstatisztika

A különböző tulajdonnév listák kiértékelésének melléktermékeként toldalékstatisztika is készült. Ezek alapján megfontolásra érdemesnek tűnik, hogy csupán a leggyakoribb toldalékok felismerését célozzuk meg, néhány egyszerű reguláris kifejezéssel. Míg az összesített tulajdonnév listában a *t* tárgyrag különböző alakjai, a *bAn* és a *nAk* és a *vAl* fordultak elő leggyakrabban (az összes toldalékolt alak 90%-a), addig a földrajzi neveknél a bAn és az On toldalék alakjai szerepeltek nagyon gyakran (az összes toldalékolt alak 73%-a).

A részletes statisztikát az 1. táblázat tartalmazza.

1. táblázat: Toldalékstatisztika

Toldalék	Standard	Funkció	Fqs1	Fqs5	FqsTOP	Fqs700	Fq NewReC	FqGeo
AKAT	K+T	PL+ACC					1	
AS	S	N→A Der					1	
AT	T	ACC	1	5			2	4
BA	BA	ILL	1	5		3	13	25
BAN	BAN	INE	1	35	187	83	45	498
BE	BA	ILL					1	13
BEN	BAN	INE		10	69	16	10	88
BÓL	BÓL	ELA		5			4	41
BŐL	BÓL	ELA					3	5
CSAL	VAL	INS			17	1		
CSEL	VAL	INS		5				
DAL	VAL	INS		5				
DEL	VAL	INS					2	
É	É	POS				1	6	4
ÉK	ÉK	Der			19	31	5	
ÉKKAL	ÉK+VAL	Der+INS		5		5		
ÉKBÓL	ÉK+BÓL	Der+ELA				2		
ÉKNAK	ÉK+NAK	Der+DAT				2		
ÉKON	ÉK+ON	Der+SUP				2		
ÉKRÓL	ÉK+RÓL	Der+DEL				1		
ÉKTÓL	ÉK+TÓL	Der+ABL				3		
EN	ON	SUP		20		11	8	300
ÉRT	ÉRT	CAU					5	9
ET	T	ACC	1	15	34	6	19	12
ETT		SUP				1		
FÉLE	FÉLE	N→A Der				5		
FAL	VAL	INS						
GAL	VAL	INS		5			6	9
GEL	VAL	INS		5				
GYEL	VAL	INS			20		1	
HEZ	HOZ	ALL					1	
HOZ	HOZ	ALL		5		28	9	13

IG	IG	TER						1
JÁT	JA+T	PERS+ACC					1	
JE	JA	PERS					1	
JÉT	JA+T	PERS+ACC					3	
KAL	VAL	INS					1	
KEL	VAL	INS						1
KÉNT	KÉNT	FOR		5				
LAL	VAL	INS					1	
LEL	VAL	INS		10			3	
LYAL	VAL	INS				8	2	
LYEL	VAL	INS				4		
MAL	VAL	INS		5		2		3
MEL	VAL	INS					1	
N	ON	SUP	1	5		6	13	79
NAK	NAK	DAT	3	40	76	119	54	31
NEK	NAK	DAT	1	40	179	26	39	4
NAL	VAL	INS		20		18	8	52
NÁL	NÁL	ADE		5		4	9	
NEL	VAL	INS	1	5			3	2
NÉL	NÁL	ADE	1	10			1	18
ON	ON	SUP	1	15		6	16	189
ÖN	ON	SUP						6
OT	T	ACC	1	15	24	10	20	41
ÖT	T	ACC		5	57			
RA	RA	SUB	1	5		18	34	6
RE	RA	SUB		5		1	2	26
REL	VAL	INS				2	5	
RAL	VAL	INS		5		5		
RÓL	RÓL	DEL		10		7	21	9
RŐL	RÓL	DEL		10			2	11
SAL	VAL	INS				11	1	
SZAL	VAL	INS					3	
SZEL	VAL	INS					1	
T	T	ACC	9	110	197	239	130	33
TAL	VAL	INS	1				3	

TEL	VAL	INS			19	3	2	
TÓL	TÓL	ABL	1	25		35	24	22
TŐL	TÓL	ABL	1			3	14	16
UK	JUK	PERS				10		
VAL	VAL	INS	2	30	15	18	13	7
VEL	VAL	INS		10		1	4	
ZEL	VAL	INS					1	
Summa			**28**	**515**	**913**	**757**	**578**	**1578**

Fqs1 Az 1 gyakorisággal előforduló (Sample 1) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fqs5 Az 5 gyakorisággal előforduló (Sample 5) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

FqsTop A leggyakoribb (legalább 15) gyakorisággal előforduló (Sample Top) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fqs700 A teljes felismert tulajdonnév lista „magyar" elemeiből 700 tétel (Sample 700) kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fq NewRec A teljes felismert tulajdonnév lista „magyar" elemeiből 700 tétel (Sample New Rec) kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

FqGeo Az augusztusban felismert földrajzi név lista (Geo Names) kézi javítása alapján észlelt összegzett toldalék gyakoriság.

3.2 Toldalékolt tulajdonnevek részaránya

Mivel a magyar morfológiai eljárások alkalmazását az EMM fejlesztői lehetőleg el szeretnék kerülni, igen fontos kérdés volt, hogy a felismert tulajdonnevek hány százaléka volt toldalékolva, azaz valójában mennyi az információveszteség abból adódóan, hogy *Bajnai Gordon, Bajnai Gordonékról, Bajnai Gordonékat* 3 különböző tétel a felismert nevek adatbázisában. A rendszer egyéves működése után készültek a teljes felismert tulajdonnév lista különböző részeiből azok a kézzel ellenőrzött listák, amelyeknek összesített és toldalékolt type/token arányát mutatja be a 2. táblázat.

2. táblázat: Type/Token arány

	Type	Token	Toldalé-kolt Type	Toldalé-kolt Token	Told Type/Type	Told Token/Token
Teljes névlista	22.464	102.041				
Sample1	100	100	28	28	0,280	0,280
Sample5	771	3. 855	103	515	0,133	0,133
SampleTop	1. 057	48 .938	37	913	0,035	0,0186

Sample 700	700	9.056	197	760	0,281	0,083
SampleNewRec	1.167	2.136	187	330	0,160	0,154
GeoNames	788	4.056	350	1581	0,444	0,389

3.3 Tulajdonnevek és idézetek felismerése

Azonos napon megjelent 11 hírből gyűjtöttem ki a bennük előfordult 100 személynevet, a rendszer csak a felét ismerte fel személynévként. Ez azzal magyarázható, hogy csak akkor tekinti az egymást követő nagybetűs szavakat személynévnek, ha a) már a rendszer által ismert személynév b) a nevet követően a személynév felismeréshez megadott különféle listák elemei közül legalább egyre illeszkednek a nagybetűs szavakat követő szavak. Így természetesen felismeri *Angela Merkel*t (15 előfordulás) és *Orbán Viktor*t (3 előfordulás), de nem ismeri fel *Csík János zenekarvezető*t, mivel ez utóbbi nem szerepelt a foglalkozásnevek listáján, ugyanezért *Varga Gábor hatóanyag-szakértő* is ismeretlen marad az EMM számára.

Az idézetek felismerésének aránya sajnos még ennél is rosszabb. Ez részben a személynév felismerés hiányosságából adódik. További probléma, hogy csak akkor tekint idézetnek egy szövegrészletet, ha idézőjelben van, és ugyanabban a mondatban szerepel a felsorolt igék (*mond, jelent* stb.) valamelyike és egy felismert személynév. Emellett túl szigorú volt az ige és a felismert személynév együttes előfordulásnak szabálya is (csak egy-egy előre megadott listán felsorolt szót engedtek meg közöttük: pl.: *tegnap, korábban, délelőtt* stb.), ezt a szabályt időközben javították (legfeljebb 3 bármilyen szó lehet az ige és a személynév között).

200 hír kézi ellenőrzésnél azt tapasztaltam, hogy az általam észlelt 120 idézetből mindössze 9-et azonosított a rendszer! 36 esetben adódott a hiba a személynév felismerés hiányából. 6 esetben az igét nem ismerte fel (például mert az elváló igekötő nem közvetlenül az ige után volt), egyes esetekben felmerült az igelista kibővítésének igénye is: *tette hozzá, zárta, véli*. Összességében úgy tűnik, a fontos közszereplők egymondatos idézeteinek van esélye a korrekt felismerésre. Bár ezek a részeredmények meglehetősen gyengének tűnnek, az EMM fő célkitűzése lényegében teljesülni látszik: a hírekben rendszeresen, gyakran szereplő fontos személyek (főként vezető politikusok) nézeteinek és kapcsolatrendszerüknek számontartása.

3.6 Témakörök szerinti keresés

A 2.5-ben definiált embercsempészet témakör keresésének eredménye

Két szabadkai embercsempészt fogtak el a rendőrök Mórahalomnál

16 felnőtt és gyermek határsértő akart két személyautóba bezsúfolódni, amikor tetten érték őket a

Szegedi Határrendészeti Kirendeltség járőrei Mórahalom külterületén

Segítséget kap Bulgária a menekültügy megoldásához

Technikai és pénzügyi segítséget nyújt az ENSZ Menekültügyi Főbiztossága és az Európai Unió is....

Hihetetlen: harminc éve rabszolgasorban tartott nőt szabadítottak ki

Három, harminc éve rabszolgasorban tartott nőt szabadított ki a brit rendőrség Londonban. Ketten külföldiek, a harmadik a fogságban születethetett. A rabszolgatartókat letartóztatta a brit rendőrség....

"Jól felszerelt" embercsempészeket buktattak le a magyar határrendészek

16 helyszínen 3 embercsempészt és összesen 91 határsértőt fogtak el a Szegedi Határrendészeti Kirendeltség járőrei a polgárőrökkel együttműködve Csongrád megye déli részén 24 óra alatt....

Éjjellátót is vitt magával a fülön csípett embercsempész

Éjjellátóval és mobiltelefonokkal szerelkezett fel egy Ásotthalomnál elfogott embercsempész....

4 Összegzés

Az EMM magyar modulja működőképes. A személynevek esetében a toldalékolt alakok aránya viszonylag alacsony (általában 10% alatt, a leggyakoribb nevek esetén mindössze 1,8%) ezért a morfológiai elemzés hiánya nem okoz jelentős problémát. A már ismert neveket biztonságosan felismeri a rendszer, az addig ismeretleneket csak akkor, ha kellőképpen fontos pozíciójuk vagy foglalkozásuk van és ez közvetlenül a név után, ugyanabban a mondatban expliciten megjelenik. A földrajzi nevek felismerésének aránya megfelelő (96%). A kiemelt témakörökre a tematikus hírkeresés használható. Jól működő nemzetközi hírfigyelő rendszerbe sikerült beillesztenünk a magyar modult.

Hivatkozások

1. Pouliquen, B.; Kimler, M. Steinberger, R., Ignat, C., Oellinger, T., Blackler, K. Fluart, F., Zaghouani, W., Widiger, A Forslund, Clive: Best Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation LREC (2006)
2. Steinberger, R., Pouliquen B., van der Goot E.: An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), Boston, USA. (2009) 1–8
3. Steinberger, R.: A survey of methods to ease the development of highly multilingual Text Mining applications. Language Resources and Evaluation Journal, Springer, Volume 46, Issue 2, (2012). 155–176
4. Steinberger, R., Maud, E., Pajzs, J., Mohamed, E., Steinberger, J. Turchi, M.: Multilingual media monitoring and text analysis - Challenges for highly inflected languages. In: Habernal, I., Matoušek, V. (eds). Text, Speech and Dialogue. 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 2013, Proceedings. Springer Lecture Notes in Artificial Intelligence LNAI 8082 (2013) 22–33
5. Steinberger, R., Ombuya S., Kabadjov M., Pouliquen, B., Della Rocca, L., Belyaeva, J., De Paola. M., van der Goot, E.: Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources and Evaluation Journal (DOI 10.1007/s10579-011-9165-9), Volume 45, Issue 3 (2011) 311–330

Magyar társadalomtudományi citációs adatbázis: A MATRICA projekt eredményei

Váradi Tamás, Mittelholcz Iván, Blága Szabolcs, Harmati Sebestyén

MTA, Nyelvtudományi Intézet,
Benczúr utca 33., 1068 Budapest
e-mail: {varadi, mittelholcz}@nytud.mta.hu
{szabolcs.blaga, harsej}@gmail.com

Kivonat A szabad szövegekből történő strukturált információkinyerés egy sajátos területe a tudományos közlemények feldolgozása. Ezen belül is különösen fontos feladat a közleményekben szöveges alakban található hivatkozások kinyerése, elemzése és adatbázisba építése.[1] Ez röviden a célja a MATRICA (*Magyar Társadalomtudományi Citációs Adatbázis*) projektnek, ami a 2010-ben forráshiány miatt megszakadt HUN-ERIH projekt folytatása.[2] A projekt megvalósítása során, különösen a majdani felhasználókkal, az egyetemi könyvtárakkal való együttműködés eredményeként az alábbi prioritások alakultak ki: 1) tudományos cikkek feldolgozása a nyers fájloktól az adatbázisig, 2) ahol lehet, ott az emberi közreműködés csökkentése, 3) ahol kell, ott a szükséges emberi beavatkozáshoz kényelmes webes felület biztosítása.

1. Bevezetés

Napjaink tudományos életében a kutatókra és könyvtárosokra egyre nagyobb terhet ró a bibligráfiai adatok rögzítése és követése. Ezért is tartotta fontosnak az MTA egy olyan technológiai lánc kifejlesztését, ami alkalmas nagy mennyiségű, elektronikus formában elérhető cikk bibliográfiai adatainak feldolgozására; számítógépes eszközökkel támogatva meg az eddig jellemzően kézi munkával végzett hivatkozásfeldolgozást. A technológiai lánc mellett fontos, hogy a társadalom- és bölcsészettudományi területen Magyarországon még nem létezett egy átfogó bibliográfiai adatbázis, amely természettudományi területen már adott. Ennek létrehozása volt a Matrica projekt másodlagos célja.

2. Kapcsolódó munkák

Az interneten szabadon elérhető és kipróbálható hivatkozásfeldolgozó szoftverek[3] elsősorban az egyéni munka segítésére hivatottak: az egyes kutatók dolgát

[1] L. [2,3]
[2] A projekt előző szakaszáról l. [1].
[3] L. többek közt az alábbiakat:
cb2bib (http://www.molspaces.com/d_cb2bib-overview.php),

könnyítik meg a saját bibliográfiájuk összeállításában. A mi célunk két dologban tér el ettől. 1) Mivel nagy mennyiségű és heterogén publikáció feldolgozását tűztük ki, ezért nem fogadhattuk el a feldolgozás olyan fokú pontatlanságát, ami a személyes használatra szánt programokat jellemzi, mivel a kézi javítás ekkora mennyiségben már nem gazdaságos. Az általunk kezelt anyag hivatkozási konvencióinak heterogenitása miatt szintén nem volt céljainknak megfelelő egy olyan szabályalapú megközelítés, amely csak néhány hivatkozási sztenderdet képes kezelni. 2) Mivel alapvetően egy közös bibliográfiai adatbázis létrehozásában gondolkodtunk, elengedhetetlen volt a kollaboratív munka támogatása egy webes felületen keresztül.

3. Nyers hivatkozások kinyerése

3.1. Fájlok

A MATRICA projektben a HUN-ERIH alatt összegyűjtött anyagot örököltük, azt már külön nem bővítettük és nem frissítettük, csak a feldolgozására koncentráltunk. A HUN-ERIH projekt vállalása a kortárs, magyarországi (de nem feltétlenül magyar nyelvű) bölcsészet- és társadalomtudományi folyóiratok feldolgozása volt öt évre visszamenőleg. Igyekeztünk minél szélesebb körből meríteni, és a kiadókkal való egyeztetések után végül 192 folyóirattól sikerült anyagot szereznünk.

A folyóiratok rendelkezésünkre bocsátott állománya nagyon heterogén, mind a fájlok terjedelmét, mind azok formátumát tekintve.

A folyóiratok egy része minden cikket külön fájlban tárolt, másik része folyóiratszámonként, harmadik része évfolyamonként bontotta fájlokra az anyagot. Ez jelentősen megnehezítette a cikkek beazonosítását és a cikkekre vonatkozó metaadatok kinyerését: a fájlokat először feldaraboltuk cikkekre, azután az egyes cikkekből egyrészt az azok azonosításához szükséges úgynevezett fejléc adatokat, másrészt a cikkhez tartozó összes hivatkozás nyers alakját nyertük ki.

Ami a fájlok formátumát illeti, a hét különböző formátum közül a HTML (43%) és a PDF (51%) bizonyult a leggyakoribbnak. A HTML-fájlokhoz képest a PDF-állományok feldolgozása jelentős többletmunkával járt.

3.2. PDF feldolgozás

A PDF fájlok szerkezete nagyon egyszerű, alapvetően minden egyes karakter abszolút geometriai pozícióját adja meg egy adott hordozón (előre megadott méretű téglalap alakú területen – papíron). Az abszolút pozíció megadása egy kétdimenziós koordináta-rendszer segítségével történik, hasonlóan az egyes karakterek kiterjedéséhez. Ezen felül szerepel a karakter mérete, amely így nem feltétlenül tölti ki a számára megadott téglalap alakú területet, valamint az

text2bib (http://text2bib.economics.utoronto.ca),
Free Cite (http://freecite.library.brown.edu/welcome),
Simple Text Query (http://www.crossref.org/SimpleTextQuery/).

aktuális karakterkódolási táblázat szerinti kódja (ami megadja, hogy milyen betű jelenjen meg vizuálisan), illetve a használt betűtípus. Karakternél magasabb rendű szövegbeli egységek (szó, sor, bekezdés, stb.) a karakterek csoportosításával jönnek létre. Ugyanakkor a legtöbb esetben a szöveg magasabb rendű felépítése nem megbízható (nem tükrözi a forrás struktúrális elrendezése a hordozón látható vizuális elrendezést), ezért a legtöbb esetben a szöveg struktúráját a karakterszintű elemek pozícióinak elemzéséből algoritmikusan kell rekonstruálni. További nehézséget jelent, hogy a PDF belső szerkezete jelentős rugalmasságot biztosít az előállításkor, így a különböző folyóiratok között szinte minden esetben, de akár a folyóiratok egyes számain belül is változhat a PDF belső szerkezete, attól függően, hogy milyen alkalmazással készítették az adott állományt. Bár a PDF belső szerkezete egységes keretet ad a dokumentumok felépítéséhez, mégis a különböző PDF-készítő programok más-más egyedi mechanizmus mentén nagyon eltérő belső struktúrájú fájlokat hoznak létre.

A belső szerkezet változásairól sok esetben a készítőnek sincs tudomása, ezért erről semmiféle analitikus információ nem áll rendelkezésünkre, tehát olyan általános feldolgozó eljárást kellett kialakítanunk, ami a PDF-fájlok egy nagyon diverz halmazára alkalmazható.

Az egyik jellemző probléma a többhasábos elrendezésű szövegek kezelése, itt sok esetben a különböző hasábokhoz tartozó azonos magasságban lévő szövegrészek egy sorként voltak tárolva PDF belső szerkezete alapján, így ezeknél a karakterek közötti térköz vizsgálatával kellett visszaállítani az eredeti többhasábos struktúrát.

Mivel egy adott PDF-állományban több cikk is szerepelhetett egyszerre, ezért a cikkek elhatárolásához és egy adott cikk metaadatainak megtalálásához olyan feltételrendszereket kellett kidolgozni, melyek egyértelműen beazonosítanak egy adott szövegrészt. A beazonosításhoz szükség volt a magasabb szövegbeli egységek helyes felismerésére, illetve a különböző formázási elemek egységes kezelésére. Itt kihívást jelentett a címben, szerzők nevének megadásánál és a hivatkozásoknál is előszeretettel használt ún. kiskapitális írásmód kezelése. Sokszor a kiskapitális írásmód PDF-en belüli megvalósítása azt jelentette, hogy a csupa nagybetűvel írt szövegben változott az egyes karakterek mérete, ez normál szöveggel, vagy esetenként egyszerű nagybetűs írásmóddal keverve nehezen kezelhető, körültekintő mérlegelést igényel a feldolgozó algoritmust paraméterezők részéről. Természetesen néhány esetben a többértelműség nem oldható fel algoritmikusan, vagy csak túlzott fejlesztési erőforrásigény mellett, ezért a manuális javítás a jobb megoldás.

További nehézséget jelent a PDF-állományok eltérő karakterkódolása. Mivel a PDF lehetővé teszi az egyes szövegrészek közötti eltérő kódolási táblák használatát, ezért ezek kezelése sokszor külön óvatosságot igényel. A legnehezebben azok az esetek kezelhetőek, mikor a karakterek kódolásából nem, csak az adott betűtípus neve és megjelenése alapján derül ki, hogy milyen karakterek vannak kódolva az adott szövegrészben. Mivel az állományokban lehetségesen használható betűtípusok száma nagyon nagy, ezért ezek az esetek is csak egyedileg, speciális cseretáblák segítségével, vagy manuális javítással kezelhetőek.

Mivel a PDF-ek belső szerkezete jelentős eltéréseket mutatott, ezért tűnt jó megközelítésnek egy lépésben megpróbálni olyan feldolgozót fejleszteni, ami minden lehetséges típusra megoldást kínál. A hatékony fejlesztés érdekében egyfajta evolúciós megközelítést használtunk, ami abból állt, hogy mindig visszavisszatérő módon fejlesztettük az algoritmusokat, hogy egyre nagyobb számú jelenséget legyenek képesek kezelni. A PDF-elemzés evolúciós fejlődése a feldolgozás előrehaladtával:

1. Dokumentumok elemzése, tipikus esetek kiválasztása.
2. A felmerült problémák kezelésére alkalmas elemző fejlesztése.
3. Az elkészült elemző alkalmazása minél többféle dokumentumtípusra.
4. Kimeneti pontatlanságok elemzése, elemző hibáinak feltárása.
5. Vissza az 1-es ponthoz.

A fejlesztési ciklusok során az egyik legfontosabb feladat annak eldöntése, hogy az adott probléma érdemes-e arra, hogy specifikus fejlesztést eszközöljünk az elemző programban, vagy hatékonyabb egyedi esetként kezelni, így spórolva a jelentős erőforrásigényű algoritmus fejlesztéssel a viszonylag ritkán előforduló „speciális" esetekben.

A feldolgozó fejlődésével párhuzamosan bővült a projektbe bevont csoportok köre, míg kezdetben csak a fejlesztői csapat dolgozott a problémákon, később a tesztelők és paraméterezők folyamatos bevonásával jelentős párhuzamosítást értünk el az egyes munkafázisokban és a csoportok egymás közti kommunikációja alapján minden csoport hatékonysága dinamikusan fejlődött. A kézi ellenőrzés jelenlegi szakaszban a nyers hivatkozások PDF-ekből való kinyerése 49,2%-os pontosságot mutat.

Az evolúciós fejlesztési ciklusok során fontos szempont a visszafelé kompatibilitás megőrzése, vagy az annak elvesztéséből származó munkaterhelés minimalizálása, ebben a tekintetben is egyensúlyra törekedtünk. Míg kezdetben gyorsan változott a feldolgozó program, a munka kiterjesztésével párhuzamosan a stabilitás is egyre fontosabbá vált.

A citációs adatbázis jövőbeni fejlődése és fenntarthatósága szempontjából jelentős előrelépés lenne, ha az egyes kiadók és szerkesztők olyan metainformációkkal látnák el kiadványaik elektronikus változatát, ami megkönnyíti az automatikus feldolgozást. Még jobb lenne, hogyha ez a formátum egységes lenne az egyes kiadványok között. A Matrica adatbázisba bekerülő cikkek esetében már bármilyen kimeneti formátum előállítható a későbbiekben.

4. Hivatkozások elemzése

A különféle formátumú fájlok feldolgozása és a nyers hivatkozások kinyerése után a következő lépésben ezen hivatkozások feldolgozása történik. A HUN-ERIH projekt során erre a célra a NooJ szoftvercsomagot[4] használtuk, amely lokális grammatikákat használ az egyes hivatkozáselemek (szerző, cím, kiadó stb.) felismerésére, majd ezek megfelelő kombinációit illeszti a hivatkozások különféle típusaira.

[4] http://www.nooj4nlp.net/pages/nooj.html

Ezzel a szabályalapú módszerrel meglehetősen alacsony F-mértékeket kaptunk egy kismintás kiértékelés során, valamint nem bizonyult elég robusztusnak a rendkívül heterogén adathalmazon. (A rendszer leírását és az eredményeket lásd az [1] cikkben.) Ezért döntöttünk úgy, hogy a projekt folytatásában statisztikai alapú gépi tanuló megoldást alkalmazunk. A maximum entrópián alapuló HunTag[5] rendszert választottuk, amelyet eddig főnévi csoportok ([4]) és tulajdonnevek ([5]) felismerésére használtak, de bármilyen szekvenciális címkézési feladatra alkalmas, így a hivatkozások parszolására is.

4.1. Az adathalmaz

A hivatkozások hasznos bibliográfiai adatmezőinek definiálásához a BibTeX szabványt vettük alapul, és az alábbi tizenkilenc mezőt határoztuk meg: szerzők, szerkesztők, cím, kötetcím, sorozat, kiadás, kiadás helye, folyóirat, kiadó, iskola (téziseknél), szervezet (konferenciáknál), intézmény (egyéb esetben), év, hónap, kötet, szám, oldalszám, megjegyzés (pl. ki fordította) és URL. Ezeken felül további öt olyan mezőt használunk, amelyeket a hivatkozások lényegi információt nem hordozó, de valamilyen pozíciót jelző elemeinek tartunk fent, mint például a szerkesztőket jelző *szerk.*, *eds.* vagy éppen *hrsg.* Hasonló mezőket definiáltunk a folyóiratszámokat és évfolyamokat jelző bibliográfiai elemeknek (pl. *vol*, *num*) és az oldalszámoknak (pl. *o.*, *p.*) is.

Tanítás és tesztelés céljára egy 12.000 hivatkozást tartalmazó mintát választottunk ki véletlenszerűen. A minta kézzel való felcímkézését diákok végezték, amit szakértő könyvtárosok ellenőriztek. Ezt az adathalmazt utólag kézzel szűrtük, hogy még tisztább tanító és kiértékelő anyaghoz jussunk, így egy kb. 10.000 hivatkozást tartalmazó gold standard korpuszhoz jutottunk. Ezt használtuk 80%/20%-os vágásban tanításra és kiértékelésre.

4.2. Jegykinyerés

A tanítás során a legfontosabb sztring értékű felszíni jegyek (karakter n-gram, a token n karakterből álló előtagja és utótagja) optimális kombinációját a teljes paramétertér bejárásával állapítottuk meg. Minden paraméterkombinációt ötszörös keresztvalidációval kimértünk, és az összesített F-mértékek alapján az 1-es n-gram, 5-ös előtag, 3-as utótag jegykombináció bizonyult a legjobbnak. Az 1-es n-gram rendre jobb teljesítményt nyújtott a többi felszíni jegy eltérő értékei mellett is, ezért elfogadtuk. A tanításhoz felhasználtunk városok, kiadók és folyóiratok neveit tartalmazó listákat is.

4.3. Kiértékelés

A kiértékelést a fent leírt gold standard adathalmazon végeztük, ötszörös keresztvalidációt alkalmazva. A táblázatban látható eredmények azt mutatják, hogy a

[5] https://github.com/recski/HunTag/

gyakori (és egyben fontos) mezők F-mértéke általában 90% felett van, míg a ritkán előforduló mezők várható módon rosszabb eredményt adnak.

mező	**pontosság**	**fedés**	**F-mérték**
szerzők	96,93	97,57	97,24
szerkesztők	91,60	91,56	91,58
cím	88,50	88,06	88,25
kötetcím	71,04	73,33	72,17
sorozat	31,91	28,86	30,31
kiadás	61,54	57,66	59,53
kiadás helye	92,02	91,37	91,69
kiadó	83,09	85,72	84,39
intézmény	53,01	54,63	53,81
szervezet	12,00	9,38	10,53
iskola	42,39	34,51	38,05
folyóirat	86,74	90,49	88,57
kötet	68,23	78,34	72,94
szám	75,62	70,12	72,77
év	97,67	94,30	95,95
hónap	65,26	55,11	59,76
oldalszám	95,79	95,10	95,44
megjegyzés	70,81	61,80	66,11
url	83,57	80,09	81,71
összesített	88,81	88,33	88,57

Külön említést érdemel két mezőcsoport: egyrészt az évfolyam és szám, másrészt a intézmény–szervezet–iskola hármas. Mindkét csoport esetében hasonló környezetekben alternáló címkékről van szó. Folyóiratok esetében nem ritka, hogy az évfolyam és a szám közül csak az egyiket adják meg, pl.

> Baumrind, D. (1978): *Parental disciplinary patterns and social competence in children.* Youth and Society. 9. 239–276.

Ebben az esetben a *9* az évfolyam és a szám is lehet, a rendelkezésre álló kontextus alapján nem derül ki egyértelműen, hogy melyik.

Hasonló a helyzet a kiadói pozícióban álló mezők esetében is; ezek: a tézisek kiadói (iskolák), a konferenciakötetek kiadói (szervezetek) és az egyéb, publikációt megjelentető, de kiadónak nem tekintett intézmények. Ezek a mezők túl azon, hogy azonos pozícióban szerepelnek, hasonló (intézmény)neveket is tartalmaznak, ami jelentősen megnehezíti a megkülönböztetésüket, nem csak a gépi tanuló algoritmus, hanem az annotátorok számára is. Ebből kifolyólag már a gold standard adathalmazban sem egységes ezeknek a mezőknek a jelölése. Ezt a megkülönböztetést az indokolta, hogy a BibTeX sztenderd mezőihez igazodtunk, de a jövőben érdemes lenne ezeket összevonni egy intézmény jellegű mező alá.

A folyamat végén azokat a hivatkozásokat, amelyek egy előre meghatározott küszöbértéknél alacsonyabb valószínűségű mezőt tartalmaznak, utólagos ellenőrzésre ajánlja fel a rendszer. Ezzel két, külön forrásból származó hibatípust is ki tudunk küszöbölni. Egyrészt lehet maga a hivatkozás valamilyen szempontból különleges, ami miatt az elemző kimenete nem elég megbízható. Másrészt ha még az első lépésben nem megfelelően történt a nyers hivatkozás kinyerése (pl. folyó szöveg vagy csonka hivatkozás lett kibányászva), azt is jelezni fogja a rendszer a hivatkozás elemzésének alacsony valószínűségével.

5. Felület

A feldolgozás hatékony párhuzamosítása érdekében egy sokfelhasználós webes felület került kialakításra. A felület célja, hogy a gépi feldolgozás irányítása, ellenőrzése, a szükséges kollaborációs feladatok kivitelezése egy egységes keretben, felhasználóbarát módon mehessen végbe. A felület funkcionalitását négy felhasználói csoport szerint lehet felbontani:

1. A létrejövő citációs adatbázis jól struktúrált megtekintése és különböző keresési funkciók megvalósítása.
2. A szükséges kézi javítások és ellenőrzések elvégzése, az adatbázis minőségének javítása, szakértői csoportok bevonása a feldolgozás minőségének javítása érdekében.
3. Új adatok bevitele, az automatikus feldolgozás körén kívül eső folyóiratok hozzáadása az adatbázishoz.
4. Az automatikus feldolgozás paraméterezése a háttérben futó feldolgozási folyamatok és azok eredményének nyomon követése, elemzése.

A webfelület minden tekintetben igyekszik a mai kor elvárásai szerint megkönnyíteni a különböző felhasználói csoportok közös munkáját. Mivel a elemzési folyamatok jelentős erőforrásigénnyel bírnak, ezért az erőforrások optimális kihasználása érdekében egy aszinkron feldolgozási mechanizmus került megvalósításra, ahol az egy időben aktív felhasználók egy globálisan meghatározott erőforráskvótán osztoznak, így nagy terhelés mellett is elkerülhető a rendszer túlzott lelassulása, a felület válaszideje kielégítő marad.

6. Összefoglalás

Az elvégzett munka eredményeként olyan technológiai lánc állt elő, amely lehetővé teszi nagy mennyiségű, heterogén elektronikus szöveg bibliográfiai adatainak félautomatikus feldolgozását. Önálló fejlesztésünk a PDF-ek kezelését megkönnyítő szoftver, a statisztikai gépitanuló modul testreszabása és felkészítése a hivatkozások parszolására, valamint a kollaboratív webes felület. Munkánk másodlagos eredménye maga a folyamatos feltöltés alatt álló citációs adatbázis, amivel reményeink szerint könnyebbé tehetjük a kutatók és könyvtárosok ezirányú munkáját, hogy érdemi feladataikra jobban koncentrálhassanak.

Hivatkozások

1. Váradi T., Pintér T., Mittelholcz I., Peredy M.: Bibliográfiai hivatkozások automatikus kinyerése. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), Szeged, Magyarország, 56-65, (2010).
2. Bergmark, D.: Automatic extraction of reference linking information from online documents. TR2000-1821 (2000)
3. Day, M.-Y., Tsai, T.-H., Sung, C.-L., Lee C.-W., Wu, S.-H., Ong, C.S., Hsu, W.-L.: A knowledge-based Approach to Citation Extraction. In: Proceedings of the IEEE International Conference on Information Reuse and Integration. (IEEE IRI 2005). Las Vegas, Nevada, USA. (2005) 50-55.
4. Recski G., Varga D.: A Hungarian NP-chunker. The Odd Yearbook, (2009)
5. Simon E. Approaches to Hungarian Named Entity Recognition. PhD disszertáció. BME, Budapest, (2013)

VII. Poszterek

Természetes nyelvi korpusz vizsgálata egyeztetés-csoport módszerrel

Drienkó László

dri@t-online.hu

Kivonat: Korábbi munkáimból, [1, 2, 3], kiindulva egy olyan disztribúciós módszert szeretnék bemutatni, amely alkalmas lehet természetes (akár nem természetes) nyelvi szekvenciákat tartalmazó adatbázisokban rejlő, az adott nyelvre jellemző bizonyos disztribúciós szabályosságok detektálására, illetve új szekvenciáknak ezen szabályosságokon alapuló feldolgozására, a szabályosságokat megtestesítő szekvencia csoportokra való "leképzésére". Az alapvető fogalmak felvázolása, illetve a korábbi eredmények rövid ismertetése után néhány megjegyzés következik a módszer alkalmazhatóságával kapcsolatban, melyek nyomán az egyeztetés-csoportok által képviselt kutatás spektruma szélesedhet.

1 Bevezetés

Az egyeztetés-csoport módszer lényege, hogy egy „training" szekvencia-halmaz elemeiből csoportokat hozunk létre, és megvizsgáljuk, milyen mértékben képezhetők le egy tetszőleges új halmaz szekvenciái ezekre a csoportokra. A csoportosítás minimális különbségen alapul, azaz minden szekvenciához megkeressük azokat a szekvenciákat, amelyek csak egyetlen elemben térnek el tőle. Az „egyeztetés"- csoport elnevezés azon megfigyelésen/feltételezésen alapul, hogy amennyiben egy mondatban egy tetszőleges szót egy ugyanolyan „lexikai" kategóriájú szóval helyettesítünk és az így kapott új mondat nyelvtanilag helyes, akkor az eredeti mondat egyeztetés-viszonyai meg kell hogy őrződjenek, mivel a behelyettesített szónak rendelkeznie kell az eredeti mondat által megkövetelt egyeztetés-jegyekkel.

Az adott csoportokat táblázatos formában reprezentáljuk, ami a nyelvtani „következtetés", azaz az új mondatok feldolgozásának az alapja lesz. Például az (1)-ben megadott egyeztetés-csoport (2)-beli táblázatos formája lehetővé teszi (3) új mondatainak feldolgozását, azaz (1)-re való „leképzését".

(1)

Adam hates football
Adam hates **basketball**
Eve hates football
Adam **dislikes** football
Charles hates football

(2)

Adam	hates	football
Eve	dislikes	basketball
Charles		

(3)

Eve hates basketball	Eve dislikes football
Charles hates basketball	Charles dislikes football
Eve dislikes basketball	Charles dislikes basketball
Adam dislikes basketball	

2 Korábbi eredmények

2.1 Egyeztetés-csoport analízis

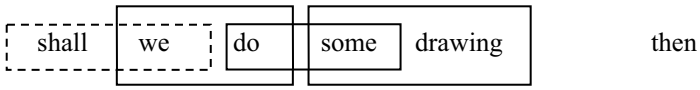
A módszert CHILDES [4] angol [5], magyar [6] és spanyol [7] gyermeknyelvi adatokra alkalmaztam. Mindegyik nyelv esetében az egyes felvételek idejét az adott anya-gyermek nyelv egy bizonyos fejlődési állapotának feleltettem meg és az addig elhangzott összes 2-5 szavas kijelentésből formáltam egyeztetés-csoportokat. Ezután megvizsgáltam, hogy a közvetlenül következő felvétel 2-5 szavas kijelentéseinek hány százaléka képezhető le az adott állapothoz tartozó egyeztetés-csoportokra. Azt találtam, hogy mindegyik fejlődési állapotban a kijelentések bizonyos hányada leképezhető a már meglévő csoportokra – enyhe növekedés is megfigyelhető. A leképzési értékek az angol esetben voltak maximálisak: egy esetben a soron következő felvétel 41% volt megfeleltethető az előzőekből nyert egyeztetés-csoportoknak. Az új kijelentések leképzési maximuma 10,3% volt.

2.2 Egyeztetés-csoport lefedhetőség

Következő lépésként [8,9] megvizsgáltam, hogyan lehet hosszabb kijelentéseket „lefedni” 2-5 szavas, egyeztetési csoportokra leképezhető szekvenciákkal. Például a

'shall we do some drawing then' mondat lefedhető a *'shall we'*, *'we do', 'do some' 'some drawing'* szekvenciákkal. Vö. (4).

(4)



Egy kijelentés lefedettség értékének kiszámítási módját (5)-ben vázoljuk.

(5)

1	2	3	4	5	6
shall	we	do	some	drawing	then
1	1	1	1	1	0

Lefedettség (Coverage): (5 szó a 6-ból): 5/6=83%

Az angol CHILDES adatokra 78% átlagos lefedettséget kaptam [8], a magyar nyelvi esetben 42%-ot [9]. A (4)-beli szaggatott vonal azt jelzi, hogy a 'shall we' szekvencia már konkrétan szerepelt a „training" halmazban, azaz nem új. Viszont megvizsgálhatjuk, hogy kifejezetten az új szekvenciák milyen mértékben járulnak hozzá a lefedettséghez. Ekkor pl. az (5)-beli érték 4/6= 66% lesz, mivel az 1. pozícióhoz 0 rendelődik. Az új szekvenciák az angol adatokra 49%-os, a magyar adatokra 29%-os átlagos lefedettséget eredményeztek.

Elvi síkon, a fent vázolt kísérletek kiindulópontjául szolgálhatnak egy olyan kétszintű nyelvi/nyelvtanulási modellnek, ahol egy alapvetőbb, elsődleges kognitív szint felel az egyszerűbb, kevésbé komplex megnyilvánulások feldolgozásáért – esetünkben ezt a szintet az egyeztetés-csoportok képviselik –, ugyanakkor a komplexebb struktúrák létrehozása, azaz az egyszerűbb megnyilvánulások egymáshoz rendelése, integrációja egy magasabb kognitív szinten történik. Hosszabb mondatoknak rövidebb fragmentumokkal való lefedhetőségét vizsgáló kísérleteink e második szintet próbálták vizsgálni.

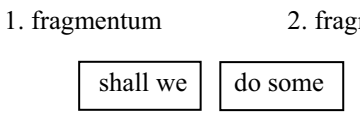
3 A módszer alkalmazhatóságával kapcsolatos megjegyzések

3.1 Fragmentum kombinációk

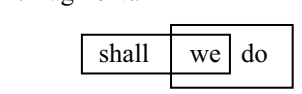
Mind nyelvelméleti, mind számítástechnikai szempontból fontos kérdés lehet a lefedhetőséget illetően, hogy milyen fragmentum kombinációk eredményezhetnek nyelvileg helyes megnyilvánulásokat. Ahogyan az (4)-ből is látszik, algoritmusunk alapjában véve kétféle fragmentum konfigurációt ismert fel, mivel feltételezte a fragmen-

tumok folytonosságát, vagyis azt, hogy bármely fragmentum bármely két eleme (szava) között nincs más fragmentumhoz tatozó elem. Ennek legkézenfekvőbb esetét mutatja (6), ahol a fragmentumok jól elhatároltan követik egymást. A folytonosság feltételezése persze nem zárja ki, hogy két fragmentum közé más elemek kerüljenek, illetve hogy bizonyos szélső elemek mindkét fragmentumhoz tartozzanak, mint például a '*we'* szó (7)-ben.

(6)



(7)

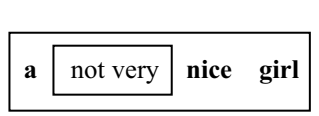


Elképzelhető azonban, hogy a fragmentum-folytonosság feltételezésének feladása nagyobb lefedettség értékekhez vezethetne, mivel összetettebb nyelvi szerkezetek is közvetlenül elérhetőek lennének. (8) például azt mutatja, hogyan ágyazhatók be egymásba fragmentumok: (8a)-ban nincs közös elem, (8b)-ben viszont a *nice* szó mindkét fragmentumhoz hozzátartozik. (8c) azt vázolja, hogyan fedhető le a klasszikus *The rat the cat the dog bit chased ate the cheese* szerkezet három fragmentummal. (9) példái egyfajta keresztfüggőség (cross serial dependency) hatást érzékeltetnek kettő (vö. 9a), illetve három (vö. 9b) fragmentummal.

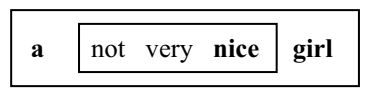
(10)-ben a beágyazás és a keresztfüggőség lehetséges kombinálására mutatunk példát: az első és második, illetve az első és harmadik fragmentum viszonylatában keresztfüggőséget látunk, ugyanakkor a harmadik fragmentum beágyazódik a másodikba.

(8)

a)



b)



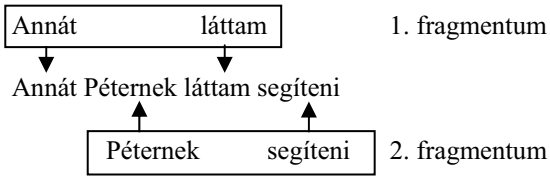
c)

1. fragmentum: **The rat ate the cheese** 2. fragmentum: *The cat chased*
3. fragmentum: The dog bit

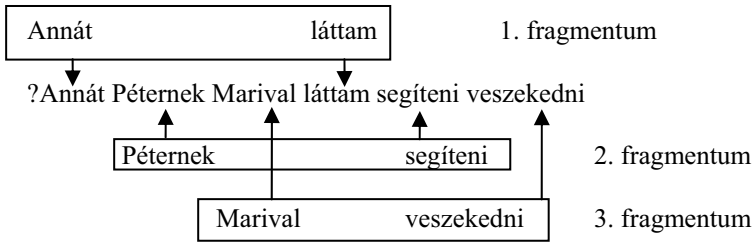
(**The rat** (*The cat* (The dog bit) *chased*) **ate the cheese**)

(9)

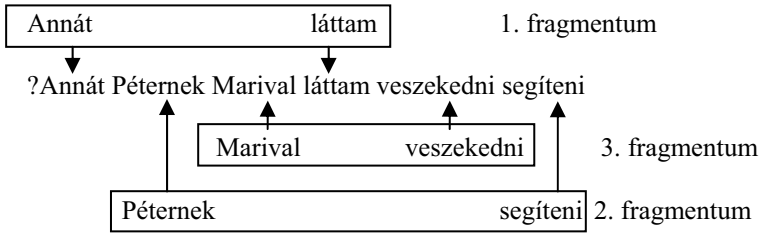
a)



b)



(10)



A fragmentum-folytonosság feltételezés feladása a jelenlegi algoritmus módosítását igényelné, ami viszont a számítástechnikai erőforrások bővítését tenné szükségessé.

3.2 Az egyeztetés-csoportok mint kognitív nyelvi struktúrák

A bemeneti adathalmaz strukturálásának eredményeképpen létrejövő egyeztetés-csoportok önmagukban is hordozhatnak releváns pszicholingvisztikai, illetve kognitív nyelvészeti információt az adott nyelv felépítését illetően. Az egyes csoportok mini-

málisan különböző szekvenciái alapjául szolgálhatnak olyan kategorizálási mechanizmusoknak, amelyek felelősek lehetnek a nyelv különböző szintű – lexikai/szintatikai, szemantikai, fonetikai, stb. – kategóriáinak kialakulásáért, továbbá a különböző nyelvi szintekhez tartozó jegyegyeztetési folyamatokért. (11a) mondataiban például az utolsó szó pozíciót igék töltik be, csakúgy mint (11b)-ben, ahol viszont szembeötlőbb az alany-ige egyeztetés – egyes szám első személy. A (11c)-beli főneveket szemantikailag egyfajta helymegjelölő funkció kapcsolja össze.

(11)

a)

you can't **reach** you can **reach** you can't **remember** you can't **know**
I can't **reach** you can't **see** you can't **play**

b)

*én nem én én nem **látom** én nem **játszok** én nem **tudom**
én nem **kéjek** én nem **szejetem** én nem **vagyok** én nem **tudok**
én nem **láttam**

c)

to the **shops** to the **hospital** to the **pub** to the **garden**
to the **seaside** to the **car** to the **farmyard**

Végül megemlítjük, hogy az egyeztetés-csoport módszer elvileg bármely olyan szekvencia halmaz esetén alkalmazható, ahol feltételezhető, hogy az egyes szekvenciák jólformáltságáért valamilyen mögöttes „szekvencia gyártó” mechanizmus felel.

Hivatkozások

1. Drienkó, L.: Agreement groups analysis of mother-child discourse. Talk presented at the 4th UK Cognitive Linguistics Conference, King's College London, UK (2012). To appear in Selected Papers from UK-CLA Meetings. Vol. 2.
2. Drienkó, L.: A linguistic agreement mapping-system model: agreement relations for linguistic processing. LAP-Lambert Academic Publishing (2012)
3. Drienkó, L.: Distributional cues for language acquisition: a cross-linguistic agreement groups analysis. Poster presentation for the 11th International Symposium of Psycholinguistics, Tenerife, Spain (2013)
4. MacWhinney, B.: The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates (2000)
5. Theakston AL, Lieven EV, Pine JM, Rowland CF.: The role of performance limitations in the acquisition of verb-argument structure: an alternative account. J. Child Lang. 28(1): (2001) 127–52
6. Réger, Z.: The functions of imitation in child language. Applied Psycholinguistics 7. (1986) 323–352

7. Montes, R. G.: Achieving understanding: Repair mechanisms in mother–child conversations. Unpublished doctoral dissertation, Georgetown University (1992)
8. Drienkó, L.: Agreement groups coverage of mother-child language. Talk presented at the Child Language Seminar, Manchester, UK (2013)
9. Drienkó, L.: Agreement groups coverage of Hungarian mother-child language. Poster presentation for the 11th International Conference on the Structure of Hungarian. Piliscsaba, Hungary (2013)

Kulcsszó-előfordulások relevanciájának vizsgálata magyar nyelvű hangzó híranyagokban[1]

Gosztolya Gábor

MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103.
ggabor@inf.u-szeged.hu

Kivonat: Kulcsszókeresés során a feladat felhasználók által beírt kulcsszavak előfordulásainak megtalálása nagyméretű hangadatbázisokban. Egy adott kulcsszókeresési rendszer pontosságának meghatározásához ismernünk kell a kulcsszavak valós előfordulásait, mely feladatra léteznek automatikus módszerek, azonban az, hogy ezek eredményei mennyire esnek egybe az emberi elvárásokkal, nem egyértelmű. Ennek vizsgálatához néhány tesztalanyt kértünk meg, hogy azonosítsák a számukra releváns kulcsszó-előfordulásokat. Válaszaikat több szemszögből elemeztük: megvizsgáltuk, használatukkal mennyire változik meg kulcsszókereső rendszerünk pontossága; elemeztük, mennyire esnek egybe a válaszok egymással; valamint azt is megnéztük, hogy az egyes alanyok jellemzően milyen jellegű előfordulásokat tartottak relevánsnak.

1 Bevezetés

A kulcsszókeresési probléma (Spoken Term Detection, STD [6]) egy viszonylag új beszédtechnológiai terület, melyben a feladat különböző, felhasználó által bevitt kulcsszavak előfordulásainak megtalálása egy nagyméretű hangadatbázisban. Bár hasonló alapokra építkezik, mint a beszédfelismerés, alapvető céljukban eltérnek: míg a beszédfelismerésé változó bemondásokhoz meghatározni a pontos szöveges átiratot, jellemzően változatlan nyelvi és akusztikus modell mellett, kulcsszókeresésben a bemondások halmaza rögzített, míg a kulcsszavak változnak a felhasználás során.

Mint a mesterséges intelligenciabeli alkalmazások általában, egy kulcsszófelismerő rendszer is hangolható annak érdekében, hogy minél inkább az elvárásoknak megfelelően működjön. Ennek során egy rögzített felvételhalmazon és rögzített kulcsszókészletet használva értékeljük ki egy konfiguráció teljesítményét valamilyen pontosságmértékkel, és ehhez hangoljuk az eljárás paramétereit. A kiértékeléshez azonban *annotált* hangfelvételekre van szükség: olyanokra, melyeknél előre meghatároztuk a kulcsszavak előfordulásainak pontos helyeit. Ez a feladat egyszerűnek tűnhet, amennyiben rendelkezésünkre áll a hanganyagok időzített szöveges átirata: ekkor azt te-

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

kintjük egy kulcsszó tényleges előfordulásának, ahol a keresett szó teljes egészében, önállóan megtalálható. Ez azonban sokszor leegyszerűsítés, teljes mértékben figyelmen kívül hagyja például az összetett szavakat. Különösen így van ez ragozó nyelvek (mint amilyen a magyar is) esetében, ekkor ugyanis a kulcsszó toldalékolt alakjait is valós előfordulásnak kell tekintenünk, melyek automatikus meghatározása nem triviális.

A felvetett problémára a legjobb megoldás az lenne, ha valamilyen emberi címkézést használnánk, azonban ennek nyilvánvaló hátránya a nagy munkaigény, emiatt nagyobb adatbázisok felcímkézése elég drága. Az is várható, hogy egyes felhasználók véleménye egy-egy konkrét esetben eltér, ugyanakkor valamiféle „objektív" listára lenne szükségünk. Kérdéses, hogy az egyes felhasználók visszajelzéseit összegezve kaphatunk-e egy széles támogatottságú listát.

Jelen cikkben a kulcsszó-relevancia problémát vizsgáltuk, elsősorban a fenti szempontokra koncentrálva. Összeállítottunk egy kérdőívet egy magyar nyelvű hangadatbázis [1] kétséges kulcszó-előfordulásairól, és felkértünk öt tesztalanyt, hogy ezek közül válasszák ki a szerintük relevánsnak tartott előfordulásokat; válaszaikat ezután több szempontból is elemeztük.

A cikk második fejezetében felvázoljuk a kulcsszókeresési feladatot és ismertetjük az abban széleskörűen alkalmazott pontossági metrikákat. A harmadik fejezetben részletesen leírjuk az alkalmazott automatikus kulcsszóelőfordulás-detektáló módszereket és a kérdőív összeállításának menetét. A negyedik fejezetben a felhasznált adatbázist és a kulcsszókereső rendszert ismertetjük; végül az ötödik fejezetben bemutatjuk az eredményül kapott pontossági értékeket, és részletesen elemezzük a különböző felhasználók válaszait.

2 A kulcsszókeresési feladat

A kulcsszókeresési feladatban felhasználók által beírt *kulcsszavak* előfordulásait keressük korábban rögzített (hang)*felvételek* egy halmazában. A kulcsszókereső rendszer előfordulás-*hipotézisek* listáját szolgáltatja, melyek mindegyike tartalmazza az előfordulás helyét (felvétel, kezdő és befejező időpontok), a kulcsszót és a hipotézis valószínűségét, mely szerint azok sorba rendezhetők. Más hasonló területekkel ellentétben a hipotézisek sorrendje nem lényeges, a valószínűség a hipotézisek szűrésére szolgál.

A hangfelvételek feldolgozása általában elég erőforrás-igényes, a felhasználó viszont joggal vár gyors választ, így tipikus a felvételek valamilyen mértékű előfeldolgozása; ez egy köztes reprezentációt eredményez, amelyben aztán a keresést végezzük. A több ismert reprezentáció közül jelen cikkünkben a legvalószínűbbnek talált fonémasorozatot használjuk, mely elég gyors keresést tesz lehetővé.

Cikkünk szempontjából persze a konkrét kulcsszókeresési algoritmus csak annyiban érdekes, hogy az általa visszaadott előfordulás-hipotézisek (melyek az egész cikkben változatlanok) hogyan illeszkednek a különféle módszerekkel meghatározott *releváns előfordulásokhoz*, és az utóbbiak hogyan befolyásolják a kulcsszókereső rendszer pontosságát. Ehhez azonban először definiálnunk kell a használt pontosságmetrikákat.

2.1 Az alkalmazott pontosságmértékek

A kulcsszókeresési probléma egy információ-visszakeresési feladat, emiatt hagyományos IR metrikákkal: pontossággal (*precision*) és fedéssel (*recall*) is mérhető egy adott algoritmuskonfiguráció teljesítménye [6]. A legtöbb információ-visszakeresési területen a két metrikát azok (parametrikus) harmonikus közepével, az F-mértékkel (*F-measure*) szokás egyetlen értékké aggregálni, azonban a kulcsszókeresés területén más metrikák terjedtek el. Leggyakrabban a Figure-of-Merit (FOM) mérőszámot használják, mely az óránként és kulcsszavanként 1, 2, … 10 hibás találat megengedése esetén elért fedési értékek számtani közepe. A másik elterjedt mérőszámot az amerikai National Institute of Standards and Technology (NIST) vezette be 2006-os kulcsszókeresési versenyén: ez az aktuális kulcsszó-súlyozott érték (*Actual Term-Weighted Value*, ATWV), mely a következőképpen definiált:

$$\mathrm{ATWV} = 1 - \frac{1}{T}\sum_{i=1}^{T}\left(P_{Miss}(t) + \beta P_{FA}(t)\right), \tag{1}$$

ahol $P_{Miss}(t)$ az adott kulcsszó eltévesztésének, $P_{FA}(t)$ pedig hibás találatának valószínűsége; azaz

$$P_{Miss}(t) = 1 - \frac{N_{corr}(t)}{N_{true}(t)} \quad \text{és} \quad P_{FA}(t) = 1 - \frac{N_{FA}(t)}{T_{speech} - N_{true}(t)}, \tag{2}$$

ahol $N_{corr}(t)$ az adott kulcsszó helyes találatainak, $N_{true}(t)$ a tényleges előfordulásainak, $N_{FA}(t)$ a hamis találatainak száma, T_{speech} pedig az átfésülendő felvételek összhossza másodpercben mérve [3]. β értéke általában 1000. Egy, a használt annotációval tökéletes összhangban működő rendszer ATWV pontszáma 1,0, egy olyané, amely egyáltalán nem ad vissza találatokat, 0,0. Feltételezve, hogy T_{speech} lényegesen nagyobb, mint $N_{true}(t)$, egy olyan rendszer, amely az összes elvárt előfordulást megtalálja, de minden kifejezésre óránként 3,6 hamis találatot produkál, szintén 0,0 értéket fog kapni, így ez a metrika jóval szigorúbb, mint a FOM. További különbség, hogy az ATWV az összes visszaadott találatot figyelembe veszi, míg FOM esetén csak a valószínűbbeket. Kísérleteink során mindkét metrikát alkalmaztuk.

3 A releváns előfordulások meghatározásának módjai

A következőkben azt ismertetjük, milyen stratégiákat alkalmaztunk, hogy meghatározzuk a kulcsszavak előfordulásainak helyeit a hangfelvételek szöveges átirata alapján.

3.1 Automatikus módszer

A legkézenfekvőbb megoldás (elvárjuk a kulcsszó előfordulását önálló szóként, ill. szósorozatként) a bevezetőben már említett okok (toldalékolás, összetett szavak) miatt nem alkalmazható, azonban annak egy módosított változata már igen: ekkor azt várjuk el, hogy a kulcsszó az átiratban teljes egészében bukkanjon fel egy szóban. Ezzel a ragozott szóalakokat is elfogadjuk. A magyar nyelv ragozási szabályait figyelembe véve a magánhangzóra végződő kulcsszavak esetében a hosszúra váltó magánhangzós változatot is elfogadtuk (pl. *Amerika – Amerikában*). Persze ez a megoldás sem tökéletes, különösen rövid kulcsszavakra jellemző, hogy sokszor fordulnak elő más szó belsejében, így sok téves riasztáshoz vezetve.

1. táblázat:
Relevánsnak minősített előfordulások száma a különböző alkalmazott módszerekkel a validációs és teszt adatbázisrészeken

Módszer	Validációs	Teszt
Automatikus	381	709
1. alany	365	690
2. alany	368	689
3. alany	396	732
4. alany	366	699
5. alany	367	697
Alanyok (többségi szavazás)	367	697
Egyértelmű	334	651

3.2 Emberi annotálás

A másik lehetőség, hogy akkor tekintünk egy előfordulást relevánsnak, amennyiben egy ember annak tekinti. Bár nyilván ez a legpontosabb módszer, hiszen pontosan akkor lesz relevánsnak minősítve egy előfordulás, amennyiben egy ember úgy gondolja, hogy az valóban releváns; nagyobb archívumok emberi annotálása azonban elég drága. Jelen cikkünkben viszont éppen arra voltunk kíváncsiak, hogy milyen változásokat okoz az emberi vélemények figyelembe vétele, és a felvétel-adatbázis sem volt túl nagynak mondható, így kísérleteinkben alkalmazhattuk ezt a megközelítést.

El szerettük volna kerülni, hogy az alanyok a többórányi hangfelvétel teljes leiratát annotálják az összes, a tesztjeinkben szereplő kulcsszóra, így automatikus módszerekkel leszűkítettük a lehetséges releváns előfordulások halmazát, és egy kérdőívre gyűjtöttük őket. Betűalapú illesztési távolságot használva megkerestük azokat a helyeket, ahol a kulcsszavakhoz hasonló betűsorozatok fordultak elő; legfeljebb az adott kulcsszó hosszának 30%-át kitevő betűbeszúrást, -törlést és -cserét engedtünk meg (tehát egy 10 betűvel leírható kulcsszó esetén legfeljebb három művelettel elő kellett tudni állítani azt). Mivel a lista még így is túl hosszú volt, azokat a potenciális előfor-

dulásokat automatikusan relevánsnak tekintettük, melyeknél *szó elején* és *teljes egészében* fordult elő az adott kulcsszó. (Ezekre az 5. fejezetben *egyértelmű* releváns előforduláskénti fogunk hivatkozni.) Mindezt abból a megfontolásból tettük, hogy ezek nagy eséllyel a kulcsszó ragozott alakjai, és habár ez nem minden esetben teljesült (pl. bizonyos összetett szavaknál), összességében elég jó közelítésnek találtuk, és hatékonyan csökkentette a kérdőív hosszát.

Végül ezt a kérdőívet tölttettük ki öt tesztalannyal; az általuk megjelölt előfordulások és az egyértelmű előfordulások halmazának unióját tekintettük szerintük releváns előfordulásoknak.

Az 1. táblázat mutatja a különböző automatikus módszerek és az egyes alanyok által relevánsnak minősített előfordulások számát. A kérdőív 111, illetve 242 potenciális előfordulást tartalmazott (a fejlesztési és a tesztelési halmazokra vonatkoztatva), melyek közül az alanyok 31-62-t, illetve 38-81-et választottak ki. A számok azt is tükrözik, hogy az alanyok (a 3. alany kivételével) alapvetően hasonlóan ítélték meg a potenciális releváns előfordulásokat (bár ehhez a kulcsszókeresési rendszer pontosságértékeit is érdemes megvizsgálni), és gyökeresen különböző módon, mint a két alkalmazott automatikus módszer. Mivel arra is kíváncsiak voltunk, hogy elérhető-e valamiféle konszenzus az alanyok között, a táblázatokban feltüntettük az egyszerű többségi szavazáshoz tartozó értékeket is.

4 Technikai megoldások

Mielőtt bemutatnánk és elemeznénk a teszteredményeket, még be kell mutatnunk, hogyan párosítjuk össze az előfordulás-hipotéziseket a releváns előfordulásokkal, valamint ismertetnünk az alkalmazott kulcsszókeresési rendszert és az adatbázist.

4.1 Az előfordulás-hipotézisek és a releváns előfordulások összepárosítása

Az irodalomban több megoldást is találunk a kulcsszóhipotézisek és -előfordulások összepárosítására. Természetesen a hipotézisnek és a tényleges előfordulásnak ugyanabban a felvételben kell lennie, és ugyanahhoz a kulcsszóhoz kell tartoznia. Mindezeken túl azt is elvárjuk, hogy a hipotézis ugyanabban az időpontban hangozzon el, mint a tényleges előfordulás, azonban ezen nyilvánvalóan nem érthetjük azt, hogy a kezdő- és végpontok is tökéletesen egybeessenek. Elvárhatjuk például, hogy ezek valamilyen határon belül legyenek; [3] esetében a tényleges előfordulás közepétől legfeljebb fél másodpercre kell esnie a hipotézisnek, míg [7] akkor párosítja össze a hipotézist egy tényleges előfordulással, amennyiben a hozzájuk tartozó időintervallumok metszik egymást. Mi az utolsó megoldást alkalmaztuk, részben tekintettel a magyar nyelv ragozó voltára, mely eléggé megnehezíti a szigorúan vett kulcsszó pontos kezdő és befejező időpontjainak meghatározását.

4.2 A kulcsszókeresési rendszer

Kísérleteinkben saját kulcsszókeresési keretrendszerünket használtuk (részletesen lásd [2]). Ebben a hangfelvételeken először beszédfelismerési lépéseket végzünk, jelen esetben egy nagypontosságú, kétmenetes neuronhálós fonémaosztályozási módszert alkalmazva [5]. Az eredményül kapott fonémasorozatot letároljuk, és erre illesztjük a beírt kulcsszó fonetikus átiratát. Az illeszkedés mértékét illesztési távolság (*edit distance*) metrikával mérjük, fonémánként eltérő műveleti költségeket használva, melyeket a fonémaosztályozó tévesztési mátrixából számítunk [4].

2. táblázat:
Kulcsszófelismerési pontosságok alakulása a különböző alkalmazott módszerek függvényében

Módszer	FOM	ATWV
Automatikus	88,72%	56,84%
1. alany	88,35%	52,32%
2. alany	87,39%	48,00%
3. alany	88,85%	60,23%
4. alany	88,15%	52,90%
5. alany	88,22%	53,05%
Alanyok (átlag)	88,19%	53,30%
Alanyok (medián)	88,22%	52,90%
Alanyok (többségi szavazás)	88,22%	53,07%
Egyértelmű	87,94%	44,77%

4.3 A felhasznált adatbázis

A kísérletekhez 70 híradót rögzítettünk nyolc tévécsatornáról (ATV, Hálózat TV, Hír TV, M1, M2, Rtl, Tv2) [1]. A felvételeket néhány mondatos blokkokra vágtuk; közülük jelen cikkben csak azokat használtuk, melyekben szépen artikulált beszéd hallható és a háttérzaj minimális. A 70 híradót 44-9-17 arányban osztottuk fel tanítási, fejlesztési és tesztelő blokkokra (időtartamot tekintve ez kb. 5 és fél óra – 1 óra – 2 óra), ügyelve arra, hogy a tévécsatornák mindegyikéből kerüljön mindegyik részhalmazba. A felvételek mindegyikét legépeltük, az ortografikus átiratot utólag is ellenőriztük. Az alkalmazott 50 kulcsszót a felvételekben gyakran előforduló főnevek közül választottuk ki; illeszkedve a felhasználói igényekhez, jelentős részük (18 darab) tulajdonnév volt. Hosszuk 6-16 fonéma, 2-6 szótag között alakult.

5 Eredmények

5.1 Kulcsszófelismerési pontosságok

A 2. táblázat tartalmazza az elért pontosságokat a különböző, a releváns kulcsszó-előfordulásokat detektáló módszerek esetén. Látható, hogy a FOM értékek gyakorlatilag változatlanok, míg az ATWV pontosságok elég nagy skálán (48,00%-tól 60,23%-ig) mozognak. A tesztalanyokhoz tartozó pontosságok nagyban eltérnek a két (*automatikus*, illetve *egyértelmű*) automatikus módszerhez tartozóktól is: ez alapján a felhasználói elvárásokhoz képest az egyik automatikus módszer jellemzően túl megengedő, a másik pedig túl szigorú. A többségi szavazással elért pontosságérték (53,07%) nagyon közel áll három tesztalanyéhoz (1., 4. és 5.), valamint az átlagos és a medián pontosságértékhez is. Ez azt jelzi, hogy egyszerű többségi szavazással valószínűleg elérhető egy, a gyakorlatban jól teljesítő konszenzusos előfordulás-lista.

5.2 A felhasználói válaszok elemzése

A pontosságértékek változásainál is érdekesebb kérdés, hogy az egyes előfordulásokat hogyan értékelték az egyes alanyok, és a vélemények mennyire hozhatók közös nevezőre. A következőkben ezeket a konkrét eseteket fogjuk körüljárni.

A csak korlátozott nyelvi információt hasznosító kulcsszókereső megközelítések ismert hátránya, hogy hajlamosak az (általában rövid) kulcsszavakat más szavak belsejében is „megtalálni", és így sok hamis riasztást generálni. Esetünkben ez a *kormány* kulcsszóval fordult elő jelentősebb számban, mely valóban megtalálható az *önkormányzat* szó belsejében, így ezeket az előfordulásokat az automatikus keresőmódszer is relevánsnak minősítette; ugyanakkor az öt alanyból négy vélte úgy, hogy ezek hamis riasztások. Kulcsszókereső rendszerünk, mely csak az akusztikus információra támaszkodhatott, természetesen szintén megtalálta ezeket az „előfordulásokat".

Az automatikus módszerben megengedtük, hogy a kulcsszó szóvégi magánhangzója hosszúra váltson (a többi magánhangzó viszont nem). A *vasút* kulcsszó esetében hasonló dolog történt, csak ellenkező előjellel: mindegyik alanyunk úgy vélte, hogy a *vasutas* szó is a *vasút* kulcsszó releváns előfordulása. Ugyanakkor, habár hangtanilag tökéletesen ugyanez az eset a *miniszter* kulcsszó és a *minisztérium* szó, a megkérdezett alanyok közül mégis mindössze egy sorolta ezt a releváns előfordulások közé.

További nagy csoport volt a kulcsszavak között bizonyos személyneveké: *Angela Merkel* (német kancellár), *Bajnai Gordon* vagy *Orbán Viktor* (magyar miniszterelnökök). Kulcsszóként a teljes név volt megadva, időnként azonban a felvételekben csak vezetékneveikkel hivatkoztak rájuk. Az összes alany egyetértett azzal, hogy ezek is releváns előfordulások, bár csak a keresett kulcsszavak fele fordult elő. Megjegyzendő, hogy mivel illesztési távolságot használva állítottuk össze a kérdőívet, azon csupán azok az előfordulások szerepelhettek, ahol a szövegkörnyezet a hiányzó keresztnévhez igen hasonló volt (pl. „amely Merkel", „Bajnai-kormány", „Orbán-kormány").

Ehhez igen hasonló eset volt a *rendőrség* kulcsszóé: többször is szerepelt a kérdőíven a *rendőr* szó, melyet az ötből három alany tartott releváns előfordulásnak annak ellenére, hogy itt a kulcsszó tartalmazta a ténylegesen előforduló szót. Ez feltehetőleg azt tükrözi, hogy ezen alanyok számára a két fogalom szorosan összekapcsolódik. Hasonló viszonyt jelez a *gázár* kulcsszó esete is: a többször is szereplő „gáz ára" szókapcsolatot ugyanis az összes alany a kulcsszó releváns előfordulásának tekintette.

A fenti példák esetében az alanyok általában egyetértettek egymással, a válaszokat azonban nemigen lehetne automatikusan megjósolni. Ha egy elhangzott szó teljes egészében tartalmazza a keresett kulcsszót, az általában releváns előfordulás; bizonyos esetekben (*kormány*) ugyanakkor nem az, máskor pedig a kulcsszó tartalmazza a ténylegesen elhangzott szót (*rendőrség*). A kulcsszó szóvégi magánhangzója hosszúra válthat, és ez időnként más magánhangzókkal is előfordulhat (*vasút*), más esetekben viszont nem (*miniszter*). A *gázár* kulcsszó esete valószínűleg egyáltalán nem kezelhető automatikusan: amennyiben kulcsszavakon belül akárhol engedélyezünk szóhatárokat, az rengeteg hamis riasztáshoz vezethet. Viszont ha ismertebb személyeket keresünk, célszerű a kulcsszót csak a vezetéknévnek választani (*Merkel*, *Bajnai*, *Orbán*).

Amikor a megkérdezett alanyok egy-egy hipotézis besorolásakor nem értettek egyet, szinte mindig négy az egyhez aránylottak a szavazatok; összesen négy helyen alakult ez három a kettőhöz. Ez azt sugallja, hogy szinte minden esetben elérhető egy elfogadott konszenzus, azaz létrehozható olyan címkézés, mely szinte teljesen egybeesik az emberek által elvárt viselkedéssel. (Ezt természetesen érdemes lenne ötnél lényegesen több alanyra is megvizsgálni.) Ezt kulcsszókeresési rendszerünk pontosságmértékei is alátámasztották: amennyiben szavazásnál azt vártuk el, hogy legalább négy alany értsen egyet az adott előfordulás megítélésében, a pontosságértékek alig változtak, egyhangú eredmény elvárása esetén viszont számottevően csökkentek.

Az emberi annotálással elért pontszámokat az automatikus módszerekéihez hasonlítva egyértelmű, hogy alapvetően különböznek: mikor csak a tiszta előfordulásokat tekintettük relevánsnak, az ATWV értéke 44,77% lett, mely a többi előforduló pontosságértékhez mérten alacsony (valószínűleg a sok hamis riasztás miatt); mikor viszont a standard automatikus módszert alkalmaztuk, az túl megengedőnek bizonyult, amely az előálló, irreálisan magas 56,84%-os ATWV értékben is tükröződik.

6 Konklúzió

Jelen cikkünkben szokatlan nézőpontból vizsgáltuk meg a kulcszókeresési problémát: azt elemeztük, hogy az automatikusan előállított kulcsszó-előfordulások mennyire egyeznek a felhasználói igényekkel. Ehhez tesztalanyokat kértünk meg, hogy jelöljék meg, mely potenciális előfordulásokat tekintik valóban relevánsnak. A válaszokat elemezve azt találtuk, hogy, habár nem volt két pontosan ugyanúgy válaszoló alany, összességében a válaszok egymáshoz nagyon hasonlónak bizonyultak, és egyszerű többségi szavazással egyértelmű konszenzus volt elérhető. A kipróbált automatikus eljárások azonban vagy túl optimisták, vagy túl pesszimisták voltak, és a tesztalanyok válaszait részletesen megvizsgálva azt sem tartjuk valószínűnek, hogy automatikus (szintaktikai) eljárásokkal azok reprodukálhatóak lennének.

Hivatkozások

1. Gosztolya, G., Tóth, L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal, Proc. MSZNY (2010) 224–235
2. Gosztolya, G., Tóth, L.: Spoken Term Detection Based ont he Most Probable Phoneme Sequence, Proc. SAMI (2011) 101–106
3. NIST: The Spoken Term Detection (STD) Evaluation Plan, National Institute of Standards and Technology (NIST), Gaithersburg, USA, http://www.nist.org/speech/tests/std (2006)
4. Szöke, I., Schwarz, P., Matejka, P., Karafiát, M.: Comparison of Keyword Spotting Approaches for Informal Continuous Speech, Proc. Interspeech (2005)
5. Tóth, L.: A Hierarchical, Context-Dependent Neural Network Architecture for Improved Phone Recognition, Proc. ICASSP (2011) 5040–5043
6. Wang, D.: Out-of-Vocabulary Spoken Term Detection, PhD thesis, Univ. Edinburgh (2010)
7. Young, S.J. et al: The HMM Toolkit (HTK) (software and manual), http://htk.eng.cam.ac.uk/ (1995)

A kondicionálisok problémája jogszabályszövegekben

Markovich Réka[1], Hamp Gábor[2], Syi[2]

[1] ELTE Filozófiatudományi Doktori Iskola, Logika Tanszék
1088 Budapest, Múzeum körút 4/I,
markovich@phil.elte.hu

[2] BME Szociológia és Kommunikáció Tanszék
1111 Budapest, Egry József u. 1.
hampg@eik.bme.hu,i@syi.hu

Kivonat: A jogszabályszövegek gépi feldolgozása során érdemes figyelembe venni a jogszabályok nyelvtechnológiai szempontból egyedi vonásait (a jogszabályok deontikus jellegét és kontextusérzéketlenségét, a szövegen belüli definíciók és a listás felsorolások gyakori jelenlétét), és ezek ismeretében kell felkészíteni a nyelvtanokat a gépi elemzés számára. A jogszabályokban kiemelt jelentősége van a kondicionalitás jelenségének, mert a jogszabályok alkalmazásának szükséges és elégséges feltételeit lehet velük kifejezni. Jelen tanulmányunkban ezért a kondicionális és retrokondicionális mondatok felismerésére alkalmas nyelvtant mutatjuk be.

Kutatásunk során a jogszabályszövegek gépi elemzésének lehetőségeit vizsgáljuk. Ehhez olyan tudáskomponenseket kezdtünk el építeni, amelyek egy jól működő mondatelemzővel együtt képesek lehetnek a gépi elemzések támogatására. Bár érdemi eredményeket csak hatékony mondatelemző birtokában lehet várni, úgy véljük, hogy amíg ez nem áll rendelkezésre, addig is lehet értelme az előkészítő munkáknak. Ezért olyan komponenseket definiáltunk, amelyek – jogalkotási, logikai, valamint nyelvtechnológiai tudásokat egybegyűjtve – a későbbiek során hasznosíthatóak lehetnek.

A természetes nyelvi mondatok gépi elemzése, az ehhez szükséges formális szemantika kidolgozása során számos nehézséggel szembesülhetünk. Elsőként kell kiemelnünk azt a problémát, hogy a logikai formalizálás következtében jópár – általában a pragmatika keretében kezelt – tartalmi összefüggés nem ragadható meg igazán. A természetesnyelv-használati praxisok körében valószínűleg a jogszabályszövegek világa az, ahol a legkevesebb teret szeretnénk hagyni a pragmatikai jelenségeknek. Nagyon furcsállanánk, ha egy bírósági döntésben arra hivatkoznának, hogy egy adott kitétel ugyan nem szerepelt a jogszabályban, de a kontextus „sugallta” azt (vagyis a társalgási implikatúra keretében jelen volt). A jog világában nincs irónia, nincsenek metaforák. Csak szószerinti jelentés van, aminek egyértelmű jele, hogy a legfontosabb fogalmak jelentését a jogszabályok értelmező rendelkezéseiben rögzítik (a jogszabály-szerkesztéstől szóló IRM rendelet [3] 2.§-a ki is mondja a jogszabályok *ellentmondásmentességére* vonatkozó elvárást). Ezért szokás úgy hivatkozni a jogszabályszövegekre, hogy azok szárazak. Viszont épp ezért gondolhatjuk, hogy ha valamely természetesnyelv-használati közeg alkalmas a gépi elemzésre, akkor a jogszabályok szövege az.

A jogszabályszövegek sajátosságai

A jogszabályszövegeknek van néhány érdekes vonása, amiket nyelvtechnológiai szempontból feltétlenül figyelembe kell vennünk. Egyrészt a szabatosságra törekvés nagyon gyakran eredményez *többszörösen összetett mondatot*, és ez a tény pontos és hatékony szintaktikai elemző nélkül megnehezíti az adekvát tagmondat-beazonosítást. Másrészt tipikus a paragrafusokon, illetve bekezdéseken belüli pontokba rendezés, mely gyakran lista típusú felsorolásokat takar. Ezek a felületes szemlélő számára tördelési megfontolások eredményének tűnhetnek, valójában strukturális rendezések, amelyeknek kihatása lehet a tagmondatok közti logikai kapcsolatokra is. A szövegszerű összekapcsolás a mondatokat *lineárisan* köti össze a *konkatenáció* műveletével, ami annyit jelent, hogy a mondatok egymáshoz fűzése az olvasás irányával egyezik meg (mondhatjuk azt, hogy *vízszintes*). Ezzel szemben a listás összekapcsolás esetében jelentkezik egy másik viszonyítási dimenzió: amikor a szöveg olvasási irányához képest ortogonálisnak (*függőlegesnek*) tekinthető irányban is kapcsolat van a tagmondatok között. Ezt a jelenséget a torontói kommunikációelméleti iskola tárta fel először. Az írásbeliség sajátosságait keresve az iskola egyik képviselője, Jack Goody [4.] az írott szöveg jellegzetességeként említette a grafonyelvi technikákat, amelyek olyan lehetőségeket kínálnak az írás befogadója számára, amire a szóbeliség nem (nem is lehet) képes. Ilyen grafonyelvi technika a lista, de ide tartozik a táblázat, a keresztrejtvény vagy a mátrix is (ezek azonban számunkra most kevéssé érdekesek). Előfordulhat, hogy a jogszabályokban megjelenő listák csak tördelési szempontból tűnnek másnak a „normál" mondatokhoz képest, bár a listaelemeket jelző (és egymástól elválasztó) *felsorolásjelek* (betűk, számok, grafikai jelek, zárójelek) szövegbe ágyazása már önmagában is megtöri a mondatszerkesztés jólformáltságát. Sok más esetben azonban akkor sem kapunk jólformált mondatokat, ha eltekintünk a felsorolásjelektől. A jogszabályokon belüli listák gyakori felbukkanása azért fontos, mert felismerésük egyfelől újszerű feladatot jelent a mondatelemzés számára, másfelől fel kell készülni arra is, hogy a listák összetevőit ugyanúgy többféle módon kapcsolhatjuk össze egymással, ahogy ez az „egyszerű", szövegszerűen összetett mondatoknál is megfigyelhető.

A jogszabályok logikai elemzésének további sajátossága abból fakad, hogy a norma nem leír, hanem előír. Az előírásoknak azonban nem állítások az eszközei, márpedig a klasszikus logika elvileg csak állításokon működik: az állítások igazságátörökítő jellege adja egy levezetés logikai érvényességét, igazsága pedig csak állításoknak van, előírásoknak nincs. Ezt a sajátosságot hivatott kezelni a deontikus logika. Ám ha a jogszabályban szereplő mondatok felszínét nézzük, akkor kijelentő mondatokat, vagyis állításokat találunk, amelyek nulladrendben deontikus logikai eszközök nélkül is elemezhetőek.

A logikai kapcsolók

A jogszabályszövegek mondatépítkezése gyakorta komplex szerkezeteket hoz létre. Az összetett mondatokat logikailag úgy ragadhatjuk meg, hogy egyszerű mondatokat

kapcsolunk össze logikai konnektívumok (logikai kapcsolók) segítségével. Elvileg tizenhat bináris logikai kapcsoló lehetséges. Ebben nincs benne a logikai tagadás, mivel az nem bináris művelet. A logikai tankönyvekben ritkán mutatják be mind a tizenhat bináris logikai műveletet, ezért érdemes itt felsorolni mindegyiküket. A műveleteket igazságtábla segítségével jellemezhetjük. Ha két propozíciót (*A*-t és *B*-t) összekapcsolunk a logikai konnektívumainkkal, akkor az összekapcsolások igazságértékei az alábbi módon alakulnak.

1. táblázat

	A	i	i	h	h
	B	i	h	i	h
1	ellentmondás	h	h	h	h
2	konnegáció	h	h	h	i
3	retroszubtrakció	h	h	i	h
4	kontraprojekció	h	h	i	i
5	szubtrakció	h	i	h	h
6	kontra-retroprojekció	h	i	h	i
7	biszubtrakció	h	i	i	h
8	exklúzió	h	i	i	i
9	konjunkció	i	h	h	h
10	bikondicionális	i	h	h	i
11	retroprojekció	i	h	i	h
12	kondicionális	i	h	i	i
13	projekció	i	i	h	h
14	retrokondicionális	i	i	h	i
15	diszjunkció	i	i	i	h
16	tautológia	i	i	i	i

Ezek közül néhányat sosem használunk a természetes nyelvben, így a fenti tizenhat kapcsoló közül „kivehetjük” az *ellentmondást* (1) és a *tautológiát* (16), valamint a négyféle *projekció* műveletét (4, 6, 11, 13) mint gyakorlatilag érdektelen eseteket. Marad tíz lehetséges kapcsoló, ám ezek közel sem azonos gyakorisággal fordulnak elő a tényleges nyelvhasználatban. Nem véletlen, hogy a logikai nyelvek praxisában, oktatásában öt-hat műveletről beszélnek a maradék tíz helyett. A leggyakrabban használt kapcsolók a *kondicionális* (12), a *diszjunkció* (15), a *konjunkció* (9) és a *bikondicionális* (10), a többi művelet – ezekhez képest – ritkábban bukkan fel. Műszaki szövegekben fontos szerepe van az olyan ritkábban használt konnektívumoknak, mint a *szubtrakció* (5) és *retroszubtrakció* (3), a *biszubtrakció* (7), a *konnegáció* (2) és *exklúzió* (8). Ritkán előforduló jelenségnek szokás tartani, ám épp a jogi szövegekben fontos szerepet kap a *retrokondicionális* művelete (14).

Kutatásunkban első körben a logikai konnektívumok közül a kondicionális megjelenéseit, beazonosíthatóságát vizsgáljuk. A ‘ha-akkor’-os szerkezetként emlegetett konnektívum a leggyakoribb következtetési eljárás, a modus ponens-szel való kapcsolata miatt kiemelt jelentőséggel bír a logikában. Ráadásul vizsgálata akár az elsőkörös intuícióinkkal szembemenő eredményt is hozhat (sok logikával ismerkedő számára nehezen elfogadható, hogy a kondicionális előtagja hamissága esetén is igaz), így érdemi kimenet remélhető a logikai elemzéstől. Az intuícióellenességgel

kapcsolatban érdemes megjegyeznünk, hogy mégha vitatható is, hogy a kondicionálisnak megfelel-e a természetes nyelvi 'ha-akkor' szerkezet, abban azért egyet szokás érteni, hogy ez utóbbit logikailag a kondicionálissal ragadhatjuk meg leginkább. Kutatásunk azonban épp arról szól, hogy mit mondhatunk erről a megfeleltetésről. Igaz-e egyrészről, hogy a jogszabályokban a 'ha-akkor'-os szerkezetek logikailag kondicionálist rejtenek, másrészről találni-e olyan kondicionálist, amelynek nem ez a felszíni szerkezete, lehet-e a kondicionális más kötőszóval összekapcsolt tagmondatok együttese, esetleg kifejezhetjük-e egyszerű mondattal is. Szintén kérdés, hogy a kondicionalitás, ami magyarul feltételességet jelent, hogyan viszonyul a tagmondatok sorrendjéhez. A logikában ismert tétel [5], hogy a klasszikus kondicionálissal elégséges, míg a retrokondicionálissal (vagy konverz kondicionálissal) szükséges feltételt állapítunk meg. Ha ezt elfogadjuk, a nyelvi vizsgálat során azt kell megvizsgálnunk, hogy ez a kötőszó-megfordítás a nyelvi szerkezet szintjén is általános-e: valóban minden 'akkor-ha'-s szerkezet retrokondicionálist rejt-e logikailag, illetve milyen más nyelvi megjelenési módozatai vannak a retrokondicionálisnak.

A fenti kérdések megválaszolása után következik az a feladat, ami az egész projektet nyelvtechnológiailag relevánssá teszi: milyen input adható (adandó) ahhoz, hogy az ember által megtalált szabályosságok gép által felismerhetővé (s ezáltal később automatikusan elemezhetővé) váljanak, s mindez milyen biztonsággal adható meg. Ez természetesen függ a rendelkezésre álló nyelvelemző szoftverek (esetünkben a NooJ) lehetőségeitől és korlátaitól, és függ a jogszabály-szerkesztési gyakorlat szabályosságától, következetességétől. A kutatás eredménye egyúttal rámutathat olyan speciális sajátosságokra, amelyek a vizsgált szöveg műfajából fakadnak (kimutatva ezzel esetleg bizonyos strukturális összefüggéseket a jogszabályszöveg mint műfaj és a logikai konnektívumok között). A kondicionális és a retrokondicionális gép általi megkülönböztethetősége pedig gyakorlati haszonnal kecsegtető eredmény: a jogalkalmazás során hasznos ismeret, hogy az adott jogszabály bizonyos tényállásokhoz, értékelésekhez elégséges vagy szükséges feltételeket állapít meg.

Az elemző nyelvtan felépítéséhez a gazdasági reklámtörvényt (Grtv.) [2] használtuk tanulókorpusznak. A törvény valamennyi mondatát annotáltuk, és azokban kerestük a kondicionálisok és retrokondicionálisok nyelvi megjelenési módozatait. A következőkben bemutatott példáink ezért mind a Grtv-ből származnak.

A kondicionalitás mintázatai

A kondicionális mondatok természetesen a 'ha-akkor' szerkezet mellett másféle felszíni formában is kifejezhetőek. Nyilvánvaló példaként hivatkozhatunk az 'amennyiben…, akkor…', 'mikor…, akkor…' nyelvi szerkezetekre. A felismerés szempontjából fontos tény, hogy a jogszabályokban többször hagyják el az 'akkor' tagot, mint ahányszor megjelenítik, ami azt is jelenti egyben, hogy a kapcsoló 'ha' tagja van kulcsszerepben. Kevésbé kézenfekvő az a tény, hogy egyszerű mondattal is kifejezhető logikai kondicionális. Erre az 'esetén' névutó használata a legjobb példa:

> „6. § (4) … Megtiltás esetén az érintett személy részére reklám közvetlen üzletszerzés útján a továbbiakban nem küldhető." (Értsd: 'Ha megtiltották, akkor nem küldhető reklám.')

Szintén kevésbé várt megoldás az, amikor nem egy, hanem két egymást követő (önmagában is összetett) mondat fed le egy logikai kondicionálist. A „kötőszó" a második mondat elején szereplő 'ebben az esetben':

> „6. § (3) Az (1) bekezdés szerinti hozzájáruló nyilatkozat bármikor korlátozás és indokolás nélkül, ingyenesen visszavonható. Ebben az esetben a nyilatkozó nevét és minden egyéb személyes adatát az (5) bekezdésben meghatározott nyilvántartásból haladéktalanul törölni kell, és részére reklám az (1) bekezdésben meghatározott módon a továbbiakban nem közölhető." (Értsd: 'Ha visszavonták a nyilatkozatot, az adatokat törölni kell és reklám a továbbiakban nem küldhető.')

A retrokondicionális az 'akkor-ha' alapeseten kívül szintén előfordulhat 'akkor' tag nélkül is. Ebben az esetben az segít a beazonosításban, hogy a 'ha' elem a második tagmondatban van. Jócskán megnehezíti a gépi felismerés számára adott instrukciót a többszörösen összetett mondat:

> „19.§ (3) Nem minősül dohánytermék reklámjának az olyan áru reklámozása, amelynek elnevezése, megjelölése vagy árujelzője valamely dohánytermékével megegyezik, ha az áru elnevezése, megjelölése vagy árujelzője egyértelműen elkülöníthető a dohánytermékétől."

A megadandó nyelvtanban számolnunk kell a retrokondicionális kontraponáltjának előfordulásával is:

> „5.§ (2) … Ilyen nyilatkozat hiányában a reklám nem tehető közzé" (Értsd: 'Ha nincs nyilatkozat, nem tehető közzé a reklám, vagyis abból, hogy jogszerűen közzétették a reklámot, következtethetünk arra, hogy előtte nyilatkoztak.')

Komoly nehézséget jelentenek – különösen a gépi feldolgozás szempontjából – azok az esetek, ahol beágyazott kondicionálisokat találunk. Előfordul azonban, hogy a kötőszó árulkodó annyira, hogy nagy bizonyossággal megadhatjuk a teljes beágyazott szerkezetet. Az alábbi mondatban a 'kivéve ha' kapcsoló felfedi azt a külső kondicionálist is, amelynek egyébként nincs felszíni jele:

> „(3) A (2) bekezdéstől eltérően, a kereskedelmi kommunikáció megjelenítési módjával összefüggő okból eredő jogsértésért az is felel, aki a kereskedelmi kommunikációt az arra alkalmas eszközök segítségével megismerhetővé teszi, valamint aki önálló gazdasági tevékenysége körében a kereskedelmi kommunikációt megalkotja vagy ezzel összefüggésben egyéb szolgáltatást nyújt, kivéve, ha a jogsértés az (1) bekezdés szerinti vállalkozás utasításának végrehajtásából ered." (Értsd: 'Ha a megjelenítési módból adódóan jogsértő a reklám, akkor csak akkor nem felel a közzétevő, ha a reklámozó utasította.' Ennek a logikai szerkezete tehát: A→(~B→C).)

Érdekes és nehézséget jelent, hogy a retrokondicionális nyelvi megjelenítésének prototípusa, az 'akkor-ha' az 'is'-sel kiegészülve megfordítja a logikai kapcsolatot, hiszen az derül ki belőle, hogy az utótagban megjelölt feltétel nem szükséges, hanem elégséges. Így épphogy nem a retrokondicionális, hanem a kondicionális kötőszavaként kell számolnunk az 'akkor is, ha' kötőszóval:

> „9.§ (2) Az (1) bekezdés szerinti vállalkozás felel akkor is, ha a kereskedelmi gyakorlatot szerződés alapján más személy valósítja meg a vállalkozás érdekében vagy javára"

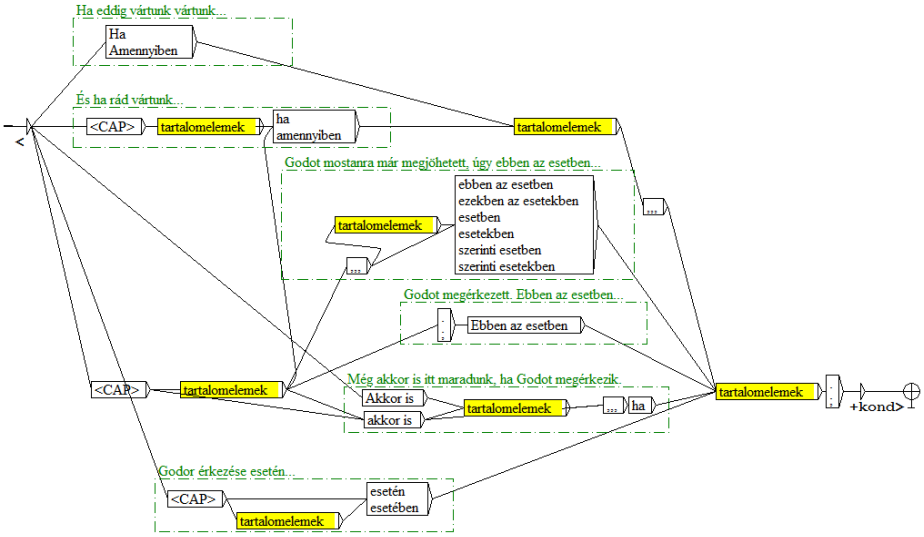
A kondicionálison és a retrokondicionálison túl a bikondicionális is előfordulhat a jogszabályszövegekben. Az ezt jellemző 'akkor és csak akkor' szerkezettel ugyan nem találkoztunk, olyan esetekkel viszont igen, amikor egy-egy kondicionális vagy retrokondicionális formáját öltő mondat valójában bikondicionálist takart. Csakhogy ezt óvatosan kell kezelnünk, mert a bikondicionalitás felismerése a lexikalitáson túlmutató háttértudást feltételez, jellemzően a jog fogalmából, eszközeinek struktúrájából tudható. Példa erre a „Ha a külön törvény eltérően nem rendelkezik, az ilyen

szabályok megsértésére e törvény rendelkezéseit megfelelően alkalmazni kell" mondat. Ennek olvasásakor nem derül ki számunkra, hogy a feltétel visszafelé is áll: abból, hogy az adott passzust alkalmazni kell, tudjuk, hogy arról a kérdésről nem rendelkezett eltérően külön törvény. Ez egy kondicionálisnál nem lenne igaz. Ez a tudás azonban nem a jogszabály szövegéből nyerhető ki, hanem a jogalkotási mechanizmus ismeretéből. Így ezeket a mondatokat nem jelöltük külön, a felszíni szerkezet szerint soroltuk be őket, s ennek megfelelően a bikondicionális kategóriájával ebben a cikkben nem foglalkozunk.

A nyelvtanok és az eredmények

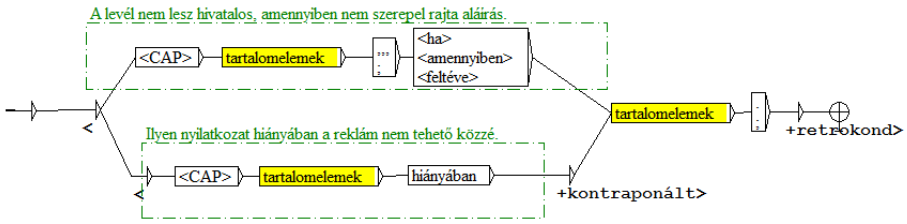
Az elemzéshez a NooJ nyelvelemző környezetet használtuk [6], aminek segítségével számos magyar nyelvű elemzés született, amelyekkel az MSzNy korábbi konferenciakötetteiben találkozhattunk. A mi célunk az volt, hogy a kondicionális típusú (kondicionális, retrokondicionális) konnektívumokat tartalmazó mondatokat megtaláljuk, illetve azonosítsuk bennük a logikai kapcsolatban álló tagmondatokat. Az elemzéshez a Grtv. szövegét úgy készítettük elő, hogy a NooJ-keresés szövegegységei mondatok legyenek, tehát a jogszabályi szövegek bizonyos elemeit (hatályosságra vonatkozó megjegyzéseket, korábbi változatoktól való eltérésekre utalások jegyzeteit), valamint a jogszabályok fontos részét képező tagolóelemeket (fejezetelés, jogszabályi belső azonosítókat stb.) elhagytuk. További része volt a szövegelőkészítésnek a felsorolások feloldása. A feloldás kifejezést azért lehet használni, mert a listák egyfajta rövidítésként foghatók fel, amelyek részben a könnyebb áttekinthetőséget, részben a tipográfiai redundanciák elkerülését szolgálják. A feloldás azonban nem magától értetődő, mivel a listák időnként más funkciójúak. A feloldások tipográfiai következménnyekkel jártak; például a felsorolások szerkezetéből fakadóan a listában szereplő elemeket, amelyek a lista elején álló egység (tagmondat vagy éppen mondatok) kiegészítései, a listaszerű felsorolás vesszővel, illetve más kapcsoló elemekkel ('továbbá', 'valamint' stb.) választja el. Ezeket az átkötő elemeket a szövegelőkészítésben mondatzáró központozással helyettesítettük. A helyettesítéseknél törekedtünk arra, hogy a kodifikációs szokások tipográfiai transzformálása ne eredményezzen jelentésmódosítást. Mivel a listák viszonylag tipikus és jól formált elemek a jogszabályi szövegekben, ezért formálisan is jól leírhatók, a feloldásuk automatizálható. Találtunk azonban olyan esetet, amely módosítást idézett elő a logikai szerkezetben: egy látszólag retrokondicionális szerkezettel rendelkező felsorolást a felsorolás ténye éppen visszahelyezett a kondicionális státuszba, mivel az egyes megjelölt feltételek nem lehetnek szükségesek, ha a felsorolás bármely elemének önálló teljesülése elegendő az első tagmondatbeli minősítéshez. A listafeloldás azonban külön mondatokat csinál belőlük, így egyenként mind retrokondicionálisként értékelhető.

A keresések ideális kivitelezése az lenne, hogy az előkészített szöveget szintaktikai elemző segítségével tagmondatokra, és azokon belül frázisokra bontjuk, majd az így felcímkézett szövegben a szövegegységeken belül megkeressük a logikai kapcsoló elemeket (a hatókörük kijelölésével együtt). A mondattani elemző modul híján azonban csak arra volt módunk, hogy felszíni illesztéseket kereső nyelvtanokat építsünk. Ezek gráfjait mutatjuk be az 1. és 2. ábrán.



1. ábra: kondicionálist tartalmazó mondatokat azonosító nyelvtan

Az 1. ábrán látható nyelvtan a kondicionálisokat tartalmazó mondatok azonosítását és felcímkézését végzi el. A tanulókorpuszban talált minták alapján a kondicionálisok négy alaptípusát különítettük el. A ‘ha… akkor…’ szerkezetek adták az első típust. Ennél az előtag tartományát kijelölő ‘ha’ – vagy szinonimája – kötelező elem, az ‘akkor’ tag opcionális. A második típus a jogszabályi szövegekben gyakran szereplő visszautaló fordulat, amely a kondicionális utótagját jelöli ki (‘ebben az esetben’, ‘a … bekezdésben felsorolt esetekben…’ stb.). Ez a szerkezet időnként több mondatot is tartalmazhat. A harmadik típus az előbbinek névutós szerkezettel kifejezett nominalizációja. Ebben az esetben – a korábban már bemutatott példán – a kondicionális elő- és utótagja nem két tagmondatban jelenik meg. A negyedik típus pedig az intuíciónkkal szembefeszülő szerkezet: ‘…akkor is…, ha…’. A nyelvtanunk a fenti szerkezeteket tartalmazó mondatokat a ‘<+kond>’ címkével látja el. Az ábrán szereplő nyelvtan tartalmaz egy beépülő gráfot (a ‘tartalomelemek’ nevűt), amely a reguláris NooJ-kifejezésekkel szóelőfordulásokat, számokat és központozásokat keres (<WF>, <D>, <P>) a nyelvtanban feltüntetett logikai elemek és konkrét kifejezések tartományában, illetve környezetében. Mivel a tartalomelemek nevű gráf elemezetlenül megtalálja ezeket, a dátumok, jogszabályi hivatkozások külön azonosítására nincsen szükség; azonban világos, hogy mind az előtag- és utótagtartományok, mind a dátum és egyéb referenciák azonosítására szükség lesz egy olyan elemzésben, ahol nem pusztán a kondicionális egységek megjelölése a cél, hanem a találatok tartalomelemeinek a kezelése is. A gráfban láthatók három vesszőt tartalmazó csomópontok. A szövegelőkészítés folyamán a keresés egyszerűsítése érdekében a tagmondathatárokat jelölő központozási jeleket (rendszerint vesszőket) három vesszőre cseréltük megkülönböztetve ezeket a felsorolásokban szereplő központozási jelektől (vesszőktől).



2. ábra: Retrokondicionálist tartalmazó mondatokat azonosító nyelvtan

A retrokondicionális nyelvtan az előbbi gráfhoz hasonlóan használja a tartalomelemek azonosításához szükséges beépülő nyelvtant, és két típust különít el. Az egyik az 'akkor-ha' szerkezetet azonosítja, amelyben a második tagmondat hatókörét kijelölő kötőszó, illetve az elhagyható 'akkor' a fontos keresőelem. A második típus a 'hiányában' névutós frázis, amely kontraponált retrokondicionálist jelöl.

A nyelvtanokat a fogyasztókkal szembeni tisztességtelen kereskedelmi gyakorlat tilalmáról (Fttv.) [1] szövegén teszteltük. Az előkészített szöveg 120 mondatból állt, az annotálás számait, valamint a két nyelvtannal lefuttatott elemzés eredményeit mutatja az alábbi táblázat.

2. táblázat

		kondicionális nyelvtan			**retrokondicionális nyelvtan**		
	esetek száma	jól felismerte	nem ismerte fel	rosszul ismerte fel	jól felismerte	nem ismerte fel	rosszul ismerte fel
kondic.	14	14					
retrokond.	13			1	13	1	
egyéb	93						

*

A tanulmány a 83887. sz. OTKA kutatás keretén belül készült. A cikk szerzői megköszönik Váradi Tamásnak a Nooj kezeléséhez nyújtott segítségét.

Hivatkozások

1. 2008. évi XLVII. törvény a fogyasztókkal szembeni tisztességtelen kereskedelmi gyakorlat tilalmáról (Fttv)
2. 2008. évi XLVIII. törvény a gazdasági reklámtevékenység alapvető feltételeiről és egyes korlátairól (Grtv)
3. 61/2009. (XII. 14.) IRM rendelet a jogszabályszerkesztésről
4. Goody, J.. Nyelv és írás. In Nyíri K., Szécsi G. (szerk.): Szóbeliség és írásbeliség. Budapest: Áron Kiadó, (1998) 189–221
5. Pólos L., Ruzsa I., Madarász Zs. A logika elemei. Budapest: Osiris (2005)
6. Silberztein, M.. NooJ Manual. (www.nooj4nlp.net) (2003)

A Humor új Fo(r)mája

Novák Attila

MTA–PPKE Nyelvtechnológiai Kutatócsoport
Pázmány Péter Katolikus Egyetem Információtechnológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
novak.attila@itk.ppke.hu

Kivonat: A MorphoLogic Humor morfológiai elemzőjéhez az utóbbi évtizedekben számos nyelven készült morfológiai adatbázis. Ezek közül némelyik igen jó lefedettséget és pontosságot ad, mások olyan nyelvekre biztosítják az automatikus morfológiai elemzés lehetőségét, amelyekre más hasonló erőforrás nem létezik. A Humor elemzőszoftver zárt licence azonban nem tette lehetővé ezeknek a nyelvi erőforrásoknak a szabad terjesztését. Ugyanakkor a Humor elemző implementációja nem teszi lehetővé az ismeretlen szavak elemzését (morphological guessing), valamint azt sem, hogy az egyes szavakhoz gyakorisági információt rendeljünk, vagy a modellt másképp súlyozzuk. Ezeket a problémákat úgy oldottuk meg, hogy a Humor morfológiai erőforrásait olyan véges állapotú leírássá konvertáltuk, amely mindezeket a problémákat megoldja és rendelkezik nyílt forráskódú implementációval is.

1 Bevezetés

A MorhpoLogic Humor elemzője [7] számára számos nyelvhez készült jó minőségű morfológiai adatbázis. Ezek között számos agglutináló nyelv szerepel: a magyar [5] mellett az következő kis finnugor nyelvek: a komi, az udmurt, a mezei mari, az északi manysi és néhány hanti dialektus [6]. Ezek mellett különböző indoeurópai nyelvekhez, a lengyelhez, az angolhoz, a némethez, a franciához és a spanyolhoz is készült Humor leírás. Ezeknek a morfológiáknak a többsége egy az elemző által használt formátumnál magasabb szintű redundanciamentes jegyalapú leírás használatával készült, amelyből a Humor adatbázis automatikusan jön létre [5, 6].

Az eredeti Humor elemzőalgoritmus nem alkalmas arra, hogy a szó végződése alapján olyan szavak lehetséges elemzéseit előállítsa, amelyeknek a töve nem szerepel az adatbázisában. Nem is lenne egyszerű az algoritmust úgy módosítani, hogy képes legyen ennek a feladatnak a megoldására. Egy ilyen ismeretlenszó-elemző integrálása az elemzőbe ugyanakkor igen hasznos eszköz lenne, hiszen minden szöveg sok olyan szóalakot tartalmaz, amelynek a töve nem szerepel az elemző szótárában.

Emellett nem lehetséges a morfológiai modellek súlyozása vagy gyakorisági információval való ellátása sem, amelyre szükség lenne ahhoz, hogy a morfológia közvetlenül alkalmas legyen adatvezérelt szövegnormalizálási feladatok (pl. automatikus helyesírás-javítás vagy beszédfelismerés) támogatására. Szintén hasznos lenne a modellek súlyozhatósága az ismeretlenszó-elemző által generált javaslatok sorrendezé-

séhez. Ezek mellett a Humor hátrányaként a morfológiaielemző-szoftver zárt licence említhető, amely nem teszi lehetővé ezeknek a nyelvi erőforrásoknak a szélesebb körben való terjesztését.

Ebben a cikkben bemutatjuk, hogy hogyan oldottuk meg a fenti problémákat a Humor formátumú morfológiai leírások forrásának véges állapotú leírássá alakításával, amelyek kompilálására és használatára nyílt forráskódú eszközök is rendelkezésre állnak. A véges állapotú reprezentáció használható végződésalapú ismeretlenszó-elemzésre, természetes megoldást kínál gyakorisági információ hozzáadására a modellhez, és lehetővé teszi a súlyozott hibamodellekkel való kompozíciót.

2 A Humor morfológiai elemző

A program 'item-and-arrangement' típusú elemzést hajt végre: egy szóalak lehetséges elemzéseit morfsorozatokként adja meg. A szót felépítő minden morfnak kiírja a felszíni és mögöttes alakját, valamint a kategóriáját (amely strukturált információt is tartalmazhat, de lehet belső szerkezet nélküli címke is).

Az elemző mélységi keresést végez az adott szóalakon a lehetséges elemzések után. Olyan morfokat keres a szótárában, amelyeknek a felszíni alakja illeszkedik a megadott szó még elemezetlen részére. A lexikon nemcsak morfokat, hanem morfsorozatokat is tartalmazhat, amelyeket az elemző így egy lépésben ismer fel.

Elemzés közben a program kétféle ellenőrzést hajt végre. Egyrészt lokális kompatibilitás-ellenőrzést végez az egymás mellett álló morfok között, másrészt azt is ellenőrzi, hogy az elemzést alkotó morfémák a nyelv lehetséges szókonstrukciói egyikét testesítik-e meg. Az utóbbi ellenőrzést a szónyelvtant leíró kiterjesztett véges állapotú automata bejárásával ellenőrzi.

A Humor elemző lexikai adatbázisa a morfémák allomorfjainak leírásából, a szónyelvtant leíró véges állapotú automatából és a szomszédos morfémák lokális kompatibilitás-ellenőrzéséhez használt kétféle adatszerkezetből áll. Ezek egyikét folytatási osztályok és bináris kompatibilitási mátrixok alkotják, amelyek az egymással összekapcsolható folyatatási osztályokat adják meg. A másik adatszerkezetet bináris tulajdonságvektorok és megszorításvektorok alkotják. Minden morf leírása tartalmaz egy jobb és egy bal oldali folytatási osztálycímkét, egy jobb oldali bináris tulajdonságvektort és egy bal oldali bináris megszorításvektort. Az utóbbi tartalmazhat olyan pozíciókat, amelyek nem számítanak az illeszkedés szempontjából.

A lokális kompatibilitás-ellenőrzés a következő módon történik: egy adott morf (tipikusan egy toldalék) akkor illeszkedik az előző morfhoz (tipikusan tőhöz), ha a tő jobb oldali tulajdonságai kielégítik a toldalék bal oldali megszorításait mind a bináris tulajdonságok, mind a releváns folytatási mátrix szempontjából.

A szószerkezet globális ellenőrzéséhez használt szónyelvtan-automata bináris kiegészítő állapotváltozókat is tartalmazhat a fő állapotváltozója mellet, amelyek segítségével az automata méretének robbanása nélkül írhatók le a nem szomszédos morfémák közötti megszorítások.

Mindezek mellett a morfológiai adatbázis tartalmaz egy olyan leírást is, amely egy a jobb oldali tulajdonságvektorok halmazáról a morfológiai kategóriacímkék halmazára történő leképezést definiál. Ezeket a címkéket használjuk a szónyelvtan-

automata éleinek címkéiként. Az adott morf fellapozását és a lokális kompatibilitás-ellenőrzést minden esetben egy a szónyelvtan-automatában végzett lépés is követ. Az adott lépés akkor lehetséges, ha az automata adott állapotában (beleértve a kiterjesztett állapotváltozók aktuális értékét is) van olyan kimenő él, amely az adott morf jobb oldali tulajdonságvektora által meghatározott morfológiaicímke-halmaz valamelyik elemével van címkézve, és nem tartozik egyéb olyan megszorítás az adott élhez, amely a kiterjesztett állapotváltozók aktuális értékével nem kompatibilis.

Az adatbázis nehezen lenne karbantartható közvetlenül abban a formában, amelyben az elemző használja, mert ez az adatbázis-reprezentáció redundáns, alacsony szintű és nehezen olvasható formátumú adatszerkezeteket tartalmaz. Ezen problémák megoldására szolgál az a nyelviadatbázis-leíró keretrendszer, amelynek segítségével az adatbázis magas szintű és redundanciamentes formában írható le, amelyet a keretrendszer automatikusan alakít át az elemző által használt formára. A nyelviadatbázis-leíró keretrendszer létrehozás után keletkezett morfológiai leírások már ennek a magasabb szintű formalizmusnak a használatával készültek.

A magas szintű leírás leképezéséhez a rendszer egy kódolási leírást használ, amely minden egyes elemi tulajdonsághoz, amely a magas szintű leírásban szerepel, megadja, hogy az milyen alacsony szintű adatszerkezetre képződjön le és hogyan. Egyes tulajdonságok a bináris tulajdonságvektorokra képeződnek le, a többi pedig együtt határozza meg a folytatási mátrixokat, amelyeket dinamikusan generál a rendszer.

3 Véges állapotú morfológiák

A legszélesebb körben használt véges állapotú morfológiai eszközkészlet a Xerox *xfst-lookup* párosa [2]. Az *xfst* compilerrel különböző formalizmusok alkalmazásával lehet számítógépes morfológiákat leíró véges állapotú transzducereket létrehozni, amelyek morfológiai elemzőként vagy generátorként való működtetésére a *lookup* program szolgál. A morfológiai leírások a Xerox formalizmusában egyrészt a morfémákat leíró lexikális adatbázisból, másrészt a morfofonológiát leíró szabályrendszerből állnak. A lexikon definiálására szolgál a *lexc* formalizmus, amelynek segítségével leírhatók és allexikonokba szervezhetők a morfémák, és a szónyelvtan folytatási osztályok segítségével adható meg. Egy *lexc* allexikon általában olyan absztrakt morfémaleírásokból áll, amely a lemma és a morfoszintaktikai címkék mellett a morféma általában fonológiailag absztrakt ábrázolását tartalmazza. Az utóbbinak a szövegekben ténylegesen előforduló felszíni alakokra vetítését egy fonológiai-helyesírási szabályrendszer végzi.

A fonológiai szabályok szekvenciális és párhuzamos szabályrendszerként is megfogalmazhatók. Az *xfst* formalizmus és compiler segítségével szekvenciális újraírószabály-rendszerek adhatók meg és alakíthatók véges állapotú transzducerekké, illetve komponálhatók egymással és a lexikont leíró, a *lexc* formátumú leírásból kompilált transzducerrel. Így egyetlen lexikális transzducer hozható létre amely közvetlenül leképezi a felszíni szóalakokat a lemmákból és morfoszintaktikai címkékből álló lexikai reprezentációkra. Egy hasonló compiler, a *twolc* használható a párhuzamos kétszintű megszorítások segítségével megadott morfológiai leírások kompilálására.

A Xerox véges állapotú eszközkészlete a Humor elemző szónyelvtan-automatájában használt kiterjesztett állapotváltozókhoz hasonló formalizmus segítségével ugyancsak lehetővé teszi az állapottér faktorizációját. Az erre szolgáló konstrukciót a Xerox terminológiájában 'flag diacritics'-nek hívják.

Bár a flag diacritics használata általában csökkenti az elemző sebességét, ez a konstrukció mégis nagyon hasznos lehet, mert használatával megelőzhető, hogy a transzducer mérete exponenciálisan felrobbanjon a morfológiában szereplő nem lokális megszorítások következtében. Emellett arra is használható, hogy akár a morfémák közötti lokális megszorításokat is a puszta folytatási osztályoknál kifejezőbb és olvashatóbb formában írjuk le. Az xfst tartalmaz egy olyan műveletet, amelynek segítségével az ilyen lokális megszorításokat megfogalmazó flagek az automataméret számottevő növekedése nélkül eliminálhatók, növelve ezzel az elemző sebességét.

A Xerox eszközkészlete igen erős formalizmust ad a bonyolult morfológiai szerkezetek leírására. Ezért nem voltak komoly kétségeink azzal kapcsolatban, hogy a Humor formalizmus felhasználásával implementált morfológiai leírások átalakíthatóak lesznek a véges állapotú leírásokká.

Ugyanakkor, bár a Xerox eszközeit kutatási célra hozzáférhetővé tették 2003-ban Beesley és Karttunen könyvének [2] publikálásakor, ezek két szempontból nem különböznek lényegesen a Humor elemzőtől: zárt forráskódúak és nem használhatóak súlyozott modellek létrehozására. Ugyanakkor az ismeretlenszó-elemzés problémájának megoldására alkalmasak. Szerencsére néhány évvel ezelőtt létrejöttek az *xfst* és a *lookup* nyílt forráskódú alternatívái. Ezen nyílt forráskódú eszközök egyike, a Foma [3] alkalmas az *xfst-lexc* formátumú morfológiai leírások kompilálására és működtetésére. Ez tehát lehetővé teszi a zárt forráskód okozta problémák kiküszöbölését. Ezen kívül a szintén nyílt forráskódú HFST-eszközkészlet [4] segítségével a Foma formátumú transzducerek OpenFST [1] formátumúakká konvertálhatók, amely implementáció viszont lehetővé teszi a súlyozott véges állapotú modellek létrehozását.

4 A Humor–lexc konverzió

Mivel a Humor formalizmusban leírt morfológiai modellek a morfológia teljes leírását tartalmazzák a morfofonológiával együtt, ezen leírások átalakításához nincs szükség sem szekvenciális *(xfst)* sem párhuzamos *(twolc)* szabályrendszer használatára. Az átalakításhoz kizárólag a *lexc* formalizmust használjuk.

Minden morf lexikai alakját és kategóriacímkéjét a morf *lexc* reprezentációjának lexikai, a felszíni alakját pedig a felszíni oldalára képezzük le. Az utóbbi a valódi felszíni alak, nem egy olyan absztrakt fonológiai reprezentáció, amit általában a *lexc* lexikonforrásokban használni szoktak. A felszíni és a lexikai alakban egymásnak megfelelő szimbólumok helyes egymáshoz rendeléséről a lexikonkonverter implementációja gondoskodik. A címkéket egyetlen többkarakteres szimbólumként ábrázoljuk.

A Humor leírásban folytatási osztályok, mátrixok, illetve bináris tulajdonság- és megszorításvektorok formájában megadott lokális morfszomszédossági megszorításokat közvetlenül *lexc* folytatási osztályokként ábrázoljuk. A leképezés egyszerű implementálásához a Humor lexikonokat generáló programot kiegészítettük egy

olyan kapcsolóval, amelynek megadása esetén a program olyan mátrixokat generál, amelyek önmagukban teljesen leírják a szomszédossági megszorításokat, így a vektorok a *lexc* lexikonokra való leképezés folyamán figyelmen kívül hagyhatóak. Minden morf lexc reprezentációjának előállításakor az allexikont, amelybe az adott morf kerül, a morf bal mátrixának neve és a bal oldali folytatásiosztály-kódja határozza meg. A lexc folytatási osztályát ugyanakkor a jobb oldali mátrixnév, folytatásiosztály-kód és a szónyelvtan-kategóriacímke együttesen határozza meg.

A Humor szónyelvtan legkönnyebben a flag diacritics formalizmus segítségével képezhető le a véges állapotú formalizmusra. A Humor automata fő állapotváltozóját egy flagre képezzük le, amelyet St-nek neveztünk el. Ugyanakkor a kiterjesztett állapotváltozók mindegyike egy-egy újabb flagre képeződik le. Hogy pontosan milyen flag diacritics élek kapcsolódnak egy-egy morfhoz, azt az adott morf szónyelvtan-kategóriája határozza meg.

A jobb oldali mátrixnév és folytatási kód alapján a morfémalexikonok bal oldalához a Humor mátrixokban leírt kompatibilitási viszonyokat közvetlenül leíró allexikonokon keresztül csatoljuk vissza a Humor szónyelvtan-automatát leíró flag élek jobb oldalát. Az St szónyelvtan-állapotflag eliminálható a leírásból, de ez az állapottér jelentős megnövekedésével járhat.

5 Eredmények

Az alábbiakban röviden összehasonítjuk magyar morfológiánk egy 144000 morfot tartalmazó változatának eredeti Humorral kompilált változatának és az átalakított xfst-vel kompilált változatnak futásimemória-igényét és elemzési sebességét. A véges állapotú lexikonból két változat eredményeit mutatjuk be. Az első változatból nem eliminátuk az St flaget, a másodikból igen.

1. táblázat: Egy 144000 morfból álló magyar morfológiai leírás Humor és xfst által kompilált változatának összehasonlítása

	Humor lexikon	lexc – St flaggel	lexc – St flag eliminálva
futási memória	3,3 MB	20,6 MB	38,5 MB
elemzési sebesség	4700 szó/s	12500 szó/s	33333 szó/s

A véges állapotúvá alakítás jelentősen növeli az elemző memóriaigényét (>11-szeresére), ugyanakkor jelentős elemzésisebesség-növekedéssel is jár (>7-szeressel). Az St flag eliminálása majdnem kétszeresére növeli a lexikon méretét, ugyanakkor igen jelentős sebességnövekedéssel is jár. A további flagek eliminálása nagyjából szintén kétszeresére növeli a véges állapotú lexikon méretét minden egyes eliminált flaggel. Ez emellett rendkívül hosszú kompilálási időhöz is vezet. Ugyanakkor gyakorlatilag semmilyen további pozitív hatással nincs az elemzési sebességre.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP-4.2.1./B-11/2-KMR-2011-0002 és a TÁMOP-4.2.2./B-10/1-2010-0014 pályázatok támogatásával készült.

Hivatkozások

1. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., Mohri, M.: Openfst: a general and efficient weighted finite-state transducer library. In Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA 2007) 11–23
2. Beesley, K. R., Karttunen, L.: Finite State Morphology. CSLI Publications, Ventura Hall (2003)
3. Huldén M.: Foma: a finite-state compiler and library. In: Proceedings of EACL (2009) 29–32
4. Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A., Silfverberg, M.: HFST - Framework for Compiling and Applying Morphologies. In Proc. SFCM (2011) 67–85
5. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szegedi Tudományegyetem (2003) 138–145
6. Novák, A.: Language Resources for Uralic Minority Languages. Proceedings of the SALTMIL Workshop at LREC 2008, Marrakech (2008) 27–32
7. Prószéky, G., Kis, B.: A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: Proceedings of the 37th Annual Meeting of the ACL, College Park, Maryland, USA (1999) 261–268

Tudásalapú ajánlórendszer adatszegény környezetben

Oravecz Csaba, Sárközy Csongor, Mittelholcz Iván

MTA Nyelvtudományi Intézet
e-mail:{oravecz.csaba,sarkozy.csongor,mittelholcz.ivan}@nytud.mta.hu

Kivonat Az ajánlórendszerek általában a felhasználói tranzakciókból és a termékekről rendelkezésre álló adatokból kinyert információkra támaszkodnak. Adatszegény környezetben azonban más információforrások felhasználására van szükség. A tanulmány olyan megoldás prototípusát mutatja be, ahol a felhasználó tevékenységét jellemző szöveges adatok automatikus feldolgozása és egy részletes ontológiában tárolt tudásbázis felhasználása segítségével válik lehetővé a releváns termékek (szolgáltatások) kiválasztása.[1]

Kulcsszavak: ajánlórendszer, tudásbázis, ontológia

1. Bevezetés

Az online felhasználók számára az igényeiknek megfelelő termékek és szolgáltatások keresése, azonosítása és beszerzése a kérdéses termék, szolgáltatás komplexitásának függvényében komoly kihívást jelentő feladat lehet, melyben a felhasználó megfelelő támogatása kritikus fontosságú. A legegyszerűbb keresőalkalmazások általában azzal a feltételezéssel működnek, hogy a felhasználó pontosan tisztában van az általa keresett termék releváns paramétereivel és kimerítően ismeri az elérhető termékek halmazát is. Ez azonban ritkán vagy így, ezért az adott döntési folyamatban helye van annak az (automatikus) asszisztenciának, amely képes a felhasználót hatékonyan segíteni.

Azokat az internetes alkalmazásokat, melyek a felhasználók érdeklődésére számot tartó termékek és szolgáltatások felderítésében és kiválasztásában nyújtanak automatikus segítséget, ajánlórendszereknek (*recommender systems*) [6] nevezzük. Esetünkben olyan ökoinnovációs intézkedések, szolgáltatások személyre szabott kiajánlását végzi egy automatikus rendszer[2], melyek alkalmazásával a felhasználók (vállalkozások) jelentős megtakarítást érhetnek el. Az alkalmazás felépítésében és funkciójában sok hasonlóságot mutat az ajánlórendszerek klasszikus típusaival, ezen belül is a tudás- és megszorításalapú rendszerekkel [1,7], de az általunk fejlesztett rendszer működési tartományának és környezetének egyúttal számos olyan paramétere van, melyek egyedileg kidolgozott megoldásokat követelnek meg, túlmutatva a klasszikus ajánlórendszerekben felhasznált módszerek és algoritmusok szolgai alkalmazásán.

[1] A kutatást a KMR_12-1-2012-0036 számú, *Piacorientált kutatás-fejlesztési tevékenység támogatása a közép-magyarországi régióban* pályázat támogatta.

[2] A továbbiakban ECOINNO rendszerként hivatkozott alkalmazás.

2. Kihívások az ECOINNO rendszerben

Alapvetően három szempont szerint érdemes megvizsgálni azokat a tulajdonságokat, melyek lényeges eltéréseket jelentenek a megszokott ajánlási paradigmához képest.

- Feladat és kontextus: Mind a kínálati, mind a keresleti oldalon speciális paramétereket kell figyelembe venni. A termékek igen komplex szolgáltatások, melyek releváns tulajdonságainak meghatározása és reprezentációja nem triviális, ennek a feladatnak automatikus, gépi módszerekkel történő megoldása nagy kihívást jelentő probléma. Másrészt a kínálati halmaz számossága nem akkora, hogy a humán szakértői beavatkozást eleve ki kellene zárni, vagyis a terméktulajdonságok (automatikusan segített) manuális annotációja reális alternatíva. Ez a megközelítés a későbbi esetleges kiterjesztés során is fenntartható, hiszen a kérdéses szolgáltatások várható jövőbeli bővülése messze nem olyan ütemű, ami a kézi feldolgozást lehetetlenné tenné.

  A keresleti oldalon megjelenő felhasználók többsége először kerül kapcsolatba a rendszerrel, illetve a rendszer által ajánlott terméktípussal. Egyrészt tehát gyakorlatilag minimális mértékben tudja explicit módon megfogalmazni a valós igényeit, másrészt kezdetben semmilyen információ nem áll rendelkezésre arra vonatkozóan, hogy korábban milyen hasonló termékeket vett igénybe. Ezen túl, a rendszer használata során sem várható olyan mennyiségű ilyen jellegű adat felhalmozódása adott felhasználóval kapcsolatban, melyre a további ajánlatokat alapozni lehetne, így ez a fajta információ csak minimális mértékben vehető figyelembe.

  Fontos szempont, hogy a rendszer alkalmazási területe jól meghatározott, zártnak tekinthető, ezért az ezzel kapcsolatos háttértudás egyértelmű meghatározása és formális rögzítése természetesen kínálkozó lehetőség.
- Információforrás: Mindkét oldalon igen változatosak a nyers, közvetlenül elérhető információ formai és tartalmi jellegzetességei.
  - Termékoldalról: A szolgáltatások strukturálatlan vagy félig strukturált[3] leírásai, melyek a legritkább esetben készültek azzal a céllal, hogy automatikus számítógépes módszerekkel feldolgozhatók, értelmezhetők legyenek.
  - Felhasználói oldalról: Minden olyan információforrás, mely a felhasználó tevékenységére, környezetére vonatkozóan tartalmaz adatot, és a rendszer számára a felhasználó minimális közreműködésével hozzáférhető, releváns lehet (weboldal URL, prospektusok, ismertetők, beszámolók, jelentések stb.). Ez a fajta információ alapvetően strukturálatlan szöveges

[3] Strukturált információ a rendszer szempontjából olyan formátumú adat, mely explicit, géppel értelmezhető formális reprezentációba közvetlenül, nyelvi, logikai feldolgozás nélkül leképezhető.

formában jelenik meg[4], és független a felhasználó és az ajánlórendszer közötti interakciótól, nem abból származik.
– Kívánt eredmény: A felhasználói igényeknek megfelelő szolgáltatások rangsorolt listája jelenik meg a rendszer kimeneteként[5].

3. A megvalósítás általános elvei

A rendszer működése során alapvetően egy információ-visszakeresési (*information retrieval (IR)*) problémát old meg, ahol a klasszikus alkotóelemek, mint a *dokumentumgyűjtemény* és a *keresési kifejezés* speciális formában jelennek meg. Ezért a megfelelő módosításokkal a IR-paradigma, illetve meghatározott sztenderd algoritmusok alkalmazása nyitva áll a feladat megoldása során. Ha kínálati oldalon elérhető termékek (mint „dokumentumok") és a keresleti oldalon megjelenő felhasználók igényei (mint „keresőkifejezés") alkalmas módon meghatározhatók és reprezentálhatók, a feladat ezen reprezentációk közötti hasonlóság kiszámítására redukálódik (mint klasszikus IR-probléma).

Két feladatot kell tehát megoldani. Egyrészt a felhasználói profil és a terméktulajdonságok olyan reprezentációját meghatározni, mely lehetővé teszi ezen reprezentációk között egy hasonlóságmérték definiálását és kiszámítását, másrészt a rendelkezésre álló információforrások adatait (és esetleg további információforrásokat) felhasználva mindkét oldalon előállítani ezeket a reprezentációkat. Nyilvánvaló, hogy az előzőekben említett zárt és korlátos domén lehetőséget ad arra, hogy a releváns háttérismeret, fogalmi rendszer és összefüggések egy formális explicit leírásban (ontológiában) megadhatóak legyenek [2,8], ennek a tudásbázisnak a felhasználása kritikus fontosságú.

4. Reprezentáció

A *kínálati oldalon* a reprezentációk előállítása a szolgáltatások deskriptív jellemzését tartalmazó szöveges leírások számítógépes nyelvészeti elemzését, felcímkézését, majd ezen címkézés manuálisan segített, a tudásbázis (ontológia) által definiált fogalmi térbe történő leképezését foglalja magában. A *keresleti oldalon* ugyanez történik azzal a különbséggel, hogy a folyamat teljesen automatikus, forrásként a felhasználóról elérhető minden lehetséges deskriptív adatot felhasznál, illetve támaszkodhat a felhasználótól irányított formában bekért további adatokra (preferenciákra).

A rendszer az alkalmazási doménre vonatkozó háttértudást, releváns objektumokat, kategóriákat, fogalmakat és relációkat egy explicit formális tudásbázisban (ontológiában) tárolja. Az ontológia az OWL Web Ontology Language

[4] Természetesen a szöveg főbb alkotóelemeire (cím, fejléc, bekezdés stb) tagolt és ebben az értelemben strukturált lehet, de ez a rendszer számára csak segédinformáció, mely további feldolgozásra vár annak érdekében, hogy a kivont információ a megelőző lábjegyzetben leírt értelemben is strukturált legyen.

[5] Ebből a szempontból az ECOINNO rendszer nem különbözik a sztenderd ajánlórendszerektől.

formalizmusban implementált, építésének munkakörnyezete a Protégé ontológiaszerkesztőre és a hozzá kapcsolódó Java API-ra épül[6].

Tegyük fel, hogy a rendszer alkalmazási tartományában lehetséges a releváns háttértudást „kimerítően" leírni.[7] Ebben az esetben mind a kínálati oldal intézkedései, mind a keresleti oldal felhasználói profiljai megfogalmazhatók, leírhatók a tudásbázis segítségével, mint olyan fogalomhalmazok, melynek elemei az ontológia adott csomópontjaihoz tartozó fogalmak. Alapvetően tehát mindkét oldalon a kérdéses reprezentáció egy többdimenziós fogalomvektor, ahol a vektor koordinátái az ontológia által specifikált fogalmak, értékei pedig kétfélék lehetnek: bináris vektor esetén 0 vagy 1, valós vektor esetén pedig az adott fogalom relevanciáját reprezentáló súlyérték. Bináris vektorok előállítása rendkívül egyszerű, amennyiben az adott fogalom hozzárendelhető az intézkedés, illetve felhasználói profilhoz, az érték 1, egyébként 0. Valós vektor esetén meg kell határozni azt a módszert, melynek segítségével az értékek kiszámíthatók.

4.1. Az intézkedésprofil előállítása

Mivel a kiajánlható szolgáltatások halmazának számossága nem kirívóan nagy, elvileg a teljesen manuális annotáció sem kivitelezhetetlen. Célszerű azonban az annotációs folyamatot automatikus módszerekkel segíteni. Ekkor a szolgáltatásokról rendelkezésre álló szöveges információt egy nyelvi elemzőlánc dolgozza fel és annotálja egy előre definiált címkehalmazból (amely az ontológia alacsony szintű specifikus fogalmainak feleltethető meg) választott címkékkel. A humán annotátor ezek után ellenőrzi és javítja a hozzárendelést, illetve a kezdeti változatban súlyokat rendel a megfelelő címkékhez. Az így kialakított reprezentáció a 4.3. részben ismertetett módon kerül további feldolgozásra.

4.2. A felhasználói profil előállítása

A feladat a felhasználói oldalon elérhető jórészt strukturálatlan adatokból az ajánlórendszer számára releváns információ kinyerése és leképezése a tudásbázis által specifikált vektortérbe. Ez több lépésben történik.

– Forrás-előfeldolgozás. Első lépés a szöveges adatok típusától függően (pl. HTML-dokumentum, PDF-ismertető, Word-dokumentum stb.) a dokumentumstruktúra elemeinek azonosítása (keletkezési idő, cím, fejléc, bekezdés stb.), mely lehetővé teszi az információ pontos lokalizálását (ezáltal pl. súlyozását) az adott forráson belül.
– Információazonosítás. Ebben a lépésben a szöveges adatok nyelvi, szemantikai elemzésére kerül sor, ahol részletes annotációt kapnak a tartalmas nyelvi elemek és szerkezetek, megtörténik a kulcskifejezések és a köztük lévő viszonyok meghatározása.

[6] `http://protege.stanford.edu/overview/protege-owl.html`

[7] Nyilván ezt közvetlenül mérni nem lehetséges, a tudásbázis minőségére gyakorlati szempontból a rendszer működési hatékonyságából, pontosságából lehet következtetni.

– Leképezés. Jelen specifikáció szerint a doménontológia alsó szintű fogalmaihoz vannak hozzárendelve azok a nyelvi elemek és relációk, melyek a kérdéses fogalmakat a szöveges adatokban instanciálják. Alapvetően tehát a tudásbázis definiálja a felhasználói adatok annotációjából az ontológiai fogalmakba történő leképezést. Mind a fogalmi csomópontok, mind a releváns nyelvi elemek azonosításában nagy szerepet játszanak azok a lexikális erőforrások, melyeket a rendelkezésre álló szöveges adatokból sztenderd statisztikai eljárások segítségével készültek (lásd 1. és 2. ábra).

```
1993.55969884557        1356    hulladék
1553.50432823931        1087    környezeti
1529.70377199569        1104    megtakarítás
1385.73629885459        1115    intézkedés
1068.88348017766        786     tonna
1065.25227892306        824     környezetvédelmi
895.661275426298        730     kft
709.212863548727        639     beruházás
675.285411574036        608     termék
...
```

1. ábra. Felhasználói dokumentumokból előállított kulcsszólista.

```
nyers<>fűrészpor<>10 12.2717 3.1616 10 14 34
kompakt<>fénycső<>15 11.7500 3.4631 12 20 41
szabadlevegős<>hűtés<>16 11.7395 3.4631 12 14 59
szerves<>oldószer<>32 10.7642 3.8708 15 29 70
mart<>aszfalt<>15 12.7571 3.1618 10 10 34
hulladékhő<>hasznosítás<>44 9.9866 3.1592 10 16 145
épület<>fűtés<>73 8.5454 3.4548 12 108 70
maradékanyag<>mennyiség<>74 8.5424 3.3077 11 16 434
...
```

2. ábra. Felhasználói dokumentumokból előállított kollokációs lista.

A felhasználói profilvektorhoz súlyokat hozzárendelni legegyszerűbben instanciagyakoriság alapján lehet:

$$w_{i,j} = \frac{freq_{i,j}}{\max_k freq_{k,j}} \tag{1}$$

ahol $w_{i,j}$ a j felhasználóprofilban instanciálódott i fogalomhoz tartozó súlyérték, $freq_{i,j}$ i előfordulási gyakorisága j-ben, $\max_k freq_{k,j}$ pedig a leggyakoribb fogalomhoz tartozó gyakorisági érték. A reprezentációkat az 1. táblázat illusztrálja.

1. táblázat. Profilreprezentációk.

	...	C_n(szennyvíz)	C_{n+1}(talaj)	C_{n+2}(zaj)	...
intézkedés$_i$	...	0.023	0.001	0	...
intézkedés$_j$	...	0	0.001	0.145	...
intézkedés$_k$	...	0	0.326	0.002	...
			...		
vállalkozás$_i$	...	0.001	0.001	0	...
vállalkozás$_j$	...	0.423	0.003	0	...
vállalkozás$_k$	...	0.003	0.377	0.005	...

4.3. A fogalmi vektorok kiterjesztése

A fogalmak közötti viszonyokat explicit módon specifikáló doménontológia lehetőséget ad arra, hogy a közvetlenül instanciált fogalmak mellett a velük meghatározott módon kapcsolatban álló további csomópontok (fogalmak, fogalmi osztályok) is hozzáadódjanak a reprezentációs vektorhoz. Ez a kiterjesztés például az ún. megszorított terjedésaktiváció (*constrained spreading activation*) alkalmazásával valósítható meg [3,4], melynek során különböző megszorítások által korlátozott módon az egyes csomópontokhoz további kapcsolódó csomópontok rendelhetők hozzá, ily módon az eredeti fogalomvektor kibővül a kapcsolódó fogalmakkal.[8]

A profilok választott vektortér alapú reprezentációja lehetővé teszi, hogy sztenderd hasonlósági mértékek segítségével természetes módon rangsorolhatók legyenek a felhasználókra szabott ajánlatok. A jelenleg alkalmazott mérték a koszinusz hasonlóság.

5. Javító stratégiák

A rendszer kézenfekvő kiterjesztését adják olyan szabályalapú megszorítások, melyek mind a felhasználói profil nyelvi szemantikai elemzéséből származó annotáció, mind az intézkedések tulajdonságai, mind a doménontológia szintjén megfogalmazhatók, és bizonyos kapcsolatokat, következményeket egyértelműen definiálnak[9]. További hasonló megszorítások származtathatók az audit kérdőívekre adott felhasználói válaszokból. Ezek egyértelmű és kiterjedt specifikálása a rendszer szempontjából kritikus fontosságú, mivel nagy mértékben leszűkíthetik az illesztési probléma keresési terét, javítva az ajánlati válaszlistát.

Nincs olyan mesterséges intelligenciára támaszkodó alkalmazás, amely kimerítően képes lenne kezelni egy adott tárgykört, tartományt. Előfordulhat, hogy

[8] A kapcsolt fogalmakhoz rendelt súly számítható pl. az eredeti súlyból lépésenként konstans érték levonásával (*decay*).

[9] Pl. adott tulajdonsággal rendelkező, vagyis adott dimenzióban nem 0 értékű vektorral jellemzett intézkedés nem járhat együtt egy másik meghatározott módon specifikált intézkedéssel.

a rendszer tudásbázisa hiányos, és nem lehet megbízható profilt, reprezentációt előállítani a felhasználóról a rendelkezésre álló adatok alapján. Ilyenkor lehetőség van arra, hogy a rendszer alacsonyabb szintű, kulcsszóalapú illesztést végezzen, illetve a tudásalapú és a kulcsszóalapú illesztést kombinálja. Ennek a megoldásnak a pontos paraméterei a részletes tesztelés során határozhatók meg.

6. Kiértékelési módszerek

Az ajánlórendszerek kiértékelésére nincs egységes, minden feladatban megbízhatóan alkalmazható módszertan [5]. Mind a tesztadatok kiválasztására, mind a felhasználói visszajelzésekből származó információ felhasználására számos megoldás lehetséges, ahol a konkrét alkalmazás teljesítményét legpontosabban mérő eljárás kidolgozása nem triviális. A projekt jelenlegi szakaszában egy mintegy 300 intézkedést tartalmazó tesztadatbázisra és 10-15 felhasználó bináris szelekciót tartalmazó válaszaira támaszkodik a kiértékelés[10]. Ebben a kontextusban a sztenderd fedés, pontosság értékek értelmezhetők, a tesztelési folyamat keresztvalidációval elvégezhető. Ez a megközelítés azonban meglehetősen durva modelljét adja a felhasználói elégedettségnek, ezért nem tekinthető végleges megoldásnak.

7. Összefoglalás és további feladatok

A tanulmányban bemutattuk egy olyan ajánlórendszer prototípusát, mely alkalmazási tartományának jellegéből nem rendelkezik azzal a jelentős méretű adathalmazzal, melyre a klasszikus rendszerek általában támaszkodnak, így a működéshez szükséges információt, tudást a sztenderd módszerektől eltérő úton kell megszerezni, illetve előállítani. Mint nagyon sok valós környezetben használt nyelvtechnológiai alkalmazás, az ECOINNO rendszer is hibrid megoldásokat alkalmaz, illeszkedve a feladat peremfeltételeihez. A rendszer alapvetően a tudásalapú megközelítéshez áll közel, de nem zárja ki más megközelítések kedvező tulajdonságainak kihasználását (pl. a felhasználói visszacsatolások figyelembevétele, elegendő felhalmozott adat esetén a felhasználóprofilok hasonlóságának monitorozása stb.).

Hivatkozások

1. Burke, R. Knowledge-based recommender systems. In: Kent, A. szerk.: Encyclopedia of Library and Information Systems, 69. kötet. New York, Marcel Dekker (2000)
2. Castells, P., Fernández, M., Vallet, D. An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering, Special Issue on "Knowledge and Data Engineering in the Semantic Web Era", **19**(2) (2007) 261–272

[10] A tudásbázis folyamatos fejlesztése miatt publikus adatok még nem állnak rendelkezésre.

3. Crestani, F., Lee, P. L. Searching the web by constrained spreading activation. Information Processing & Management, **36**(4) (2000) 585–605
4. Griffith, J., O'Riordan, C., Sorensen, H. A Constrained Spreading Activation Approach to Collaborative Filtering. In: Gabrys, B., Howlett, R.J., Jain, L.C. szerk. Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, 4253. kötet. Berlin, Heidelberg, Springer (2006) 766–773
5. Gunawardana, A., Shani, G. A survey of accuracy evaluation metrics of recommendation tasks. The Journal of Machine Learning Research, **10** (2009) 2935–2962
6. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. szerk. Recommender Systems Handbook. Springer (2011)
7. Thompson, M., Göker, C., Langley, P. A Personalized System for Conversational Recommendations. Journal of Artificial Intelligence Research, **21** (2004) 393–428
8. Vallet, D., Fernández, M., Castells, P. An Ontology-based Information Retrieval Model. In: Proceedings of the 2nd European Semantic Web Conference (ESWC 2005), Heraklion, Greece (2005) 455–470

4FX: félig kompozicionális szerkezetek automatikus azonosítása többnyelvű korpuszon

Rácz Anita[1], Nagy T. István[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, TTK, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail: raczanita89@gmail.com, nistvan@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat Jelen tanulmányunkban ismertetjük gépi tanulási módszeren alapuló megközelítésünket, mely segítségével négynyelvű párhuzamos korpuszon automatikusan azonosítottuk a félig kompozicionális szerkezeteket (FX). Elsőként felderítettük a lehetséges jelölteket a magyar, angol, német és spanyol jogi szövegekben, majd egy gazdag jellemzőkészleten alapuló bináris osztályozó segítségével azonosítottuk e szerkezeteket. Ennek az alapvetően adatvezérelt módszernek az alapja a manuálisan annotált 4FX korpusz. Ezáltal lehetőségünk nyílik az FX-ek nyelvspecifikus sajátosságainak vizsgálatára. A 4FX korpusz, illetve a négy nyelvre megvalósított automatikus detektáló hozzájárulhat számos számítógépes nyelvészeti alkalmazás, például gépi fordítók hatékonyságának javításához is.
Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, felszíni szintaktikai elemzés

1. Bevezetés

A félig kompozicionális szerkezetek (FX) az összetett kifejezések egyik típusa, melyek egy igei és egy főnévi komponensből épülnek fel. A főnév főként a szemantikai funkciókért, míg az ige többnyire csupán a szerkezet igeiségéért felel [1], mint például *irányt ad*, *forgalomba hoz* vagy *ajánlatot tesz*. Az FX-ek emellett szintaktikai, lexikai, szemantikai, pragmatikai vagy statisztikai szempontból idioszinkratikus tulajdonságokkal bírnak [2]. Ezen jellemzők, valamint relatív gyakoriságuk miatt tehát számos természetesnyelv-feldolgozó alkalmazás számára kulcsfontosságú lehet e szerkezetek folyó szövegben történő azonosítása.

A számítógépes nyelvészet számára ugyanakkor ez komoly kihívást jelent, hiszen az FX-ek *(segítséget kap)* felépítése szintaktikailag gyakorta egybeesik egyéb (kompozicionális) szerkezetekével (*pénzt kap*), valamint idiomatikus kifejezésekével (*vérszemet kap*). Másrészt mivel jelentésük nem teljes mértékben kompozicionális, így összetevőik különálló lefordítása is csak ritka esetben eredményezi az FX adott idegen nyelvi megfelelőjét. A nyelvek FX-einek eltérő sajátosságai pedig további nehézségeket jelenthetnek az automatikus azonosítás számára.

Ezen sajátosságok figyelembevételével törekedtünk jelen munkánkban az FX-ek négy nyelven történő automatikus azonosítására. Kiindulópontunkat a magyar, német, angol és spanyol jogi szövegekből felépülő 4FX elnevezésű párhuzamos korpusz képezte, amelyben első lépésként a félig kompozicionális szerkezetek kerültek manuális annotálásra. Ahogyan azt a következőkben bemutatjuk, a kézi annotáció által nem csak a különböző nyelvek FX-einek összehasonlítására nyílt lehetőségünk, hanem nyelvspecifikus tulajdonságaik feltárására is, melyek egy gépi tanuló alapjait képezték. Ezen adatvezérelt megközelítés a magyar nyelvre már bemutatott eljáráson alapszik [3], melyet sikeresen adaptáltunk a három másik nyelvre azok sajátosságainak figyelembevételével. A módszer szintaktikai elemzésre épülő FX-jelöltkiválasztó megközelítésre épül, mely a potenciális FX-ekről egy gazdag jellemzőtérre támaszkodó gépi tanuló algoritmus segítségével hoz döntést.

2. Kapcsolódó munkák

A félig kompozicionális szerkezetek automatikus felismerésére, valamint a főnév + ige szerkezetek azonosítására már számos nyelvben kísérletet tettek, például az angolban [4–7], a hollandban [8], a németben [9], valamint a baszkban [10].

A többszavas kifejezések identifikálásában rendkívüli fontossággal bírnak a párhuzamos korpuszok. Ennek kapcsán Caseli és munkatársai [11] egy olyan osztályozáson alapuló módszert dolgoztak ki, mely portugál-angol párhuzamos korpuszból képes kinyerni az FX-eket.

Samardžić és Merlo [9] angol és német nyelvű párhuzamos szövegállományban található félig kompozicionális szerkezeteket vizsgálva jutott arra a megállapításra, hogy az FX-ek párhuzamosításánál a gyakorisági adatok mellett nyelvi jellemzők is fontos szerepet játszanak.

Zarrieß és Kuhn [12] bemutatta, hogy a többszavas kifejezések hatékonyan detektálhatóak a parallel szövegekben fordítási párhuzamok alapján.

Attia és munkatársai [13] pedig arab többszavas kifejezések azonosításakor támaszkodtak a Wikipedia-bejegyzések párhuzamos címeiben található aszimmetriákra.

Ismereteink szerint az itt bemutatott az első olyan négynyelvű párhuzamos korpusz, amelyet a többszavas kifejezések egyidejű azonosítására használtak fel. A továbbiakban részletezzük a felhasznált korpusz tulajdonságait, valamint az FX-ek nyelvspecifikus sajátosságait.

3. A korpusz

A korpusz kialakítása során a JRC-Acquis [14] párhuzamos korpuszból indultunk ki, mely európai uniós jogi szövegeket tartalmaz. E szöveggyűjtemény angol nyelvű megfelelőjéből véletlenszerűen választottuk ki a szövegeket, amíg a korpusz mérete a százezer tokent meg nem haladta. Ezen angol nyelvű szövegek, valamint ezek német, magyar, illetve spanyol párhuzamos megfelelői kerültek

manuális annotálásra. Az így létrejövő korpusz képezte a manuális annotáció alapját. A műveletet két magyar anyanyelvű nyelvész végezte el, akik magas szintű német, angol és spanyol nyelvtudással rendelkeztek. Az egyes nyelveken annotált korpuszok méretét az 1. táblázat mutatja be.

1. táblázat. Az egyes részkorpuszok méretei.

	Angol	**Német**	**Spanyol**	**Magyar**	**Összesen**
Mondatok száma	5220	6392	5369	4927	21908
Szavak száma	100169	99258	111266	89338	400031

Ahogy a 1. táblázat mutatja, szavak számának tekintetében a német és az angol korpusz közel megegyező, a spanyol szövegállomány tokenszáma ennél csaknem 10 százalékkal több, míg a magyaré körülbelül ugyanennyivel kevesebb. A mondatok és szavak számát egybevetve ugyanakkor megfigyelhető, hogy az angolhoz képest a spanyol nyelvben jóval hosszabb mondatok jellemzőek, a németben inkább a „több rövidebb mondat" elve érvényesül, míg a magyar mondatok hosszúsága az angoléhoz közelít. A nyelvek közötti eltérések ugyanakkor nem csak e tekintetben válnak nyilvánvalóvá, hanem, amint azt a következőkben bemutatjuk, az FX-ek számát és felszíni formáját illetően is lényeges különbségek állapíthatók meg.

4. Annotációs elvek

Az FX-ek minél egységesebb annotálása érdekében bizonyos alaptételeket tartottunk szem előtt. Ezek a SzegedParalellFX [15] kialakítása során alkalmazott irányelveket foglalták magukban, azaz olyan kérdéseket, mint például *A főnévi komponenssel morfológiailag megegyező tövű főige képes-e helyettesíteni a szerkezetet?, Az ige elhagyásával rekonstruálható-e az eredeti cselekvés?, A szerkezet nominalizálható, illetve passzivizálható-e?*. Ezen kérdéseket a magyar és az angol nyelv mellett a németben és a spanyolban is felhasználtuk.

A korpusz építése során a másik lényeges alapelv volt, hogy nem csupán a prototipikus felépítésű igei FX-eket jelöltük (VERB, pl. *forgalomba hoz*), hanem a melléknévi igenévi (PART, pl. *forgalomba hozott*), illetve a főnévi (NOM, pl. *forgalomba hozatal*) alakokat is. Emellett a félig kompozicionális szerkezetek nem folytonos változatait (SPLIT, pl. *hozta a vállalat forgalomba*) is bevontuk az annotálásba. Erre vonatkozó adatainkat a 2. táblázat mutatja be.

Az itt közölt gyakorisági statisztikák annál is inkább figyelemre méltóak, minthogy azonos szövegállomány különböző nyelvű párhuzamos variánsai képezték kiindulópontunkat. Az adatokból kitűnik például, hogy a spanyol korpuszban csaknem kétszer annyi FX található, mint azok angol megfelelőiben. Ez pedig egyértelműen alátámasztja az FX-ek különbözőségét az annotálásba bevont nyelvek között, ugyanakkor a nyelvek között is különbségek tapasztalhatóak az FX-

2. táblázat. Manuálisan annotált FX-ek gyakoriságai különböző nyelveken.

	Angol	**Német**	**Spanyol**	**Magyar**	**Összesen**
NOM	24	241	73	160	498
	5,47%	18,24%	8,34%	19,98%	17,24%
VERB	186	216	494	300	1196
	42,37%	27,94%	56,46%	37,45%	41,42%
SPLIT	79	214	119	68	480
	18,00%	27,68%	13,60%	8,49%	16.62%
PART	150	102	189	273	714
	34,17%	13,20%	21,60%	34,08%	24.72%
Összesen	439	773	875	801	2888
	100,00%	100,00%	100,00%	100,00%	100,00%

ek tekintetében. Ezen eltérések okainak és az FX-ek nyelvspecifikus jellemzőinek pontos feltárása pedig az első lépés lehet azok automatikus azonosításában.

5. FX-ek nyelvspecifikus sajátosságai

A kézi annotáció eredményeinek elemzése egyértelműen rámutat az imént említett nyelvspecifikus sajátosságokra.

A 2. táblázat egyik szembetűnő eredménye például, hogy a négy nyelv közül a németben a leggyakoribbak a nem folytonos FX-ek. Kötött szórendű nyelvről lévén szó itt az ige alapvetően a második helyen áll, argumentumainak pozíciója azonban már jóval rugalmasabb. Az FX-ek esetében ez azt eredményezi, hogy a főnévi komponens gyakran a mondat utolsó tagjaként mintegy keretes szerkezetet alkot az igével, pl.:

Diese Verordnung ***tritt*** *am 31. März 2006* ***in Kraft****.*
Ez a rendelet 2006. március 31-én ***lép hatályba****.*

E tulajdonságának köszönhetően a németben a legmagasabb a SPLIT konstrukciók száma, melynek aránya megközelíti a folytonos szerkezetekét. Ugyanakkor a német nyelv további sajátossága, hogy a főnévi alakok (NOM) száma messze meghaladja a többi nyelvben találhatókét, mely jelenségre a német szakszövegekre gyakran jellemző nominális stílus (Nominalstil) adhat magyarázatot. Elemzéseink statisztikailag is alátámasztották tehát azt a tényt, amelyet a német szakirodalom az FX-ek kapcsán gyakorta megemlít: a jogi nyelvezet sajátja a főnevesítést előtérbe helyező kifejezésmód, melynek egyik legtipikusabb indikátora a félig kompozicionális szerkezetek alkalmazása is. Ezen értékeiben a magyar nyelvhez áll a legközelebb a német [16].

Ugyanakkor ezt leszámítva azonban nem állapítható meg nagy egyezés a magyar nyelvvel. Továbbá érdekes tény például, hogy a magyarban messze a legalacsonyabb a SPLIT-es szerkezetek aránya. Ennek oka lehet, hogy egyrészt

nincsen előre meghatározott szórend, és a szavak egymásutánisága a mondat információs struktúráját tükrözi, így a nem folytonos FX-ek esetében általában a közbeékelődő információra helyeződik a hangsúly. Valószínűleg a jogi szövegek tárgyilagosságra törekedve kerülhetik bizonyos információk kihangsúlyozását, melynek köszönhetően előnyben részesítik a folytonos FX-eket. A hangsúlyok eltolódását a következő mondatok jól szemléltetik:

*A **bírságot** a kellő visszatartó hatásnak megfelelő mértékben **szabják meg**.*
*A kellő visszatartó hatásnak megfelelő mértékben **szabják meg a bírságot**.*

Az adatok emellett szembetűnően mutatják, hogy a spanyol nyelv alkalmaz leggyakrabban félig kompozicionális szerkezeteket, melyek száma csaknem kétszerese az angol FX-ekének. A szerkezetek jelentős része folytonos, ennek kapcsán pedig egy különös sajátosságát is szükséges megemlíteni a spanyol FX-eknek. Korpuszunkban több példát is találtunk ugyanis a kettős FX-ekre, melyeket a következő szerkezetek példáznak:

lleva a cabo la aproximación (közeledést hajt végre)
da lugar a malentendidos (félreértéseknek ad helyt)

Könnyen belátható, hogy a magyar nyelv számára sem idegen konstrukciókról van szó, mivel azonban ezekkel nem találkoztunk sem a német, sem az angol nyelvű korpusz annotálása során, így feltételezhető, hogy ténylegesen egy nyelvspecifikus tényezővel van dolgunk.

3. táblázat. Gépi tanuló megközelítés eredményei a különböző nyelveken.

	Szótárillesztés			**Gépi tanuló**		
	Pontosság	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Angol	78,46	29,48	42,86	70,87	61,78	66,01
Német	82,5	7,61	13,92	58,81	46,91	52,19
Spanyol	57,22	32,71	41,65	65,7	45,48	53,75
Magyar	77,65	25,09	37,93	78,55	62,79	69,79

6. Gépi tanuló megközelítés az FX-ek automatikus azonosítására

Az FX-ek folyó szövegekben való automatikus azonosítására alapvetően a [3] megközelítést alkalmaztuk. A módszer először különböző morfológiai és szintaktikai jellemzőkre alapozó jelöltkiválasztó módszerek segítségével választja ki a potenciális FX-ket folyó szövegekből, majd egy gazdag jellemzőkészleten alapuló

döntési fa mesterséges intelligencia algoritmus alapján szelektálja ki a jelöltek közül az FX-eket. A módszert alapvetően angol, valamint magyar nyelvre valósították meg, továbbá az alap jellemzőkészlet mind a két nyelv esetében ki van egészítve nyelvspecifikus jellemzőkkel. Ezt a megközelítést alkalmaztuk az angol, valamint a magyar részkorpuszon, valamint adaptáltuk spanyol és német nyelvre. Ehhez jelöltkiválasztó módszereket definiáltunk a spanyol, valamint a német nyelvre, ami az angol és magyar nyelvű módszerekhez hasonlóan történt. Továbbá szükséges volt az alap jellemzőkészletet az aktuális nyelvhez igazítani és implementálni, valamint mind a két új nyelv esetében kiegészítettük a jellemzőkészletet nyelvspecifikus jellemzőkkel. Így a német és a spanyol esetében új morfológiai jellemzőként definiáltuk a főnevek nemét, míg német esetében az összetett főneveket. A rendszert minden nyelv esetében tízszeres keresztvalidációval értékeltük ki az aktuális részkorpuszon.

A gépi tanuló megközelítésünket minden nyelv esetében összevetettük egy szótárillesztési alapmegközelítéssel. Ebben az esetben azokat az FX-eket jelöltük, amelyeket a különböző jelöltkiválasztó algoritmusok választottak ki a folyószövegből, valamint egy adott FX listában szerepelnek. A megközelítés eredményei a 3. táblázatban találhatók.

7. Eredmények

Ahogy az a 3. táblázatban is látható, gépi tanuló megközelítésünk német és spanyol nyelven elért eredményei valamivel szerényebbek az angol és magyar nyelvű részkorpuszokon elért értékekhez képest. Ennek megfelelően a legjobb eredményeket a magyar és angol nyelvű részkorpuszon értünk el 69,79-os, valamint 66,01-os F-mértékkel, melyekhez viszonylag magas pontosságértékek tartoztak. Ezzel szemben német és spanyol nyelven csupán valamivel 50-es F-mértéket meghaladó eredményeket kaptunk, melyek elsősorban a gyenge fedési eredményeknek volt köszönhető. A német nyelvet leszámítva a szótáralapú megközelítés 40-es F-mérték körüli értékeket ért el. Továbbá érdemes megemlíteni, hogy a magyar és a spanyol nyelv esetében a gépi tanuló megközelítés magasabb pontosságértéket tudott elérni a szótárillesztésnél.

8. Az eredmények értékelése, összegzés

Jelen munkákban bemutattuk 4FX elnevezésű korpuszunkat, melyben a JRC-Acquis párhuzamos, többnyelvű korpusz négy különböző nyelven manuálisan annotált FX-ei találhatók. A korpuszon egy már meglévő, gépi tanuló algoritmuson alapuló megközelítés segítségével automatikusan azonosítottuk a folyó szövegekben az FX-eket. Mivel a megközelítés korábban csak angol és magyar nyelvű FX-ek azonosítására volt képes, ezért szükséges volt azt spanyol és a német nyelvre adaptálni. Ahogy az a 3. táblázatban látható, az általunk alkalmazott gépi tanuló megközelítés robosztusnak tekinthető, mivel az négy különböző nyelven is képes volt felülmúlni a szótárillesztési alapmódszerünket. Ehhez a különböző nyelvspecifikus jellemzők is hozzájárultak.

Az egyes nyelvek közti egyértelmű eltérés alapvetően a fedésértékben mutatkozik meg. Így a német és spanyol nyelven elért gyengébb eredményekért elsősorban a fedésértékek felelnek, ami alapvetően a jellemzőkinyerő megközelítések gyengébb teljesítményének a következménye.

Német nyelvben az azonosításkor például komolyabb problémát jelentett a szabad szórend lehetőségéből fakadó nem folytonos szerkezetek magas száma, amit a szótárillesztő megközelítés meglehetősen alacsony fedésértéke is mutat. E szerény adatok azzal is magyarázhatóak továbbá, hogy bár a produktív módon képzett főnévi FX-ek a magyar mellett itt fordulnak elő a legnagyobb gyakorisággal, azonosításukra azonban még nincsen teljes mértékben felkészítve az itt bemutatott megközelítésünk.

A magyarban az azonosítási hibák a főnevek problematikája mellett főként a nyelv morfológiai sokszínűségéből fakadtak, hiszen itt az igei alakok a szám, személy, igeidő és igemód függvényében számos eltérő ragokat kaphatnak. Ugyanez érvényesnek tűnik a spanyol tekintetében is, ahol a morfológiai gazdagság miatt az igéken túl a melléknévi igenevek azonosítása, valamint a korábban bemutatott kettős FX-ek felismerése is gyakori hibaforrásnak számít. Az angol nyelv esetében a hibák egy további jellemző csoportját szükséges kiemelnünk, mégpedig a homonim alakokat. Itt ugyanis a szerkezet főnévi alakja *(to have a walk)* több esetben megegyezik a szerkezetet helyettesítő főigével *(to walk)*, ami hibát jelenthet az automatikus szófaji egyértelműsítés számára, és ez szintén növeli a hibaforrások számát.

A nyelvek tehát szembetűnő eltéréseket mutatnak az FX-ek tekintetében, ami meglehetősen eltérő nyelvspecifikus jellemzők definiálását teszi szükségessé. Továbbá, ahogyan a 2. táblázat is mutatja, az FX-ek gyakorisága is jelentősen eltér a különböző nyelvekben. Összességében azonban megállapítható, hogy a nyelvi specifikumok ellenére is lehet létjogosultsága az általunk kidolgozott megközelítésnek, melynek további finomítása jövőbeli terveink között szerepel.

Köszönetnyilvánítás

A kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében zajlott. Nagy T. Istvánt a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosítószámú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt támogatta. Mindkét projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Vincze, V.: Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses. PhD thesis, Szegedi Tudományegyetem, Szeged (2011)
2. Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of LREC-2002, Las Palmas (2002) 1934–1940

3. Vincze, V., Nagy T., I., Farkas, R.: Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In: Proceedings of ACL (Volume 2: Short Papers), Sofia, Bulgaria, ACL (2013) 255–261
4. Cook, P., Fazly, A., Stevenson, S.: Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, ACL (2007) 41–48
5. Bannard, C.: A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, ACL (2007) 1–8
6. Vincze, V., Nagy T., I., Berend, G.: Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, ACL (2011) 116–121
7. Tu, Y., Roth, D.: Learning English Light Verb Constructions: Contextual or Statistical. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, ACL (2011) 31–39
8. Van de Cruys, T., Moirón, B.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, ACL (2007) 25–32
9. Samardžić, T., Merlo, P.: Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden, ACL (2010) 52–60
10. Gurrutxaga, A., Alegria, I.: Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, ACL (2011) 2–7
11. Caseli, H.d.M., Villavicencio, A., Machado, A., Finatto, M.J.: Statistically-driven alignment-based multiword expression identification for technical domains. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore, ACL (2009) 1–8
12. Zarrieß, S., Kuhn, J.: Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore, ACL (2009) 23–30
13. Attia, M., Toral, A., Tounsi, L., Pecina, P., van Genabith, J.: Automatic Extraction of Arabic Multiword Expressions. In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, China, Coling 2010 Organizing Committee (2010) 19–27
14. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of LREC 2006. (2006) 2142–2147
15. Vincze, V., Felvégi, Zs., R. Tóth, K.: Félig kompozicionális szerkezetek a Szeged-Paralell angol–magyar párhuzamos korpuszban. In Tanács, A., Vincze, V., eds.: MSzNy 2010, Szeged, Szegedi Tudományegyetem (2010) 91–101
16. Duden: Der Duden in 12 Bänden. Das Standardwerk zur deutschen Sprache: Duden 06. Das Aussprachewörterbuch: Unerlässlich für die richtige Aussprache. Betonung ... Namen: Bd 6 (Duden Series Volume 6)): Band 6. Gebundene Ausgabe (2006)

Az utónevek eredetleírásának formalizálása az Utónévportálon

Sass Bálint, Raátz Judit

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu, raatz.judit@nytud.mta.hu

Az MTA Nyelvtudományi Intézet feladata, hogy az anyakönyvi bejegyzésre alkalmasnak minősített női és férfi utónevek jegyzékét kezelje, illetve bővítse. Folyamatos igény mutatkozik a mind újabb utónevek felvétele iránt, a lista évente 60-70 női és 40-50 férfi utónévvel egészül ki.

Az Intézet útjára indítja az Utónévportált, melynek célja, hogy a klasszikus Ladó-Bíró féle utónévkönyv [1] utódjaként szabadon hozzáférhetővé tegye az utónevekkel kapcsolatos közérdeklődésre számot tartó információkat a folyamatosan karbantartott utónévlistának megfelelően. A portál hátterét adó naprakész, hiteles utónév-adatbázis tartalmazza az egyes nevek összes fontos adatát, beleértve az aktuális gyakorisági értékeket is. A portál a már korábban is elérhető havi frissítésű jegyzéken kívül kiegészítő szolgáltatásokat nyújt.

Az adatbázisra épülő részletes keresési lehetőség elsősorban a névadáshoz kíván segítséget adni, magában foglal számos olyan szempontot, melyek a névadáskor szóba kerülhetnek. Ilyenek: a nevek hosszúsága, szótagszáma, hangrendje, névnapja mellett a részletesen kidolgozott eredet- és jelentésinformáció, adott hangzású, ritmusú vagy magánhangzókombinációval bíró nevek keresésének lehetősége, illetve a konkrét aktuális adatokra támaszkodó gyakorisági keresés.

A cél az, hogy a utónevekhez tartozó adatok értelmük, szemantikus tartalmuk szerint legyenek kereshetők. A nevekhez jópár egyszerű, explicit, atomi adat tartozik (pl. szótagszám, névnap), ami könnyen kezelhető. Az eredetleírás, vagyis az utónevek etimológiája azonban nem ilyen. Itt a szabad szövegként megfogalmazott leírás feldolgozására, az információ kivonatolására és explicit megadására van szükség ahhoz, hogy a célt megvalósítsuk, és a leírásban meglévő akár rejtett aspektusokra is külön-külön rákereshessünk.

Azt a megoldást választottuk, hogy a több mint 3500 eredetleírás mindegyikét egy szigorúan meghatározott formális alakra hozzuk, amely minden fontos tartalmi információt magában foglal, és melynek automatikus kezelése már kézenfekvő. Az ötlet egy korábbi cikkből [2] származik, mely az etimológiák formalizálására ad általános javaslatot.

A formalizálási eljárás két szakaszra bomlik. Először egységesítettük a leírásokat, azaz a különféle megfogalmazású, de azonos jelentésű fordulatokat egy szabályalapú megoldással egységes alakra hoztuk (pl.: a *becéző formája*, *becéző rövidülése*, *becéző alakja* egyaránt *becézője*-ként jelenik meg).

A második szakasz a tényleges formalizálás, ami valójában az eredetleírásban szereplő nyelvi elemek, illetve az egyik elemet a másikba alakító műveletek felfedését jelenti. Egy példa az 1. ábrán látható.

eredetleírás	*A latin Dominicus név magyar formája*
formalizált alak	latin:Dominicus [megfelelője] magyar:∼

1. ábra. A Domonkos név leírása és formalizált alakja. A leírásban két elemet (nevet) kapcsol össze a [megfelelője] művelet.

A tényleges formalizálás félautomatikus úton történt. Az egyszerűbb nevek (a nevek 55%-ának) formalizált alakját automatikusan állítottuk elő az egységesített alakból, a gyakori mintázatok alapján kialakított szabályrendszer segítségével. Az összetettebb nevekhez (pl.: Lionella *a francia 'kis oroszlán' jelentésű Leon férfinév angol Lionel változatának női párja*) manuális úton – részletes útmutató szerint dolgozó két párhuzamos annotátorral – rendeltük hozzá a formális alakot.

A formalizált leírásnak köszönhetően az eredetleírásokon belül szemantikus keresés válik lehetővé az Utónévportálon. Ennek segítségével olyan kérdésekre is választ kaphatunk, hogy közvetlenül mely nyelvből került át adott név a magyarba, vagy hogy melyek a névalkotással létrehozott neveink.

Hivatkozások

1. Ladó, J., Bíró, Á.: Magyar utónévkönyv. Vince Kiadó, Budapest (1998)
2. Sass, B.: Javaslat az etimológiai minősítés egységesítésére. In: II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2004), SZTE, Szeged (2004) 203–208

Magyar nyelvű webes szövegek számítógépes feldolgozása

Varga Viktor[1], Wieszner Vilmos[1], Hangya Viktor[1],
Vincze Veronika[2], Farkas Richárd[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{viktor.varga.1991,vilmos.wieszner,hangyav}@gmail.com,
rfarkas@inf.u-szeged.hu

[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatjuk a magyar nyelvű webes szövegek elemzésével kapcsolatos nehézségeket, elsősorban Facebook-bejegyzésekre és kommentekre támaszkodva, valamint tárgyaljuk ezeknek lehetséges javítási módjait. A webes szövegek elemzése a belőlük kinyerhető információ miatt fontos, azonban a szabályos szövegeken tanult elemzők nem képesek hatékonyan feldolgozni ezeket. A megoldást az eddigi angolra alkalmazott, illetve a magyar nyelv sajátosságaira finomhangolt módszerek hozhatják meg.

1 Bevezetés

Az emberek életének évről-évre egyre nagyobb részében van jelen az internet, főként a rajta átáramló kommunikáció (gondoljunk csak a Twitterre vagy a Facebookra). Nagy mennyiségű adat jön létre a felhasználók egymással való kommunikációja folytán, és ez sok számítógépes nyelvészeti alkalmazás számára hasznos lehet, például az információ- és véleménykinyerésnél. Az utóbbi időben ezért jelentős fontosságra tett szert a webes szövegek, főként az ún. közösségimédia-szövegek (felhasználók által írt szövegek: blogok, állapotjelentések, chatbeszélgetések, kommentek) feldolgozása.

A közösségimédia-szövegekkel (*social media texts*) és azok elemzésével foglalkozó kutatások ugyanakkor rávilágítottak, hogy nagy nehézséget okoz ezen szövegek ún. nem sztenderd nyelvhasználata, jelentősen lecsökkenti a meglévő, szabályos szövegen (mint amilyen a Szeged Korpusz [1] is) tanult elemzők hatékonyságát. Az ezzel kapcsolatos kutatások legnagyobb része angol nyelvre született ([2, 3, 4]) és ezeknek magyarra való alkalmazása – mint az a sztenderd szövegek elemzésénél is megállapítható – nem hozna tökéletes eredményt. A magyar és az angol nyelv közötti morfológiai és szintaktikai különbségek ugyanis más megközelítést, más típusú szabályok bevezetését követelik meg. Az alapvető lépések hasonlóak, normalizálni, standardszerűvé kell a szöveget, ennek kivitelezése több módon történhet.

Cikkünk célja, hogy összefoglaljuk a közösségimédia-szövegek elemzésével kapcsolatos (elsősorban a Facebook-kommentekből és -posztokból álló tesztkorpuszon végzett) eredményeket, főbb hibakategóriákat és lehetséges megoldási módjaikat.

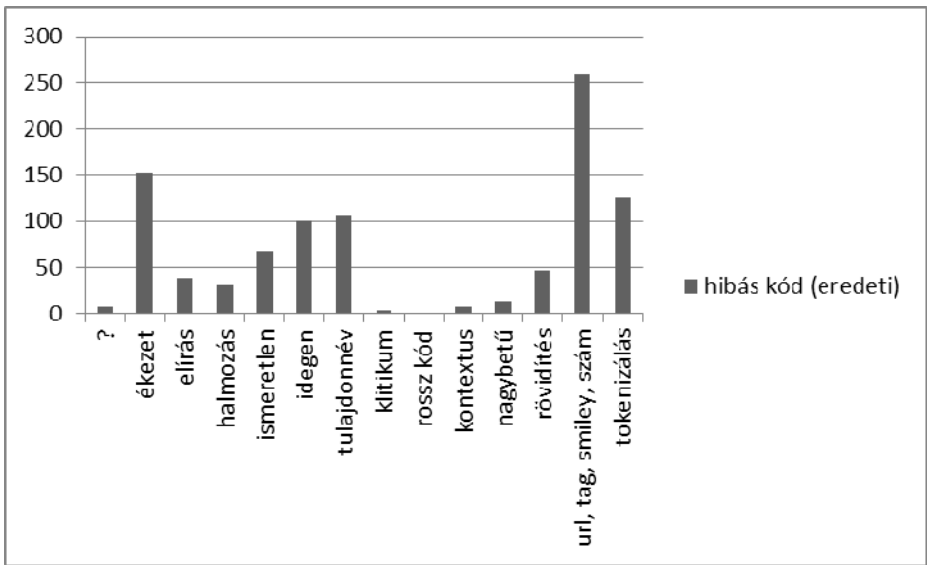
2 Problémák

A webes, azon belül a közösségimédia-szövegek nagy részének alapvető jellemzője, hogy írásbeli formájuk ellenére beszélt nyelvi sajátosságokat mutatnak. A szituációval ez könnyedén magyarázható: a szóbeli kommunikáció valósidejűségét (online) és multimodalitását egyszerre törekszik megtartani, így többek között az élmény (vagy vélemény) megosztásának gyorsasága és az érzelemkifejezés jelentős szerepet játszik a szövegekben, a hibák nagy része is ezeknek tudható be. A gyorsaságot ugyanis – a bevitelből adódóan – a gépelés gyorsításával lehet elősegíteni: többek között ékezetek mellőzésével (*ugyse* /úgyse/, *hat* /hát/, *lehet egy hulye kerdesem?)*, központozás és nagybetűk hanyagolásával, rövidítésekkel (*h*, *sztem*, *lécci*), egybeírással (*nemtom*, *énis*), valamint többnyire nem szándékoltan félregépeléssel (*mindegyekinek* /mindegyiknek/). A hétköznapi szóbeli kommunikációban elengedhetetlen érzelemkifejezés megnyilvánulhat a nagybetűhasználatban, a betű- és központozáshalmozásban (*jóóó*, *lehet ezekkel dolgozni???*), és az emotikonok használatában. Egyéb „zajok" a hezitáció explicitté tétele (*őőő*, *khm*), a nyelvi kreativitás termékeinek, illetve angol szavaknak és rövidítéseknek (*cool*, *wtf*, *pls*) a használata. Mindezek egyénenként és regiszterenként, illetve környezetenként változnak.

Az általános jellemzőkön kívül megállapítható, hogy a hibák szempontjából a közösségimédia-szöveg sem homogén kategória, az elemzők számára vannak könnyebben (blogok, Facebook-állapotjelentések) és nehezebben feldolgozható szövegek (kommentek, chat, mikroblogos bejegyzések). A blogok nagy részére jellemző a helyesírási szabályok lehetőség és képesség szerinti betartása, így ezekkel jobban boldogulnak, mint a beszélt nyelvre inkább hasonlító (akár több résztvevős) chatszövegnél, ahol a mondatra szegmentálás is problémát okoz az írásjelek és nagybetűk következetlen használata miatt.

Következő lépésben a tesztkorpuszt (150 Facebook státuszüzenet és 350 komment) a magyarlanc morfológiai és szintaktikai elemzővel [6] leelemeztük, majd kézzel részletes hibaellenőrzést végeztünk, ezután a hibákat a fentebb megállapított kategóriákba soroltuk. A különböző morfológiai hibakategóriák a nyers szövegben az 1. ábrán látható arányban fordultak elő. A számok a hibásan kódolt (X kódú, azaz le nem elemzett, illetve hibás szófaji kóddal ellátott) szóalakokat jelzik.

Az adatok azt mutatják, hogy az elemző a legtöbb hibát webcímek és egyéb kiszűrhető elemek miatt ejtette, a következő leggyakoribb a tokenizálással (szavak egybe- és különírása és egyéb szóközhiány), majd az ékezetekkel kapcsolatos hibák. Mint várható volt, az ismeretlen, de létező szavak (a diagramon *ismeretlen*, *idegen*, *tulajdonnév*, *rövidítések, kontextus* címszavak alatt) miatt történő hibák is jelentős számúak, valamint az elírás és a betűhalmozás is gyakori jelenség. A hibák természetesen halmozottan is előfordulhattak, az összetett hibákat a megfelelő hibakategóriákba külön-külön soroltuk be.



1. ábra: Morfológiai hibatípusok gyakorisága.

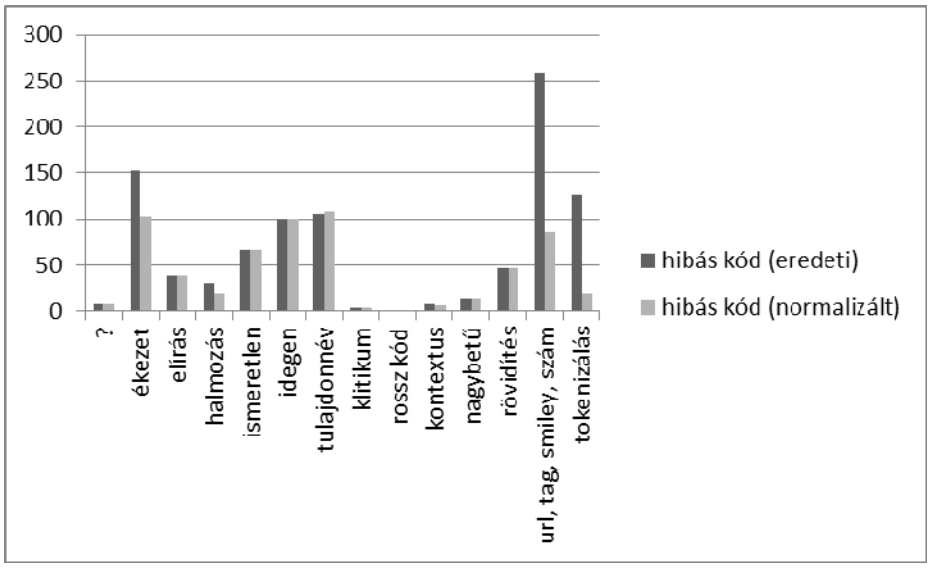
Látszik tehát, hogy a fentebb említett jelenségek a tokenizálásban és az automatikus morfológiai egyértelműsítésben problémát jelentenek, az elemző a számára ismeretlen szavakat nem tudja kiértékelni, vagy helytelen kódot ad. A kutatás egyelőre a morfológiára koncentrált, a NER tulajdonnév-felismerő [5] és a szintaktikai elemző eredményének kiértékelése folyamatban van. Annyi már látható, hogy a morfológiai hibák ezekre is hatással voltak: a helyes szintaktikai elemzéshez nélkülözhetetlen a pontos morfológiai egyértelműsítés, ami nem teljesül; a névelem-felismerő nem tudja kezelni a tiszta kisbetűvel írt neveket, a nagybetűvel írtakat – amelyeket nem látott a tanító adatbázison (pl. *Kedves Barátaim*) – pedig sokszor automatikusan névelemnek könyveli el.

3 Megoldások

A felmerült problémákat több oldalról is meg lehet közelíteni. Elméleti szempontból a hibák két csoportra oszthatók: amelyek benne vannak a tanulókorpuszban, de az elemző más alakban találkozik vele a szövegben; és amelyek semmilyen formában sincsenek a korpuszban. Az előbbire a forrásszöveg szabályalapú normalizálása (standard szöveghez hasonló formájúvá alakítása), utóbbiak nagy részére a szótár bővítése kínálhat megoldást.

Első lépésben a mondatra és tagmondatokra szegmentálást segítő, csere alapú szabályokkal (emotikonok és hiperhivatkozások egységes kezelése, szóköz és központozás helyzetének rögzítése) javítottuk a tokenizálás eredményeit. A legnagyobb problémát egyértelműen az ékezetek használata jelenti, a többi szabály elsődlegesen erre a problémakörre irányul. Az idegen ékezetek magyarra cserélése mellett toldalékokra

vonatkozó, nyelvészeti jellegű cseréket állítottunk fel (*-ság, -szerű, -ő* stb), illetve gyakori szótövek ékezetesítése (*és, csinál, tehát*, stb.). A másik normalizálási kísérlet a betűhalmozásokra irányult, ugyanis a magyarban kettőnél több azonos betű nem fordulhat elő egymást követően. A szabályok alkalmazása utáni elemzési eredmények a 2. ábrán találhatók.



2. ábra: Morfológiai hibatípusok gyakorisága a normalizálási lépések után.

Mint látható az ábrán, a kiszűrhető elemek (webcím, emotikon stb.) okozta kódolási hibák nagy része az egységes kezelés segítségével eltűnt, mint ahogy a tokenizálással kapcsolatos hibák is. A toldalék- és tőalapú ékezetesítés nem hozott akkora eredményt, azonban egy helyesírás-elemző ezzel együtt várhatóan jobb eredményt fog mutatni, mint ahogy a betűhalmozási problémák esetén is.

A szótár bővítése főként az emotikonokra, magyar és angol rövidítésekre és gyakori szavakra nyújthat megoldást, ez a munkafázis jelenleg is folyamatban van.

4 Összegzés

A közösségimédia-szövegekből kinyerhető információ egyre nagyobb jelentőségű lesz, ezek elemzése azonban – zajosságuk miatt – nem egyszerű, a standard szövegen tanult elemzők nagy hibaszázalékkal futnak le. Kutatásunk a közösségimédia-szövegekkel kapcsolatos elemzési problémák feltérképezését tűzte ki célul, számba vettük a morfológiai hibalehetőségeket és lehetséges megoldási módjukat. A kutatás jelenlegi eredményei már megkönnyíthetik egy helyesírás-elemző munkáját, ami a szöveg standardizálásának szempontjából jelentős eredményt hozhat.

Köszönetnyilvánítás

A kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123–131
2. Khan, M., Dickinson, M.: Does Size Matter? Text and Grammar Revision for Parsing Social Media Data. In: Proceedings of the Workshop on Language Analysis in Social Media (2013) 1–10
3. Liu, Fei, Weng, Fuliang, Jiang, Xiao: A Broad-Coverage Normalization System for Social Media Language. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2012) 1035–1044
4. Mott, Justin, Bies, Ann, Laury, John, Warner, Colin: Bracketing Webtext: An Addendum to Penn Treebank II. Guidelines. URL (2013. 11. 25.) = http://catalog.ldc.upenn.edu/docs/LDC2012T13/WebtextTBAnnotationGuidelines.pdf
5. Szarvas, Gy., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Discovery Science (2006) 267–278
6. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013. Hissar, Bulgaria (2013) 763–771

Morfológiai újítások a Szeged Korpusz 2.5-ben

Vincze Veronika[1,2], Varga Viktor[2], Simkó Katalin Ilona[2],
Zsibrita János[2], Nagy Ágoston[2], Farkas Richárd[2]

[1] MTA-SZTE, Mesterséges Intelligencia Kutatócsoport
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{vinczev,zsibrita,nagyagoston,rfarkas}@inf.u-szeged.hu
{viktor.varga.1991,kata.simko}@gmail.com

Kivonat: A Szeged Korpusz a legnagyobb, kézzel annotált adatbázis, amely a szóalakok lehetséges morfológiai kódjait és lemmáit is tartalmazza. Ebben a munkában bemutatjuk a korpusz újabb változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot.

1 Bevezetés

A Szeged Korpusz a legnagyobb, kézzel annotált magyar adatbázis, melyben a szavak lehetséges és a szövegkörnyezetnek megfelelő morfológiai kódjai, illetve a szavak lemmái kézzel be vannak jelölve [1]. A korpusz 2.0 verziójában található morfológiai kódok az MSD kódrendszernek felelnek meg [2]. Ebben a munkában bemutatjuk a korpusz újabb változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is kézzel hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot.

2 Harmonizált morfológiai kódok

Egy korábbi munkánkban már lefektettük a KR [3] és MSD [2] kódrendszerek harmonizálásának alapelveit [4]: a harmonizálás során arra törekedtünk, hogy az új morfológiai kódoknak olyan (és csak olyan) információkat kell tartalmazniuk, amelyek a későbbi feldolgozás (szintaxis, különféle alkalmazások) szempontjából hasznosak.

A 2.5 verzióban így a korábbi 2.0-s verzióhoz képest az alábbi morfológiai újítások találhatók:

- a gyakorító, ható és műveltető igék lemmája a képző nélküli igető lett, és a kódban jelöljük azt, hogy az ige milyen alakban áll;
- a melléknévi igenevek önálló kódot kaptak (korábban a melléknevek és az igenevek nem voltak elkülöníthetők MSD-kóduk alapján);
- tulajdonnév és köznév elkülönítésének megszüntetése;
- a személyes névmási határozószóknak a névmási rendszerbe való beillesztése.

A fenti esetekben az egyes szóalakok mellé felvettük az új morfológiai kódokat, valamint szófajilag is egyértelműsítettük a szövegeket, azaz manuálisan kiválasztottuk, hogy melyik lehetséges kód illik az adott szövegkörnyezetbe. Az alábbiakban részleteiben is ismertetjük az egyes morfológiai újításokat.

2.1 Gyakorító, ható és műveltető igék

A KR kódrendszer a gyakorító és műveltető igéket (pl. *olvasgat*, *futtat*) az alapalakból képzett igének tekinti, tehát a gyakorító és műveltető szuffixumokat képzőként kezeli. A ható igék (*mehet*) toldaléka ezzel szemben inflexiós toldaléknak számít a KR rendszerében. Az MSD kódrendszer eredetileg mindezen toldalékokat a lemma részeként kezelte, azaz míg például az *olvastak* és *olvashattak* morfológiai kódja azonos volt (Vmis3p---n), addig lemmájuk eltért: *olvas* és *olvashat*. A harmonizációnak köszönhetően a Szeged Korpuszban is jelöljük azt, hogy az ige gyakorító, műveltető vagy pedig ható-e. Az igei MSD-kód második pozíciójában jelenítjük meg ezeket az információkat, lemmának pedig az ige toldalékolatlan alakját tüntetjük fel. Arra is figyelmet fordítottunk, hogy ezen toldalékok nem zárják ki egymást, tehát egy adott igealak lehet egyszerre például műveltető és ható is. Így a toldalékok lehetséges kombinációját is meg tudjuk jeleníteni a harmonizált kódrendszerben. Az alábbi táblázat mutatja be a harmonizált kódokat:

1. táblázat: Igei harmonizált kódok.

Leírás	**Kód**	**Toldalék**	**Példa**
fő (main)	m	-	*megy*
segéd (auxiliary)	a	-	*fogok (menni)*
ható (modal)	o	-hAt	*mehetek*
gyakorító (frequentative)	f	-gAt	*pofozgat*
műveltető (causative)	s	-(t)At	*etet*
gyakorító+ható	1	-gAthAt	*boncolgathat*
műveltető+ható	2	-(t)AthAt	*fektethet*
műveltető+gyakorító	3	-(t)AtgAt	*etetget*
műveltető+gyakorító+ható	4	-(t)AtgAthAt	*futtatgathat*

Az igék újrakódolásakor különös figyelmet fordítottunk a kétértelmű esetekre, amikor ugyanaz az igealak jeleníti meg a műveltető és nem műveltető alakot. Ez el-

sődlegesen a múlt idejű igealakoknál fordult elő, amikor például a *festetted* alak jelölheti a *fest* és a *festet* múlt idejű E/2. tárgyas ragozású alakját is, kontextustól függően.

2.2 Melléknévi igenevek

Míg a KR kódrendszer a melléknevektől elkülönítve kezelte a melléknévi igeneveket, addig az MSD-ben az A szófaji kód vonatkozott a melléknevekre és a melléknévi igenevekre egyaránt. Azonban a melléknevek és a melléknévi igenevek morfológiai és szintaktikai viselkedése eltérő vonásokat mutat: a melléknevek fokozhatók, míg a melléknévi igenevek nem, vö. *az okos fiú – az okosabb fiú* és *az éneklő fiú* - **az énekelőbb fiú*, továbbá a melléknévi igenév igen gyakran megőrzi az eredeti ige vonzatszerkezetét: *a slágert jó hangosan éneklő fiú*. Mivel úgy gondoljuk, hogy e különbségek kihatással vannak a mondatok szintaktikai elemzésére is, a harmonizált kódrendszerben is bevezettük e megkülönböztetést. A melléknévi MSD-kód második pozíciójában jelenítjük meg azt az információt, hogy melléknévről vagy melléknévi igenévről van-e szó, illetve utóbbi esetben megadjuk a melléknévi igenév típusát is (folyamatos, befejezett vagy beálló). A kódokat az alábbi táblázat részletezi:

2. táblázat: Melléknévi (igenévi) harmonizált kódok.

Leírás	**Kód**	**Képző**	**Példa**
melléknév	f	-	*friss*
folyamatos melléknévi igenév	p	-Ó	*sétáló*
befejezett melléknévi igenév	s	-t/-tt	*megvásárolt*
beálló melléknévi igenév	u	-AndÓ	*felveendő*

Bizonyos szóalakok mind melléknévként, mind melléknévi igenévként használatosak, vö. *égető kérdések – a kertben tüzet égető gondnok*. Az egyértelműsítés során is a fenti különbségeket (fokozás, vonzatok) használtuk nyelvi tesztként.

2.3 Köznevek és tulajdonnevek

Az MSD kódrendszer korábbi verziójában a köznevek és tulajdonnevek külön kóddal rendelkeztek. Azonban úgy gondoljuk, hogy a köznév-tulajdonnév elkülönítés nem bír jelentőséggel a morfológia szintjén, így egy morfológiai elemzőnek nem is lehet feladata a tulajdonnevek felismerése, meghagyva az a névelem-felismerő alkalmazásoknak. Mindezekből kifolyólag a Szeged Korpusz 2.5-ös változatában eltöröltük a köznév-tulajdonnév megkülönböztetést, így minden főnévi kód egységesen Nn- kezdettel rendelkezik.

2.4 Személyes névmási határozószók

A magyar nyelvben a hagyományos terminológiával személyes névmási határozószóknak hívott szóalakok két csoportra bonthatók. Az első csoportot azok alkotják, amelyek etimológiájukat tekintve határozóragra vezethetők vissza (*bennem*, *neki*). A második csoportba azok tartoznak, amelyek névutóból eredeztethetők (*szerinted*, *mögöttünk*). Az eredeti MSD-rendszerben e szóalakok egységesen a határozószavak egy alosztályát képezték, míg a KR rendszerében mindkét csoport főnévként szerepeltek (bár a morfológiai kód felépítése eltért a két esetben).

A harmonizált kódrendszerben egyik megoldást sem vettük át, hanem névmásként kezeljük ezeket az alakokat, a személyes névmási rendszerbe illesztve. A névutóból eredeztethető alakok esetében lemmaként a névutót tüntetjük fel, a határozóragból eredeztethető alakoknál pedig a személyes névmást. Néhány példát mutatunk az alábbiakban:

3. táblázat: Névmási harmonizált kódok.

Szóalak	**Lemma**	**Morfológiai kód**
szerintem	szerint	Pp1-sn
nálunk	mi	Pp1-p3

Ezek az alakok automatikusan lettek átcímkézve, esetükben nem volt szükség kézi egyértelműsítésre.

2.5 Írásjelek

Az írásjelek morfológiai kódolásán szintén változtattunk. Az alábbi 8 írásjelet tekintjük relevánsaknak (az írásjelek mögött az ASCII kódjuk szerepel): !(33) ,(44) -(45) .(46) :(58) ;(59) ?(63) –(8211).

A releváns írásjelek lemmája maga az írásjel lesz, morfológiai kódja szintén. Egyéb nem releváns írásjelek (olyan karaktersorozatok, melyek nem tartalmaznak sem betűt, sem számot) lemmája szintén maga az írásjel lesz, de kódja K (központozás) lesz.

2.6 Elváló igekötők

Az elváló igekötőt tartalmazó igei elemek (igék, főnévi, melléknévi és határozói igenevek) lemmájában megjelöltük az igekötő-igei elem közti morfémahatárt. Mivel bizonyos szintaktikai műveletek hatására az ige és igekötő elválhat egymástól, úgy döntöttünk, hogy ezekben az esetekben jelöljük a morfémahatárt a lemmában.

3 Helyesírási hibák javítása

A morfológiai javítások mellett figyelmet fordítottunk a helyesírási hibák javítására is. A korpusz 2.0 változatában külön MSD-kóddal rendelkeztek a rossz helyesírású

(elírt, elgépelt) szavak (pl. *kiráj*), illetve azok, melyek értelmes magyar szavak, azonban a szövegkörnyezetbe nem illettek bele (*mer úgy gondolom* vs. *mert úgy gondolom*). Amennyiben a helyes és az elírt alak azonos tokenszámú egységet tartalmazott, úgy a helyesírási hibát vagy elírást tartalmazó szóalakok mellé felvettük azok helyes alakját is annak lehetséges MSD-kódjaival együtt, majd a szövegkörnyezetnek megfelelően kiválasztottuk az aktuális kódot. Azokban az esetekben pedig, ahol a helyes és helytelen alakok tokenszáma között eltérés mutatkozott (pl. *areggel* vs. *a reggel*), a fő szóalak morfológiai kódját vettük fel (pl. egy egybeírt névelő és főnév esetén a főnévi címkét).

4 Statisztikai adatok

A Szeged Korpusz 2.0 verziója 1,2 millió tokent tartalmazott (egy tokennek számítva a többtagú tulajdonneveket). Ezek közül 11 461 token minősült ismeretlen vagy rossz helyesírású szónak. A 2.5-ös verzióban e szavak száma mindösszesen 1563 lett, azaz a morfológiai elemzés számára problematikus szavak aránya 1%-ról 0,13%-ra csökkent, ami jelentős – egy nagyságrendnyi – változást jelent: a problémás szavak 86,4%-át sikerült kijavítani.

A korpusz jelen változatában az ismeretlen szavak legnagyobb része angol számítástechnikai terminus. Ez arra vezethető vissza, hogy a számítógépes szövegek alkorpuszban gyakran szerepelnek az eredeti angol megnevezések is a felhasználói kézikönyvek szövegeiben.

A korpusz 2.5 változatában összesen 1315 morfológiai kód szerepel. Az alábbi táblázat mutatja be az újonnan bevezetett kódok előfordulásait:

4. táblázat: Új kódok gyakorisága

Leírás	**Kód**	**Előfordulás**
Folyamatos melléknévi igenév	Ap*	23483
Befejezett melléknévi igenév	As*	12588
Beálló melléknévi igenév	Au*	520
Melléknévi igenév összesen	Ap*, As*, Au*	36591
Műveltető ige	Vs*	1698
Ható ige	Vo*	8415
Gyakorító ige	Vf*	327
Műveltető/ható/gyakorító kombinációja	V1*, V2*, V3*, V4*	67
Műveltető/ható/gyakorító igék összesen	Vs*, Vo*, Vf*, V1*, V2*, V3*, V4*	10057

A személyes névmási határozószók újrakódolása további 8232 tokent érintett. Ha összegezzük tehát a megváltozott kódú szavakat (melléknévi igenevek, műveltető/ható/gyakorító igék, személyes névmási határozószók, javított helyesírási hibák), akkor összesen 64 788 szóalak kódja változott meg, ami a korpusz szavainak 4,36%-a.

5 Morfológiai elemző

A Szeged Korpusz 2.5 változata lehetővé tette, hogy a *magyarlanc* nevű adatvezérelt nyelvi elemző [5] morfológiai és szófaji egyértelműsítő moduljait az új adatbázison tanítsuk be, létrehozva ezzel az elemző újabb változatát, mely a morfológiai elemzés és szófaji egyértelműsítés végeredményeként az új harmonizált morfológiának megfelelő kódokat ad vissza.

A korpusz teljes állományát véletlenszerűen osztottuk fel tanító és kiértékelő adatbázisra 80:20 arányban, majd a tanítást követően értékeltük a szófaji egyértelműsítő teljesítményét. Akkor fogadtuk el helyesnek a *magyarlanc* által adott elemzést, ha mind a lemma, mind pedig a morfológiai kód egyezett az etalon korpuszban lévővel. Eredményeink szerint a *magyarlanc* szófaji egyértelműsítő modulja az új kódrendszer használatával 96,32%-os pontosságot ér el, ami megegyezik a korábban publikált, Szeged Korpusz 2.0 verzión tanított rendszer eredményességével [5], vagyis az elemzés minőségét nem befolyásolja érdemben a megnövekedett kódhalmaz.

6 Összegzés

Ebben a munkában bemutattuk a Szeged Korpusz 2.5 változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot. A korpusz lehetővé tette azt is, hogy a magyarlanc morfológiai elemző és szófaji egyértelműsítő modulját az új szófaji kódokra tanítsuk be. Eredményeink alapján a szófaji egyértelműsítés minősége változatlanul magas a megnövekedett kódhalmaz ellenére is.

A korpusz kutatási és oktatási célokra szabadon hozzáférhető a http://www.inf.u-szeged.hu/rgai/SzegedTreebank oldalon.

Köszönetnyilvánítás

A kutatás – részben – a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123-131
2. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf

3. Kornai, A., Rebrus, P., Vajda, P., Halácsy, P., Rung, A., Trón, V.: Általános célú morfológiai elemző kimeneti formalizmusa. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2004) 172–176
4. Farkas, R., Szeredi, D., Varga, D., Vincze, V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (2010) 349–353
5. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013)

A határozott és határozatlan ragozás hibáinak automatikus felismerése magyarul tanulók szövegeiben

Vincze Veronika[1], Zsibrita János[2], Durst Péter[3], Szabó Martina Katalin[4]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
zsibrita@inf.u-szeged.hu

[3] Szegedi Tudományegyetem, Hungarológia Központ
durst.peter@gmail.com

[4] Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
szabomartinakatalin@gmail.com

Kivonat: Jelen munka célja, hogy a HunLearner magyar nyelvtanulói korpuszban automatikusan azonosítsuk a határozott és határozatlan igeragozásban elkövetett nyelvtanulói hibákat. A hibaelemzés rámutat a nyelvtanulók számára nehézséget okozó nyelvtani szerkezetekre, ami az adott jelenségek célzott oktatásában és gyakorlásában hasznosítható a nyelvoktatás felől nézve, számítógépes oldalról pedig egy nyelvhelyesség-ellenőrző továbbfejlesztésében lehet hasznos.

1 Bevezetés

A jelen dolgozatban a HunLearner magyar nyelvtanulói korpuszban [1] folyó munkálatok egyik részfeladatáról számolunk be. A projekt a határozott és határozatlan igeragozásban elkövetett nyelvtanulói hibák automatikus azonosítását tűzte ki célul. Az általunk vizsgált ragozásnak több elnevezése is elterjedt (*tárgyas ragozás*, *határozott ragozás*, *határozott tárgyas ragozás*, vö. [2]), ebben a dolgozatban *a határozott tárgyas ragozás* terminust használjuk.

Munkánkban először röviden ismertetjük a határozott és határozatlan tárgyak típusait. Ezek után bemutatjuk a vizsgálatunk alapjául szolgáló HunLearner korpusz bővített változatát, majd megmutatjuk, miként lehetséges automatikus eszközökkel azonosítani a határozott ragozásban elkövetett hibákat. A leggyakoribb hibatípusokról végül statisztikai elemzéseket is adunk.

2 Határozott tárgyak

A magyar nyelv sajátosságai közül kiemelkedik a határozott tárgyas ragozás, amely kifejezetten kevés nyelvben figyelhető meg. Széles körben elterjedt elnevezése a rövidebb tárgyas ragozás terminus, a grammatikák azonban inkább határozott tárgyas ragozásként említik [2]. A határozott igei paradigma használati szabályainak elsajátítása és alkalmazása gyakran okoz nehézséget a magyar nyelv tanulói számára, ráadásul a határozott tárgy különböző típusai is eltérő mértékben okoznak nehézséget a nyelvtanulás során. A határozott ragozást a struktúrában megjelenő ún. határozott tárgy hívja elő, tehát a tárgy határozottságát jelölni kell az igén. Ezt harmadik személyű tárgyakkal tudjuk kifejezni teljes paradigmában, a második személyű tárgyak jelölésére csak hiányos ragozási sor áll rendelkezésre a magyarban (vö. *ismerem őt*, *ismered őt* vs. *ismerlek téged*).

A határozott ragozás több nyelvi szinten átívelő jelenség, amelynek lényegét M. Korchmáros nyelvtanában [3] így foglalja össze: „Általában akkor beszélünk a magyar igeragozás szempontjából megkülönböztetett határozott tárgyról, ha az a beszélő és a hallgató tudatában egyforma mértékben azonosított egyedi vagy annak tekintett objektum(ok)at jelöl." Ez az egyébként nagyon pontos megfogalmazás azonban még nem ad elég fogódzót sem a magyar nyelv határozott tárgyas ragozásának elsajátításához, sem pedig annak számítógépes feldolgozásához; mindenképpen szükség van a határozott tárgyas ragozást megkövetelő határozott tárgyi tömbök pontos és részletes bemutatására. A leggyakoribb és a nyelvtanulók számára is a legkisebb nehézséget jelentő határozott tárgyak a következők:

1. A tárgy tulajdonnév:
Ismerem Klárit.

2. A tárgy határozott névelővel álló névszó:
Megesszük az almát.
Elviszem a pirosat.

3. A tárgy főnévi mutató névmás:
Ezt kérem.

4. A tárgy birtokos személyjellel vagy *-é* birtokjellel álló névszó:
Mindenki ismeri a testvéremet.
A Katiét vették meg.

5. A tárgy visszaható / kölcsönös / birtokos névmás:
Mindenki magát látja a tükörben.
Szeretik egymást.
A mienket ne vidd el.

6. A tárgy harmadik személyű személyes névmás:
Ismerem őt.

Érdekesség, hogy a személyes névmások közül csupán a harmadik személyűek számítanak határozott tárgynak, hiszen a határozott tárgyas ragozás alapvetően csak harmadik személyű tárgyra tud utalni.

7. A tárgy *-ik* kijelölő jellel áll:
Csak az egyiket kérem.
Melyik könyvet olvastad?
Hányadikat eszed már?

Meg kell jegyezni, hogy a *Melyik?* és a *Hányadik?* kérdőszón kívül más kérdő névmás nem minősül határozott tárgynak.

8. A tárgy egy mellékmondat:
Tudom (azt), ki vagy.

A tárgyi alárendelő mellékmondatok több formában is előfordulhatnak, hiszen a főmondatban nem jelenik meg szükségszerűen az *azt* utalószó. Ez a változatosság mind a nyelvtanulók, mind a számítógépes nyelvfeldolgozás szempontjából igen problematikusnak tekinthető.

9. A tárgy a *mind* vagy a *valamennyi* névmás:
Mind elolvasta.
Valamennyit megették.

A *valamennyi* névmást illetően fontos hangsúlyozni, hogy az csupán annak 'összeset' jelentésében jár határozott ragozással. Ennek következtében a szerkezet használatának elsajátítását tovább nehezíti, hogy esetében csak a szövegkörnyezet segítségével lehet eldönteni, hogy milyen ragozást kell használni.

10. **A tárgy explicit módon nem jelenik meg a mondatban:**
Add ide!
Tegnap vettünk egy esernyőt. Ma elvesztettük.

Az explicit módon nem realizálódó határozott tárgy főleg a párbeszédes formájú szövegekre jellemző, és, mivel az adott szerkezetben fonológiailag nem realizálódik, az adott kontextus mutatja meg a szerkezetben való létezését. Ilyenkor vagy egy a szövegben már korábban említett, vagy pedig egy nyelven kívüli eszközökkel (pl. rámutatás) azonosított tárgyról van szó.

3 Kapcsolódó irodalom

A számítógépes nyelvfeldolgozás szempontjából a határozott tárgy kezelése problematikusnak tekinthető, ugyanis mint láttuk, a határozott tárgyi tömbök morfológiai megjelenése nem egységes, emiatt automatikus felismerésük bizonyos esetekben akadályokba ütközik. A témához kapcsolódó korábbi korpuszalapú kutatások között találunk kínai anyanyelvűekkel végzett, szóbeli mintavételen alapulót [4], eltérő anyanyelvű válaszadókkal végzett kérdőíves tesztelést [5], valamint egy ugyancsak kérdőíven alapuló vizsgálatot homogén, mordvin anyanyelvű csoporttal [6]. Ugyanakkor meg kell említenünk, hogy – a jelen projekttől eltérően – egyik esetben sem használtak még automatikus eszközöket a határozott tárgy, valamint a határozott ragozásban vétett nyelvtanulói hibák feldolgozásának céljából.

4 A HunLearner korpusz

A HunLearner korpusz magyar mint idegen nyelv szakos egyetemi hallgatók fogalmazásait tartalmazza [1]. Horvát anyanyelvű diákok három nagyobb témában írtak esszét: *Egy szimpatikus ember*, *Nehézségek a magyar nyelv tanulásában*, *illetve Magyar bevándorlók Angliában*. A korpuszban a főneveket érintő morfológiai hibákat kézzel javítottuk, és minden hibához automatikusan hozzárendeltük annak típusát.

A korpusz néhány új szöveggel bővült a közelmúltban. Ezeket észt diákok írták az *Egy szimpatikus ember* témában. A korpusz jelen, kibővített változatában 1427 mondat és 22 000 token szerepel.

5 Határozott ragozási hibák a korpuszban

A HunLearner korpusz szövegeit a magyarlanc szoftverrel [7] automatikusan elemeztük, majd a morfológiai és szintaktikai elemzés alapján szabályokat definiáltunk az tárgy-ige egyeztetés különböző típusaira. Ezek alapján automatikusan össze tudtuk gyűjteni azokat az eseteket, amelyekben eltérés mutatkozott a tárgy típusa által indikált és a tényleges igeragozás között. Például: megvizsgáltuk, hogy a köznévi tárgy rendelkezik-e névelővel. Amennyiben rendelkezik határozott névelővel, az igeragozásnak határozottnak kell lennie.

Az alábbi példában a főnévi igenév mutató névmási tárgya határozott ragozást váltana ki a *szeret* igén, azonban a nyelvtanuló határozatlan ragozást használ: *Végül mindenkinek **szeretnék** azt mondani, hogy Angliában tök jobb életem van, mint Magyarországban.*

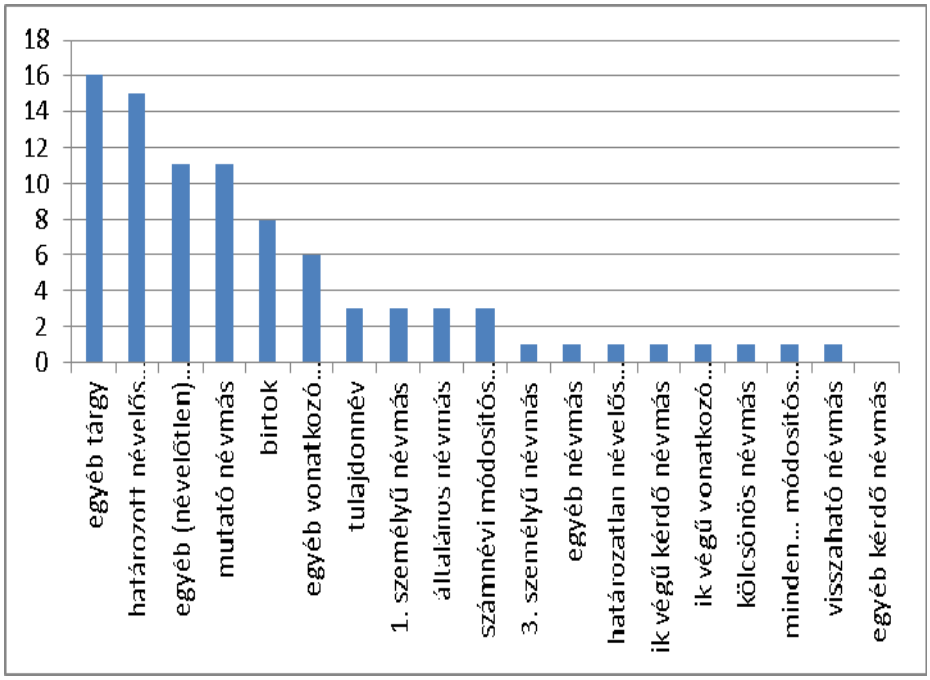
Az 1. táblázat mutatja a vizsgálat számszerű eredményeit. Jelen cikk keretei között csak azokat az eseteket vizsgáltuk részletesebben, ahol a tárgy fonológiailag is jelen van a mondatban (*Van tárgy a mondatban* oszlop), tehát egyelőre nem foglalkozunk azokkal az esetekkel, amikor a névmási tárgy jelenléte pusztán a határozott ragozású igéből lenne kikövetkeztethető. Az alárendelő mellékmondati tárgyakat is kizártuk a vizsgálatból, hiszen a tárgyi szerepet betöltő mellékmondatok automatikus azonosítására jelenleg nem képes a magyarlanc szintaktikai modulja. Kizártuk a vizsgálatból továbbá azokat a morfológiailag többértelmű igealakokat is, ahol a határozott és határozatlan ragozás egybeesik (pl. múlt idő E/1. alakban, vö. *olvastam*), itt ugyanis nem eldönthető, hogy a nyelvtanuló határozott vagy határozatlan ragozást kívánt-e használni.

A szűrések után kapott 87 esetet további vizsgálatoknak vetettük alá. Az eredmények szerint a leggyakoribb hibaforrás a határozott névelős köznévi tárgy: ez határozott ragozást váltana ki, azonban a hibák 17%-ában határozatlan ragozású igével szerepel együtt. Két másik gyakori hiba a mutató névmási tárgy és a névelőtlen köznévi tárgy, melyek a hibák 13-13%-ában a nem megfelelő ragozású igével fordulnak elő. A birtokos személyjellel ellátott tárgyakat érintő hibákat is ideszámítva elmondhatjuk, hogy a határozott ragozást érintő hibák 50%-áért a fenti hibák felelnek.

1. táblázat: Ragozásbeli eltérések.

Alkorpusz	Igék száma	Ragozásbeli eltérés	Van tárgy a mondatban	Egyértelmű igealak
Nehézségek	1018	149	42	32
Anglia	564	74	46	16
Szimpatikus	841	149	47	39
Összesen	2423	372	117	87

Az 1. ábra mutatja a hibásan használt igeragozást kiváltó tárgytípusok gyakoriságát.



1. ábra: Hibás igeragozást kiváltó tárgyak.

Az eredmények egyben azt is mutatják, hogy jóval több a határozott tárgy-határozatlan igealak típusú tévesztés (59%), mint a határozatlan tárgy-határozott igealak típusú.

6 Az eredmények felhasználása

A vizsgálat eredményeit egyrészt kitűnően hasznosíthatja a nyelvoktatás, hiszen a hibák statisztikai elemzése lehetőséget nyújt arra, hogy a nehezebbnek bizonyuló szerkezeteket célzottan gyakorolhassák a diákok a nyelvórán. Másrészt számítógépes

nyelvészeti oldalról nézve az egyeztetési hibák automatikus hibajavítása előtt is megnyílik a lehetőség, hiszen a tárgy típusa alapján meg lehet határozni az elvárt igealakot, és amennyiben nem a megfelelő szerepel a szövegben, egy nyelvhelyesség-ellenőrző program javítási javaslatokat tehet az igealakra nézve.

7 Összegzés

Ebben a munkában bemutattuk számítógépes nyelvészeti eszközökön alapuló megközelítésünket, mely a határozott és határozatlan ragozásban elkövetett hibák automatikus azonosítását célozza. A vizsgálatból kiderült, hogy melyek azok a nyelvtani szerkezetek, amelyek problémát jelentenek a magyart mint idegen nyelvet tanulók számára. Ezen eredmények haszna elsődlegesen a nyelvoktatásban mutatkozik meg, hiszen a nyelvtanulók így célzottan gyakorolhatják a problémásabb szerkezeteket, mindemellett a határozott és határozatlan ragozás hibáinak automatikus azonosítása egy nyelvhelyesség-ellenőrző programban is jó szolgálatot tehet.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze V., Zsibrita J., Durst P., Szabó M. K.: HunLearner: a magyar nyelv nyelvtanulói korpusza. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 97–105
2. Pete I.: A határozott tárgyas ragozásról. Magyar Nyelvőr, Vol. 130. (2006) 317–324
3. M. Korchmáros V: Lépésenként magyarul. Magyar nyelvtan – Nem csak magyaroknak.. Szegedi Tudományegyetem, Szeged (2006)
4. Langman, J., Bayley, R.: The acquisition of verbal morphology by Chinese learners of Hungarian. Language variation and Change, Vol. 14 (2002) 55–77
5. Durst P.: A magyar főnévi szótövek és egyes toldalékok elsajátításának vizsgálata magyarul tanuló külföldieknél. Hungarológiai Évkönyv, Vol. 11. Pécs (2010)
6. Durst, P., Janurik, B.: The Acquisition of the Hungarian definite conjugation by learners of different first languages. Lähivõrdlusi. Lähivertailuja 21. Tallinn: Estonian Association for Applied Linguistics (EAAL) (2011) 19-44
7. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013) 763–771

VIII. Laptopos bemutatók

Magyar hangsúly-adatbázis az interneten kutatáshoz és oktatáshoz

Abari Kálmán[1], Olaszy Gábor[2]

[1] DE Pszichológiai Intézet, Szociál- és Munkapszichológiai Tanszék
abari.kalman@arts.unideb.hu

[2] BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Kivonat: Hangsúlyadatbázis eddig nem készült magyar nyelvre. Beszédtechnológiai kutatásokban és az oktatásban is nagy igény lenne egy referenciaként használható, helyes hangsúlycímkéket tartalmazó mondatgyűjteményre. Fejlesztésünkkel ezt a hiányt kívántuk pótolni. A webes lekérdezőfelülettel rendelkező hangsúlyadatbázisunk 1869 kijelentő mondatot tartalmaz, amelyekben minden szó hangsúlypozícióját megjelöltük. A honlapon számos szempont alapján kereshetünk és a találati lista is több formában (szöveges, kép és hang) jeleníthető meg. A vizsgált magyar mondatok hangsúlymintázatainak gyakorisága is lekérdezhető. A honlap a http://magyarbeszed.tmit.bme.hu/hangsuly címen érhető el.

1. Bevezetés

Az itt bemutatott adatbázis a magyar hangsúlyozás szöveges tartalmon való előrejelzéséhez nyújt interaktív támogatást. Az adatbázis hangsúlycímkékkel ellátott magyar mondatokat tartalmaz szöveges formátumban, melyekhez képi megjelenítések és a meghangosított mondatok hangállományai társulnak. Az adatbázis olyan mondatkorpuszra támaszkodik, amelynek két kiindulópontja van. Az alapot egy korábbi gépi beszédfelismerési kutatáshoz alakították ki [5] úgy, hogy fonetikailag kiegyensúlyozott mondathalmazt hoztak létre irodalmi művek szövegeiből. Az adatbázis másik forrása a BME TMIT-en erre alapozott és elkészített párhuzamos, precíziós beszédadatbázis 12 beszélővel [2]. Mindezekből logikusan adódott az a gondolat, hogy erre a mondathalmazra alapozva elkészítsük a mondatok hangsúlyozási jelekkel kibővített szöveges változatát is, ami egyfajta támaszt adhat későbbi hangsúlykutatásokhoz, valamint felhasználható az oktatásban is. A munka 3 évig tartott. Az adatbázis 1869 kijelentő mondatot tartalmaz. Mivel a hangsúlyozással kapcsolatos nyelvészeti irodalom szerteágazó, el kellett döntenünk, hogy milyen formában közelítünk a témához. A hangsúly jelölésének a legegyszerűbb változatát választottuk, bináris felépítésben gondolkodtunk, vagyis azt jelöltük, hogy van hangsúly (W) vagy nincs hangsúly (N). A másik egyszerűsítésünk, hogy csak szóhangsúlyokat jelölünk a szövegben, ezt is következetesen, azaz minden szó kap egy W vagy N címkét. (Itt megjegyezzük, hogy ezeket a címkéket csak a kutatható adatbázisban láthatja a felhasználó, a mondatlistá-

ban a W címkét kiemeléssel helyettesítjük, ahogy ezt a formát alkalmazzuk e tanulmány sok példájában is a könnyebb olvashatóság kedvéért. A hangsúly mintázatok bemutatásánál pedig a W-t H jelöléssel helyettesítjük, az N-t pedig a - karakterrel.) A harmadik egyszerűsítésünk, hogy a hangsúlyok fizikai kivitelezésénél csak az alapfrekvenciát használtuk a hangsúly élmény megvalósításához, úgy, hogy az első szótagi magánhangzókon erőteljes alapfrekvencia (F0) emelést hajtottunk végre, majd a második magánhangzóig visszavezéreltük az alapfrekvenciát a mondat dallamvonulatát meghatározó alapra. A negyedik egyszerűsítés a mondatok szöveges ábrázolását érinti. A feldolgozás menete során csak kisbetűket alkalmazunk, így a mondatlista mondataiban minden karakter csak kisbetűvel szerepel, és mondatvégi írásjel sincs a mondatok végén.

A hangsúlyadatbázis referencia vizsgálatát külön kutatásban végeztük el [3], az interneten közreadott változat már tehát egyfajta referenciának tekinthető, amely közvetlenül tanulmányozható három formában. Ezek közül a legmegfoghatóbb a szövegbe elhelyezett hangsúlycímkék állománya. A címkék és a mondat tartalma lehetőséget ad keresésekre és csoportosításokra is. A hangformátumot beszédszintetizátorral állítottuk elő az adott mondat hangsúlyjelei szerint. Ez adja az akusztikai megjelenítést. Az F0 változást képben is megjelenítjük, tehát az összevethető a hangzó és az írott, címkézett formával. Az adatbázis hangsúly jelölése nagy pontosságúnak tekinthető, amit úgy kell érteni, hogy nincs benne címkézési hiba, vagyis ahol hangsúlyt jelöltünk az adott szóra, ott a hangsúlyos ejtés nem okoz megértési zavart, furcsa ejtést és fordítva. Vannak olyan mondatok, amelyek többféle hangsúlykiosztással is ejthetők az értelmezés, illetve a közlési szándék szerint. Ezeknél a mondatoknál az egyik helyes formát adják a jelölések.

Az adatbázis érdekessége, hogy kontrasztokat is bemutat hangban, tehát az érdeklődő tanulmányozhatja a jó hangsúlyozással megvalósított mondatot, valamint ugyanannak a mondatnak két másik változatát is. Az egyikben a hangsúlyozást létrehozó alapfrekvencia csúcsokat megszüntettük, ez egyfajta neutrális szerkezetet eredményez, ami érzeti szempontból nem biztos, hogy élesen érzékelhető a hangzásban. A másik kontrasztban a rossz hangsúlyozást próbáltuk megvalósítani, többnyire megfordítottuk a hangsúly kiosztás címkéit, azaz azokra a szavakra tettünk hangsúlyt (F0 csúcsot), amelyek az eredeti mondatban hangsúlytalan (N) címkével voltak ellátva. Mindhárom forma tanulmányozható az adatbázisban. Példákat az 1. táblázat tartalmaz.

1. táblázat: Példák az adatbázis elemeire

A) jó hangsúlyozás	[:N]a [:W]híradások[:W]annak[:N]idején [:W]röviden [:N]számoltak [:N]be [:N]az [:N]ügyről
B) neutrális forma	[:N]a [:N]híradások[:N]annak[:N]idején [:N]röviden [:N]számoltak [:N]be [:N]az [:N]ügyről
C) rossz hangsúlyozás	[:N]a [:N]híradások[:N]annak[:W]idején [:N]röviden [:W]számoltak [:W]be [:N]az [:W]ügyről

Fontos tudni, hogy a fenti három változatban a hangzó mondat szegmentális szerkezete (hangidőtartamok, szünetek) ugyanaz, csak az alapfrekvencia csúcsok meglétében/hiányában, illetve a helyében (melyik szón van) különböznek a mondatok fizikai megvalósításai. Az ilyen célzott F0 változtatásokat a Profivox beszédszintetizátor speciális alkalmazási lehetősége biztosította [4]. Az összes mondat mindhárom változatát kézi feldolgozással készítettük el.

Az adatbázis webes lekérdező felülettel is rendelkezik. A honlap (http://magyarbeszed.tmit.bme.hu/hangsuly) minden funkciója 2014 januárjától érhető el.

2. A hangsúlyadatbázis szerkezete

A hangsúlyadatbázis fő komponense egy MySQL adatbázis, amely az 1869 mondat három különböző hangsúlyozással címkézett szöveges formáját tartalmazza. Az SQL adatbázist WAV és PNG állományok egészítik ki, amelyek a mondatok meghallgatását és a képek megjelenítését teszik lehetővé. A hangsúlyadatbázis utolsó komponense a PHP/HTML forráskódú állományok gyűjteménye, amely a honlap oldalainak megjelenítéséért és a keresés megvalósításáért felelős.

3. A honlap felépítése

Az Első magyar hangsúlyadatbázis az interneten bárki számára hozzáférhető, használatához egy böngésző szükséges. Az adatbázisból a keresőgépekhez hasonló, könnyen kezelhető felületen keresztül kapjuk meg a helyesen hangsúlyozott mondatok listáját, de a honlap sok egyéb funkciót is tartalmaz.

A honlap funkcionálisan 4 részből áll: 1) keresés az adatbázisban, 2) mondatok listája, 3) leírás a kutatható adatbázisról és 4) segítség a honlap használatához. Az adatbázis keresési lehetőségeiről a következő fejezetben részletesen beszámolunk. A honlap második, mondatokat listázó része az adatbázis 1869 mondatáról ad teljes áttekintést: az összes mondat (egyfajta) helyesen hangsúlyozott listáját mutatja meg. A honlap harmadik részében a kutatásra ingyenesen elérhető adatbázisról kapunk tájékoztatást. A honlap használatáról – hangsúlyosan a keresőfelület működéséről – az utolsó, 4-es pontban találunk információt. A következőkben csak a keresési lehetőségeket mutatjuk be.

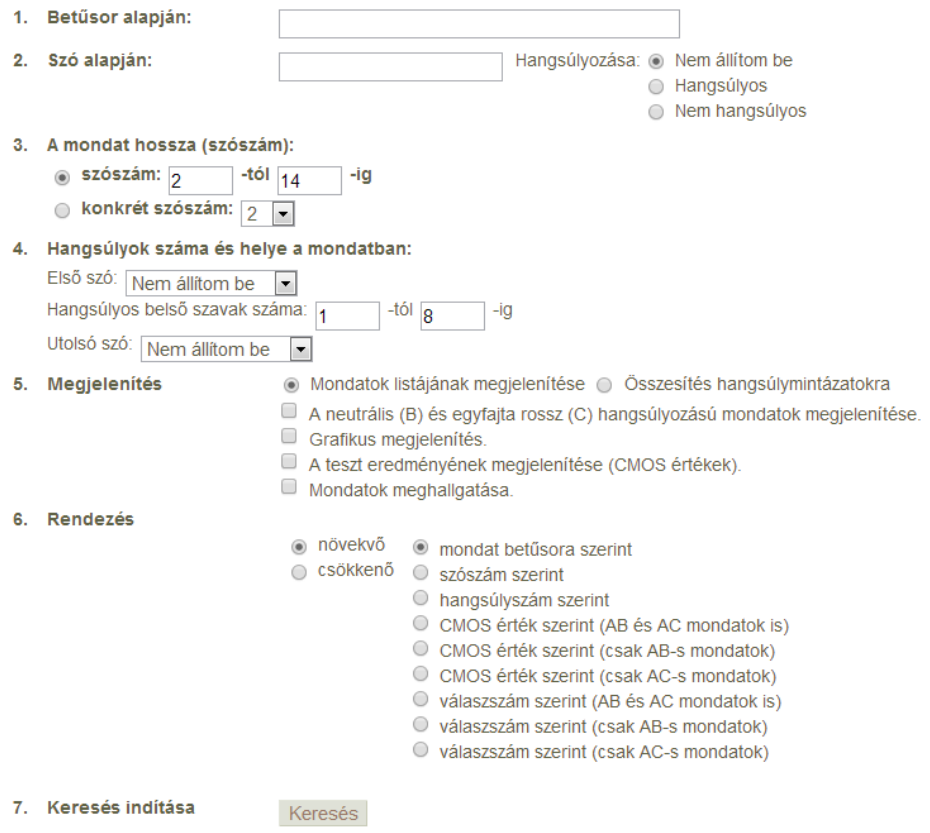
3.1. Keresés az adatbázisban

A hangsúlyadatbázis webes keresőfelületét az 1. ábra mutatja be. A hét részre tagolt felhasználói felület első négy pontja az 1869 mondat szűrésére, azaz a találati lista több szempontú szűkítésére használható, míg az 5. pontban a találati lista megjelenési módjaiból választhatunk. A 6. pontban a találati lista rendezettségét állíthatjuk be, a 7.

pontban pedig a keresést indíthatjuk el. A következő hét alfejezetben az 1. ábra hét pontját mutatjuk be részletesen.

3.1.1. Keresés betűsor alapján

Tetszőleges karaktersorozat megadásával az 1869 mondat ortografikus karaktereiben végezhetünk keresést. Rákereshetünk egy korábban vizsgált teljes mondatra pl. az *a világosság felé fordult, és belebámult az üveg papírnyomóba* keresőkérdésre egyetlen mondatot fog tartalmazni a találati lista (*a **világosság** felé fordult, és **belebámult** az **üveg** papírnyomóba*). A *világ* keresőkérdésre egy 27 elemű találati lista a válasz, amely a fenti mondatot éppúgy tartalmazza, mint pl. az *a **mai** világban **nem** sikk betegeskedni* mondatot is. A keresés során a keresési mezőbe gépelt karaktersorozatokat tehát úgy értelmezzük, hogy azt tetszőleges karaktersorozat előzheti meg vagy követheti.



1. ábra. Az Első magyar hangsúlyadatbázis honlapjának keresőfelülete

3.1.2. Keresés szó alapján

Ebben a pontban tetszőleges szó előfordulására kereshetünk rá, miközben a szó hangsúlyhelyzetét is beállíthatjuk. Választhatunk hangsúlyos és hangsúlytalan előfordulások között, valamint dönthetünk úgy, hogy nem vesszük be a szűrőfeltételbe ezt az opciót. Például a *több* szó hangsúlytalan pozícióban 3 mondatban fordul elő (az egyik mondat: *ez pedig* ***nem*** *több egy* ***közepes nyugati*** *egyetem költségvetésénél*), míg hangsúlyos helyzetben 10 találat jelenik meg (az egyik ezek közül: ***öt*** *nap alatt* ***több*** *mint* ***kilencven*** *órát dolgozott*). A keresés a keresőmezőbe írt karaktersorozat pontos előfordulásán alapul.

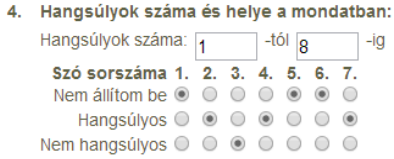
3.1.3. Keresés a mondat hossza alapján

Az adatbázisban előforduló szavak száma 2 és 14 között változik. Ebben a pontban a szavak száma alapján szűkíthetjük a találati listát. Erre két módunk van. A tartomány alapú mondathossz-beállítás a legkisebb és legnagyobb szószám megadását követeli meg. A másik lehetőség konkrét szószám megadása. Ebben az esetben csak az itt beállított szószámmal rendelkező mondatok jelennek meg a találati listában.

3.1.4. Keresés a hangsúlyok száma és helye alapján

A hangsúlyok számának és helyének beállítása az előző pontban specifikált szószám megadásától függ. Amennyiben ez tartomány alapú, akkor a hangsúlyok helyét a mondat három pozíciójában, a mondat első és utolsó szavában, illetve a mondat belsejében állíthatjuk be. Külön dönthetünk tehát az első és utolsó szó hangsúlyos vagy hangsúlytalan pozíciójáról, illetve a mondatbelseji hangsúlyos szavak számáról. Ez utóbbi egy intervallum megadásával lehetséges. Az 1. ábra 3. pontja ezt a hangsúlymegadási formát mutatja. Tegyük fel, hogy a 3. pontban a szószám intervalluma 2–5, a 4. pontban pedig az első és az utolsó szó is hangsúlyos és a hangsúlyos belső szavak tartománya 1–1. A 8 elemű találati lista ekkor tartalmazza a ***szeretném****, ha* ***néhány*** *percre* ***elfordulna*** és az ***üresre*** *facsarunk, aztán* ***megtöltünk önmagunkkal*** mondatokat is.

Ha a 3.1.3. pontban konkrét szószámot állítunk be, akkor a hangsúlyok pontos, szavankénti megadására is lehetőségünk van. A 2. ábra a keresési felület 4. pontját emeli ki hét szavas szószám megadás esetén. A hangsúlyok számát a teljes mondatra specifikálhatjuk egy intervallum megadásával. Mivel az adatbázis összes mondatára vonatkozóan a hangsúlyszámok 1 és 8 között változnak, a 2. ábrán látható beállítás nem jelent szűkítést. Az újdonság az ábra további részében figyelhető meg. A mondat mind a hét szavára beállíthatjuk a hangsúlyos vagy hangsúlytalan pozíciót, illetve eltekinthetünk az opció beállításától. A 2. ábra alapján a találati listában csak azok a mondatok jelennek meg, amelyekben a 2., 4. és 7. szó hangsúlyos, a 3. pedig nem hangsúlyos. A többi szó hangsúlypozíciója tetszőleges lehet. A találati lista 7 mondatot tartalmaz (két példa a listából: *a* ***szegénység*** *és* ***betegség*** *együtt járása* ***közismert***, illetve ***ehhez tegnap*** *a* ***tőzsdetanács hozzá*** *is* ***járult***).



2. ábra. A keresési felület 4. pontja konkrét szószám megadása esetén

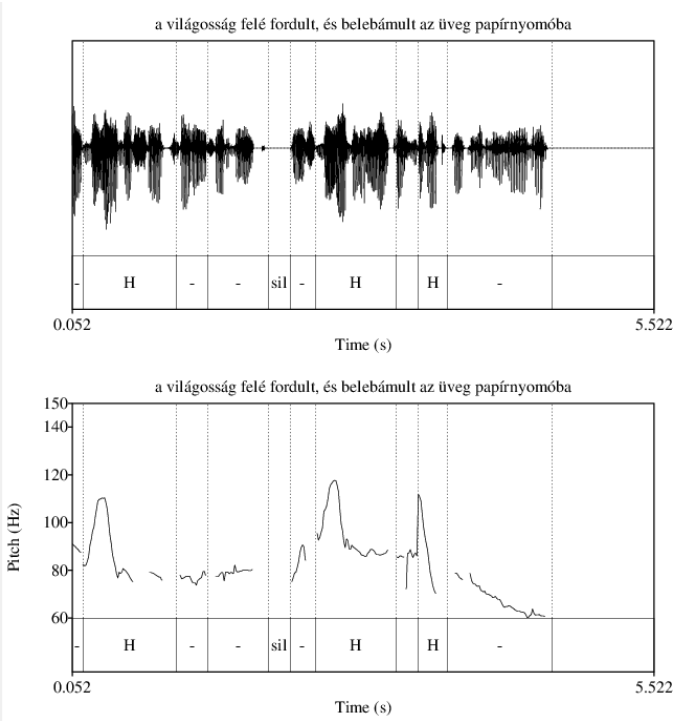
3.1.5. A megjelenítés beállításai

A találati lista az előző négy pont keresőfeltételei alapján áll össze. Alapesetben a találati lista elemei mondatok, melyek megjelenését ebben a pontban határozhatjuk meg (l. 1. ábra 5. pont). Ez az alapértelmezett `Mondatok listájának megjelenítése` opció választásával kezdeményezhető. A másik lehetőség a találatok megjelenítésére, hogy a hangsúlymintázatokra összesítve kérjük a mondatokat (`Összesítés hangsúlymintázatokra` opció). Ez utóbbi lehetőséget az alfejezet végén tárgyaljuk.

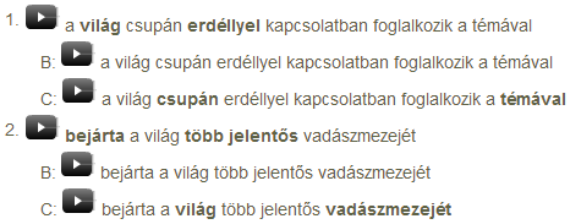
A mondatok listája alapértelmezetten a keresés során kiválogatott helyes hangsúlyozású mondatokat tartalmazza (a hangsúlyadatbázisból). Azonban kezdeményezhetjük ugyanazon mondatra a neutrális (B) és egyfajta rossz hangsúlyozású (C) változat megjelenítését is, melyek a helyes hangsúlyozású mondat alatt a B és C pontok után jelennek meg a találati listában (4. ábra).

A grafikus megjelenítési opció segítségével a mondat szerkezetének és a hangsúly kiosztásnak az összefüggéseit lehet tanulmányozni a hangsúlyozás fizikai megvalósulásának tükrében. A találati lista minden eleme ekkor tartalmazza rezgésképet és az alapfrekvencia görbét szinkron megjelenítésben, szóhatárokkal és a hangsúlycímkékkel kiegészítve (3. ábra). A `H` címke a hangsúlyos, a `-` címke a hangsúlytalan szavakat jelöli. A mondatbelseji szüneteket a `sil` karaktersorozattal jelöltük.

A mondatok mindhárom típusa meghallgathatóvá tehető a megfelelő jelölőnégyzet bekapcsolásával (`Mondatok meghallgatása`). A 4. ábrán a *világ* keresőszóra adott kételemű találati lista látható. A megjelenítési opciókból az `A neutrális (B) és egyfajta rossz (C) hangsúlyozású mondatok megjelenítése` és a `Mondatok meghallgatása` jelölőnégyzeteket kapcsoltuk be. A lejátszás gombra kattintva hallgathatjuk meg a megfelelő mondatokat.



3. ábra. Grafikus forma a találati lista egy elemére



4. ábra. A *„világ”* szókeresésre adott találati lista mindhárom mondattípus és a meghallgatási opció beállítása után

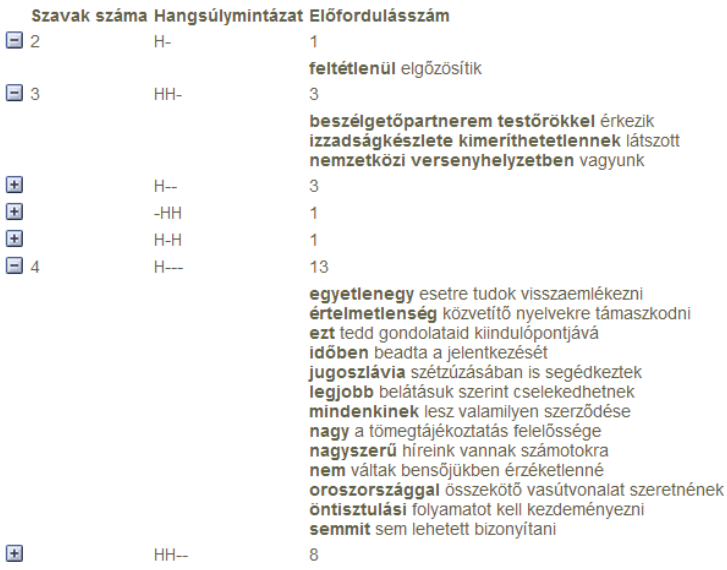
A hangsúlyadatbázis nyilvános tesztelésének eredményeit [3] itt is közreadjuk (több száz tesztelő hallgatott meg 40-40 mondatpárt). Az eredményeket az `A teszt eredményének megjelenítése (CMOS értékek)` opció kiválasztásával jeleníthetjük meg. A teszt során az A-típusú (jó hangsúlyozásúnak tartott) mondatokat kellett összevetniük a tesztelőknek vagy a B (neutrális), vagy a C (rossz hangsúlyozású) ugyanazon mondattal egy-egy mondatpárt meghallgatva. Például az *olasz* keresőszóra adott találati lista egyetlen mondatot jelenít meg (***olasz*** *klub csak* ***elvétve*** *igazolt akkoriban magyar labdarúgót*), amely a fenti opció beállítása után táblázatos formában tartalmazza a teszt eredményeit is erre a mondatra. A megjelenő CMOS értékeket [1] kiemeltük az 1. táblázatba.

1. táblázat: Az adatbázis egy mondatának átlagos CMOS értékei

CMOS	CMOS AB	CMOS AC
0,86 (N=7)	0,75 (N=4)	1 (N=3)

Hogyan kell értelmezni a CMOS adatokat? A vizsgált mondatra adott CMOS pontszámok 1, 0 vagy –1 értéket vehetnek fel, és a tesztalany döntésén alapulnak. Az A-típusú mondatra vonatkozó kedvező ítélet esetén 1 értéket kap a mondat. A (B) vagy a (C) mondatra vonatkozó döntés esetén pedig –1 értéket rögzítünk. Más esetekben (ha mindkettőt egyformának tartja, tehát nem dönt egyik mellett sem), akkor 0 értéket adunk. Az 1. táblázat több tesztelőre vonatkozó, átlagos CMOS értékeket tartalmaz. Az AB mondatok meghallgatása során az A-típusra vonatkozó preferenciát a CMOS AB pontszám tartalmazza (0,75). A zárójelben lévő 4-es érték a minta elemszámát jelenti, vagyis összesen 4 tesztalany találkozott (mindegyik egyszer) a fenti mondattal. Három tesztalany az A-típusú mondatot részesítette előnyben, egy pedig egyformának ítélte a B-típusú mondattal. Azaz 0,75=(1+1+1+0)/4. Az ugyanezen AC mondatok esetén mindhárom tesztalany az A-típusú mondatokat preferálta, mivel a CMOS AC értek 1. Az 1. táblázat CMOS oszlopában az összesített, az AB és AC mondatokra egyaránt vonatkozó ítéletek átlaga szerepel (0,86). A fenti mondattal összesen 7 tesztalany találkozott.

A találatok eddigi formájától jelentősen eltérő megjelenítést kapunk, ha az `Összesítés hangsúlymintázatokra` opciót választjuk. Hangsúlymintázatnak nevezzük a mondat szavaira vonatkozó hangsúlyjelek sorozatát balról jobbra értelmezve. Ha a mondatot a hangsúlymintázatával jellemezzük, akkor annyi jel van a hangsúlymintázatban, ahány szó van a mondatban (a névelők is szónak számítanak). Egy két szavas mondat hangsúly mintázata például a `H-` képpel fejezhető ki, de lehet `-H` is, vagy `HH` is. Ezen opciónál a megjelenítés egyéb beállításait és a következő pontban szereplő rendezési szempontokat is figyelmen kívül hagyjuk. Az 1–4. pontokban szereplő keresési feltételeknek megfelelő mondatokat a szavak száma és a hangsúlymintázatok szerint csoportosítjuk. A találati lista így ezen csoportosítás mellett a csoportok elemszámát és kérésre, a csoportba tartozó mondatokat tartalmazza (5. ábra).



5. ábra. Egy találati lista része, amikor összesítést kértünk a hangsúlymintázatokra

3.1.6. A találati lista rendezése

A találati lista 9 szempont szerint rendezhető. A rendezés iránya minden esetben lehet növekvő és csökkenő is. Rendezhetünk a mondatok jellemzői alapján: ábécé szerint, szószám vagy hangsúlyszám szerint. A nyilvános teszt eredményeit is felhasználhatjuk, így a CMOS értékek és a válaszok száma alapján is rendezhetünk. Ez utóbbi két szempont esetén választhatunk az összesített vagy külön az AB és AC mondatpárokra vonatkozó CMOS értékek és teszt válaszszámok között.

3.1.7. Keresés indítása

A `Keresés` gomb megnyomásával kezdeményezhetjük a találati lista megjelenítését, amely a keresési felület alatt jelenik meg. A találati elemeinek összetevőit a 3.1.5. pontban részleteztük.

4. Összefoglalás

Jelen cikkben az első magyar hangsúlyadatbázisra alapozott webalapú felhasználói felület felépítését és használatát mutattuk be. A hangsúlycímkékkel ellátott mondatok sok szempont alapján lekérdezhetők, a találatként megjelenő mondatokhoz pedig többféle ábrázolást választhatunk. A lekérdezése széles tárháza jó adatbányászati, elemzési alapot nyújt a kutatóknak. Az elkészült hangsúlyadatbázis alkalmas külön-

böző hangsúlyjelölő algoritmusok tesztelésére, a beszédszintézis továbbfejlesztésére. Segítheti a gépi beszédfelismerést, használható az oktatásban és általában is új lendületet adhat a hangsúlykutatásokhoz, a hangsúly és a mondatszerkezet viszonyának vizsgálatához.

Támogatás: Az adatbázis létrehozását a Paelife (Grant No. AAl-08-01-2011-0001) és az EITKIC_12-1-2012-001 projektek támogatták

Hivatkozások

1. ITU-T: P.800 Methods for subjective determination of transmission quality (1996)
2. Olaszy G.: Precíziós, párhuzamos, magyar beszédadatbázis fejlesztése és szolgáltatásai. In: Gósy M. (szerk.): Beszédkutatás. MTA Nyelvtudományi Intézet, Budapest (2013) 261–270
3. Olaszy G., Abari K., Bartalis M. In: Gósy M. (szerk.): Beszédkutatás 2014. MTA Nyelvtudományi Intézet, Budapest (2014) Megjelenés alatt.
4. Olaszy, G., Németh, G., Kiss, G.: Hungarian audiovisual prosody composer and TTS development tool. In: Puppel S., Demenko, G. (szerk.): Prosody 2000. Poznan, Poland (2001) 167-178
5. Vicsi K., Víg A.: Az első magyar nyelvű beszédadatbázis. In: Gósy, M. (szerk.): Beszédkutatás 1998 MTA Nyelvtudományi Intézet, Budapest (1998) 163–177

Dokumentumkollekciók vizualizálása kulcsszavak segítségével

Berend Gábor, Erdős Zoltán, Farkas Richárd

Szegedi Tudományegyetem,
TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.,
e-mail:berendg@inf.u-szeged.hu, erdos.zoltan91@gmail.com, rfarkas@inf.u-szeged.hu

Kivonat A dokumentumkollekciókban történő eligazodás kapcsán hasznos segítséget képesek nyújtani a különféle intelligens adatvizualizációs eljárások. Egy lehetőség, ha a dokumentumkollekció alkotóelemeire mint irányítatlan, súlyozott gráf csúcsaira tekintünk, a köztük lévő kapcsolatok erősségeit pedig a dokumentumpárokat jellemző hasonlóságértékek adják, majd az előzőek szerint definiált gráfot jelenítjük meg valamilyen gráfrajzoló eljárás segítségével.

Az efféle megközelítések alkalmazása során azonban – különösen nagy méretű adatbázisok esetében – gyakorlati nehézségekbe ütközhetünk. Nagy számú csúccsal és éllel rendelkező gráfok esetében nehézkessé válhat azok áttekinthetősége, valamint a csúcsok koordinátáinak meghatározásáért felelős optimalizációs számítások konvergenciája is lassú lehet.

Az általunk megvalósított alkalmazás – korábbi munkáinkra [1,2] is építkezve – dokumentumkollekciók vizualizációját hajtja végre azok kulcsszavaira támaszkodva. Előnye, hogy a vizualizációs szempontból nehézséget jelentő méretű korpuszok megjelenítését is lehetővé teszi azáltal, hogy a dokumentumok hierarchikus klaszterekbe történő besorolásának elvégzése után bizonyos csomópontokat összevonva ábrázol. A tematikus dokumentumegyüttesek aprólékos felépítésének megismerésére pedig felhasználói interakció útján nyílik lehetőség.

A teljes dokumentumgráf megjelenítésével kapcsolatos nehézségek olyan módon kerültek tehát áthidalásra, hogy az alkalmazás inicializálása során a dokumentumok kulcsszavaik alapján történő klaszterezéseként előálló főbb témák – melyek száma jellemzően jóval elmarad a dokumentumok számától – kirajzolása történik meg. A klaszterezés végrehajtása egy különösen jól skálázódó algoritmus segítségével [3] történik, amivel akár százezres nagyságrendű dokumentumkollekciók klaszterezése is megoldható az alkalmazás inicializálása során. A vizualizálandó korpusz klasztereinek feldolgozását megkönnyítendő, a dokumentumklasztereket összegző, azokat a többi klasztertől megkülönböztető kulcsszavak kiválasztása és megjelenítése történik meg információelméleti megfontolások mentén.

Demóalkalmazásunk a Magyar Nemzeti Szövegtárban található újságcikkek vizualizációját hajtja végre. Amiatt ugyanakkor, hogy az alkalmazás bemenetéül egy egyszerű – a megjelenítendő dokumentumok kulcsszavait tartalmazó – szöveges állomány szolgál, így adaptálása más jellegű

szövegekre könnyen végrehajtható. Természetesen a bemeneti állomány a kulcsszavak mellett tartalmazhat egyéb adatokat is (pl. dokumentumközi hivatkozással kapcsolatos információkat), így ezek beépítése sem okozna nehézséget a vizualizációs eljárásba.

Kulcsszavak: automatikus kulcsszókinyerés, dokumentumvizualizáció

Köszönetnyilvánítás

Berend Gábor publikációt megalapozó kutatása a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosítószámú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése országos program című kiemelt projekt keretében zajlott. A projekt az Európai Unió és az Európai Szociális Alap társfinanszírozásával valósult meg. A további szerzőket a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta.

Hivatkozások

1. Berend, G., Vincze, V., Farkas, R., Zsibrita, J., Jelasity, M.: Kulcsszókinyerés alapú dokumentumklaszterezés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 251–262
2. Farkas, R., Berend, G., Hegedűs, I., Kárpáti, A., Krich, B.: Automatic free-text-tagging of online news archives. In: Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, Amsterdam, The Netherlands, The Netherlands, IOS Press (2010) 529–534
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008+

Információkinyerés magyar nyelvű önéletrajzokból a nexum Karrierportálhoz

Farkas Richárd[1], Dobó András[3], Kurai Zoltán[3], Miklós István[1], Miszori Attila[3], Nagy Ágoston[1], Vincze Veronika[2], Zsibrita János[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3] nexum Magyarország Kft.

Kapcsolat: rfarkas@inf.u-szeged.hu

Kivonat

Számos nagyvállalatnak komoly problémát jelent az alkalmazottak toborzása. A legnagyobb gondot az jelenti, hogy akkora mennyiségben jelentkeznek ezekhez a cégekhez egy adott kiírásra, hogy nincs arra elegendő emberi erőforrás, hogy a rengeteg, sokszor több ezernyi beérkezett önéletrajzot egyesével végignézzék. Ezért a bevett szokás az, hogy az adatbázisban lévő önéletrajzok közül tulajdonképpen véletlenszerűen választanak pozíciótól függően néhány tízet vagy néhány százat, mert csak ezek végigolvasására van idejük.

A Szegedi Tudományegyetem Nyelvtechnológiai Csoportja és a nexum Magyarország kft. közös kutatás-fejlesztési projektje során egy olyan módszer kifejlesztésén dolgozunk, mely egy adott álláslehetőséghez megadott lekérdezést és egy önéletrajzhalmazt inputként kapva visszaadja az önéletrajzok rendezett sorozatát az illetők adott pozícióra való alkalmassága alapján. Mivel a gyakorlatban az látszik, hogy minden önéletrajz egyedileg szerkesztett, mindegyik más és más struktúrájú, ezért az önéletrajzok megfelelőségük szerint közvetlenül nem rangsorolhatók. Ahhoz, hogy ez megvalósítható lehessen, szükség van arra, hogy a munkavállalók adatait egy egységes adatstruktúrába ki tudjuk nyerni az önéletrajzokból.

Ezért első lépésként egy olyan módszert fejlesztettünk ki, mely alkalmas arra, hogy egy tetszőleges önéletrajzból a munkavállaló legfontosabb adatait, mint például a nevét, a születési dátumát, az elérhetőségeit, a tanulmányi adatait, a munkatapasztalatait, a nyelvismeretét és további lényeges adatait kinyerje. A legtöbb nagyvállalat karrierportáljában az önéletrajz beadása mellett a munkakeresőknek egy űrlapot is ki kell tölteniük, melyen az önéletrajzi adataikat strukturáltan megadják. Mivel az algoritmusunk alkalmas arra, hogy ezeket az adatokat az önéletrajzból kinyerje, ezért az űrlapkitöltési folyamat automatizálásában is fel lehetne használni módszerünket úgy, hogy a munkakeresőnek az adatait csak ellenőriznie kelljen, és ne neki kelljen mindent manuálisan bevinnie az űrlapba.

Ennek a feladatnak az első problémáját az okozta, hogy a munkavállalók önéletrajzaikat rendszerint különböző fájlformátumban küldik be a nagyvállalatok karrierportálján keresztül. Hogy algoritmusunk egységes formátumú önéletrajzokat kapjon bemenetként, a különböző formátumokból először egységesen PDF formátumot ké-

szítettünk, majd a PDF formátumú önéletrajzokból egyszerű szöveges dokumentumokat gyártottunk. Ezek a szöveges dokumentumok a formázásokat ugyan mellőzik, de a dokumentumok elrendezését megtartják.

A feladat ezek után a célinformáció kinyerése a szöveges (de strukturált) állományokból. Kézzel annotált tanító dokumentumok hiányában azt a megoldást láttuk kézenfekvőnek, hogy megpróbálunk egy olyan módszert kidolgozni, mellyel tanító adatok automatikusan generálhatók. Ez az álláskeresők által a karrierportálon az önéletrajz feltöltése mellett kitöltött űrlapok alapján megvalósítható. Mivel az adatok az űrlapon strukturáltan kerültek felvitelre, és elméletileg ugyanazok az adatok szerepelnek az önéletrajzban is, ezért az űrlap adatainak önéletrajzra való mappelésével automatikus tanító önéletrajzokat kaphatunk.

Rendszerünk az önéletrajzok előfeldolgozása – amely magában foglalja a szöveg normalizálását és a dokumentumok struktúrájának egy belső fareprezentációba történő illesztését – után automatikusan tanító adatokat generál az önéletrajz-űrlap párosokból. Habár kezdetben ez viszonylag egyszerű feladatnak tűnt, számos problémával kellett szembenéznünk. Először is, rengeteg önéletrajz-formátum, -struktúra fordul elő a beadott önéletrajzok között, és sok egyáltalán nincs is vagy csak alig van strukturálva. Másodszor, bár úgy gondoltuk, hogy a feltöltött önéletrajz és a vele egy időben kitöltött űrlap adatai megegyeznek, valójában sok helyen különböznek, egyes adatok a két helyen különböző formában szerepelnek, illetve sok adat pusztán az egyik helyen szerepel. Ezen kívül a rengeteg elgépelés is nagyban nehezíti a munkát. Megoldásként az adatokat próbáltuk normalizálni, a felismerésben különféle mintákat használtunk, és a különböző adatosztályokhoz külön annotátorfüggvényeket készítettünk. Az így automatikusan generált tanító adatok mellett kézzel annotált dokumentumokat is felhasználtunk a tanításban.

A tanító adatok elkészítése után egy MEMM szekvenciajelölő modellt tanítunk [1], melyhez számos különféle jellemzőt definiáltunk, többek közt különféle reguláris kifejezéseket, listákat, szóalaki jellemzőket, mondat- és szövegbeli elhelyezkedést és a dokumentum struktúrájában elfoglalt pozíciót, kézzel gyűjtött doméntaxonómiákat stb. Az így tanított modell egy még annotálatlan önéletrajzot megkapva képes az önéletrajzban található fontosabb adatok jó minőségű kinyerésére. Természetesen a kinyerés minősége nagyban függ a kapott önéletrajz strukturáltságától és minőségétől is.

Demónkban lehetőség nyílik a rendszer megismerésére éles működés közben, továbbá a fejlesztés során megoldott gyakorlati nyelvtechnológiai problémák megvitatására.

Hivatkozások

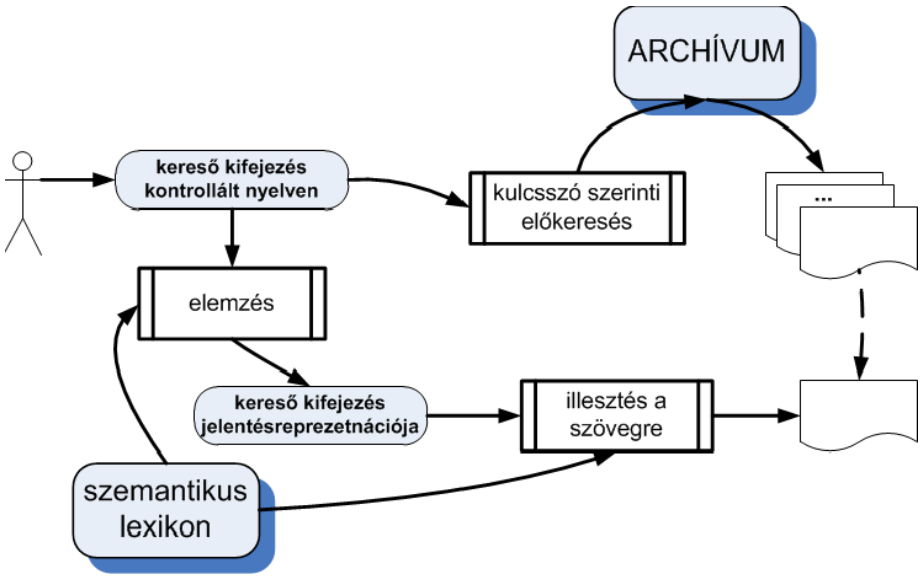
1. McCallum, Andrew, Freitag, Dayne, Pereira, Fernando: Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. ICML (2000) 591–598

MASZEKER: szemantikus keresőprogram

Hussami Péter[1]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy j. u. 7
hussami@all.hu

Az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjával, valamint Könyvtár- és Humán Információtudományi Tanszékével közösen fejlesztette a MaSzeKer magyar szemantikus keresőt (projektazonosító: TECH_08_A2/2-2008-0092). A projekt eredményeképpen olyan szoftvermegoldás született meg, amely különböző műfajú forrásokat szemantikus reprezentációs alakra konvertál, és ezekben a reprezentációkban keres. Ennek előnye, hogy a rendszer így morfológiailag eltérő, de szemantikailag megegyező tartalmakat is össze tud kapcsolni. A rendszer áttekintő architektúrája az 1. ábrán látható.



1. ábra: A MASZEKER rendszer áttekintő architektúrája

Az ábrának megfelelően a releváns dokumentumok keresése a következő lépésekből áll:

1. a felhasználó egy kontrollált nyelven adja meg a kereső kifejezést,
2. szintaktikus és szemantikus elemzés előállítja kereső kifejezés jelentésreprezentációját,
3. szavak szerinti keresés előszűri az archívumot,
4. azokra a szöveg szegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a kereső kifejezés jelentésreprezentációját.

Az MSzNy VIII konferencián tartott előadáson ([1]) ismertették a fenti elemek megvalósítására vonatkozó elméleti alapelveket, elsősorban egy módszert ajánlva a folyamat szemantikus reprezentáció fölé építésére. Az MSzNy VIII-IX konferenciákon tartott bemutatókon ([2, 3]) a rendszer első változatát mutattuk be, amely főnévi csoportos, illetve teljes (mondat) keresőkifejezéseken működött. Ezek a korábbi verziók szabadalmi igénypontokra lettek tervezve – a rendszer ennek megfelelően ezekre lett optimalizálva.

A jelen demó egy viszonylag eltérő nyelvezetű dokumentumhalmazon is működik, nevezetesen angol nyelvű klinikai ajánlásokon. Ezek, a szabadalmi igénypontokkal szemben lényegesen több segédigét / deontikus operátort tartalmaznak, és szabadabb szórendet használnak. Az ehhez tartozó szemantikus reprezentációt shallow parsing elv alapján dolgozzuk fel ([4]).

A demo használata: a felhasználó a kontrollált nyelven adhat meg keresőkifejezést. A keresőkifejezés több mondatból ill. főnévi kifejezésből állhat, egy kontrollált angol nyelven megfogalmazva. A megszorítások az egyértelműséget biztosítják, a tipikusan nehezen egyértelműsíthető fordulatokat akartuk kizárni. Legfontosabb korlátozások (a teljes definíció [5]-ban hozzáférhető):

- csak kijelentő mód, jelenidejű mondat írható,
- titott a mellérendelő mellékmondat (viszont a mondatok AND, OR kapcsolóval kapcsolhatóak, zárójelezhetőek),
- tiltott az alárendelő mellékmondat bármiféle lerövidítése (pl. igeneves utómódosítók),
- az alárendelő mellékmondatnak a „which” vonatkozó névmással kell kezdődnie, és ennek a közvetlenül megelőző főnévi csoport fejére kell vonatkoznia,
- tiltottak az igeneves előmódosítók,
- felsorolás, koordináció csak főnévi csoportok közt megengedett, ezeket a felhasználónak jelölnie kell.

A felhasználói interfész segíti a kontrollált nyelv szabályainak betartását, és a morfoszintaktikai elemzés eredménye alapján a rendszer ellenőrzi a szabályok betartását. Mivel teljes szabályrendszer nem ellenőrizhető, a generált jelentésreprezentációt grafikusan bemutattatik – ha szükséges, a felhasználó módosíthatja a keresőkifejezést. Ez a megjelenítés segíti egyértelműsíteni az egyes szavak jelentésének megállapítását is, ha több frame/synset van egy csomóponthoz rendelve, a felhasználó választhatja a megfelelőt.

A rendszer a keresőkifejezéshez illő frázisokat keres az igénypontok szövegében, és az eredményt a grafikus interfészen megmutatja, kiemelve azokat a szavakat, amelyekből álló frázist a keresőkifejezés egy szegmenséhez hasonlónak talált. Míg a keresőkifejezés feldolgozásánál maximálisan törekszünk a pontos jelentésreprezentációra, a keresés fázisában az aktuális szövegrészlet vizsgálatánál csak azt ellenőrizzük, hogy jelentheti-e a keresőkifejezés valamely frázisát.

Hivatkozások

1. Szőts M., Csirik J., Gergely T., Karvalics L.: MASZEKER: projekt szemantikus kereső technológia kidolgozására. In: Tanács A., Vincze V. (szerk.): MSzNy 2010 – VII Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudomány Egyetem, Szeged (2010) 159–167
2. Hussami P.: MASZEKER: szemantikus kereső program. In: Tanács A., Vincze V. (szerk.) MszNy 2011 – VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged (2011) 321–322
3. Hussami P.: MASZEKER: szemantikus kereső program. In: Tanács A., Vincze V. (szerk.) MszNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem, Szeged (2013) 302–304
4. Gyarmathy Zs., Simonyi A., Szőts M.: Felszíni szintaktikai elemzés és a jóindulatú interpretáció elve információ-visszakeresésben. In: MszNy X – X. Magyar Számítógépes Nyelvészeti Konferencia, megjelenés előtt
5. A kontrollált nyelv definíciója http://www.maszeker.hu/?page=download

ReALIS1.1

Nőthig László, Alberti Gábor, Dóla Mónika

PTE BTK Nyelvtudományi Tanszék
ReALIS Elméleti, Számítógépes és Kognitív Nyelvészeti Kutatócsoport
nothig.laszlo@gmail.com,
{alberti.gabor,dola.monika}@pte.hu

Kivonat: A laptopos bemutatásra szánt ReALIS1.1 program elsősorban nyelvészek (mint „belső felhasználók") számára hivatott eszköztárat adni olyan nyelvfragmentumok építésére, amelyek jól ragadják meg a természetes nyelvek sajátosságait, elsősorban a kompozicionális jelentésösszegződést. A definiálható jelentések olyan pragmatikai-szemantikai leírások, amelyek megfelelnek a (reprezentacionalista dinamikus diskurzusszemantikák családjába tartozó) ReALIS releváns definícióinak. A felépített nyelvet alkalmazókat külső felhasználóként határozhatjuk meg. Lényegében egy sajátosan megsokszorozott adatbázist kapnak, ami a való világ modellje mellett annak alternatíváit is felkínálja. A ReALIS alapállása szerint ezek a formális szemantikából ismerhető „lehetséges világok" mindig odaköthetőek a világmodellben jelen lévő humán ágensekhez mint azok (tév-) hiedelmei, vágyai, szándékai, álmai. A külső felhasználó a program használata során (lépésről lépésre) lexikai egységeket kap választásra, ezekből mondatokat építhet, a felépített mondatoknak pedig megkapja az igazságértékelését egy általa kiválasztott vagy feltöltött világmodell alapján. Az igazságértékelést olyan „konstruktivista" módon kibővítve értjük, hogy a program az „igaz" válaszon túl megadja mindazt az információt, ami alátámasztja e választ. Nemcsak a nyelvleírás „próbára tételét" szolgálhatja tehát a program, hanem adatgyűjtésre és -rendszerezésre is használható.

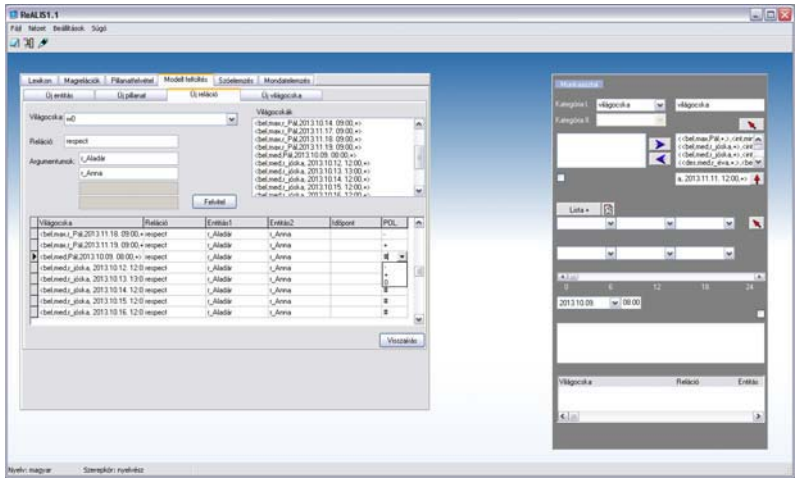
1. Felhasználók, felhasználások

1.1. A laptopos bemutatásra szánt ReALIS1.1 program elsődlegesen *nyelvészek* számára hivatott eszköztárat adni olyan (tetszőleges nyelvű) nyelvfragmentumok építésére, amelyek jól ragadják meg a természetes nyelvek sajátosságait, elsősorban a kompozicionális jelentésösszegződést [10]. A definiálható jelentések olyan pragmatikai-szemantikai leírások, amelyek megfelelnek a (reprezentacionalista dinamikus diskurzusszemantikák [9] [11] családjába tartozó) ReALIS releváns definícióinak [2]. A ReALIS egy olyan modell, amely a generatív szintaxiselméletek formális pontosságát [1] egyesíti az optimalistáselméletek és a DRT dinamikus megközelítésével [3], figyelembe véve a kognitív nyelvészek holisztikus nézőpontját is [8].

Belső felhasználónak fogjuk nevezni az 1.1-ben meghatározott felhasználói kört.

1.2. A felépített nyelvet alkalmazókat *külső felhasználóként* határozhatjuk meg. A külső felhasználó a program használata során (lépésről lépésre) lexikai egységeket kap választásra, ezekből mondatokat építhet, a felépített mondatoknak pedig megkapja az igazságértékelését egy általa kiválasztott vagy feltöltött [12] világmodell alapján.

Külső felhasználó lehet például egy nyomozó vagy bíró, aki állítások igazságát értékeltetheti. Az igazságértékelést olyan „konstruktivista" módon kibővítve értjük, hogy a program az „igaz" válaszon túl megadja mindazt az információt, ami alátámasztja e választ. A külső felhasználó tehát adatgyűjtésre is használhatja a programot.



2. A belső felhasználók számára felkínált használati esetek

2.1. A belső felhasználó definiálhat egy w_0 külvilágot, aminek u_i entitásokból álló univerzumán relációkat adhat meg [10]. A reláció egyik tagja szükségszerűen diszjunkt időintervallumok egy sorozata. A program folyamatos kérdésekkel levezényli a külvilág felépítését: újabb és újabb relációkat kér, egy adott reláció esetében pedig időintervallumokat két végpontjuknál meghatározva – míg a felhasználó nem választja azt az opciót, hogy az adott művelettípust már nem kívánja ismételni. Bármikor vissza lehet kérni egy említett művelettípust módosításra.

Olyan relációkat lehet így definiálni, amelyek homogének abban az értelemben, hogy bármely intervallumon belüli időpontban igazak (pl. *szeret, úszik, utazik,* szemben olyanokkal, mint *megszeret, átúszik, hazautazik*). Egy argumentumhelyhez (egy, vagy akár több) megszorító reláció rendelhető a már korábban definiált relációk köréből. Például az *utazik* cselekvői argumentumához hozzárendelhetjük, hogy *ember*.

2.2. A belső felhasználó az addig definiált világocskákhoz képest újabbat határozhat meg, ahol a w_0 külvilág alkotja e definíció bázisát. Egy w' világocskához képest az (1a) pontban megadott címkesorozattal vagy annak alternatívájával adható meg egy w" világocska.

1. példa. *A világocskák címkézése*

a. ⟨bel,max,$r_{Jóska}$,τ",+⟩
b. ⟨⟨bel,med,$r_{Jóska}$,τ,+⟩, ⟨int,max,r_{Peter},τ',+⟩, ⟨bel,max,$r_{Jóska}$,τ",+⟩⟩
c. +/–/∅/0/~~0~~

Az (1a) például egy emberi lény ($r_{Jóska}$) biztosnak tartott (max) tudását (bel) hordozó világocskát definiál. Alternatívák a bel („hiedelem") címkéhez: int („szándék"), des („vágy") stb. Alternatívák a max címkéhez: kisebb intenzitási fokozatok (pl. med: „közepes"). A címke negyedik tagja egy időpillanat, amely a világocskához rendelt információ származási ideje. Az ötödik tag a polaritás, amelynek értékei a fenti (1c) pontban vannak felsorolva (értelmezésüket később adjuk meg).

A program felhasználói kérésre megadja, hogy egy világocskához milyen definíciós lépéseken keresztül juthatunk el a külvilágtól. Az (1b) pontbeli címkesorozat például olyan információ gyűjtőhelyeként szolgáló világocskát határoz meg, amelyet nyelvileg így ragadhatunk meg: „Jóska úgy sejti, hogy Péter leghatározottabb szándéka őt rávenni arra, hogy biztosra vegye azt, hogy…"

2.3. A belső felhasználó információt rendelhet a világocskákhoz, amit az alábbi módon kell a programnak levezényelnie.

A felhasználónak elsődlegesen egy időpillanatot kell megadnia, amit kiegészíthet egy reláció megadásával, illetve annak egyes argumentumait is specifikálhatja. A program erre kiírja a relációkhoz tartozó időintervallumok alapján, hogy az adott pillanatban mely relációk állnak fenn mely entitások között. Ha a felhasználó megadott egy relációt néhány argumentummal, akkor csak a további argumentumok kiírása a feladat. A kiírás egysége a (külső) infon [15]: egy infon azt az információt jelenti, hogy bizonyos entitások egy bizonyos relációban állnak a megadott pillanatban (pl. Péter éppen szereti Marit, vagy éppen utazik).

A belső felhasználó a fenti módon előállított infonokat (egyesével vagy csoportosan) világocskákhoz rendelheti – nevezzük ezt *pillanatfelvételnek*, majd bármely paraméterüket módosíthatja – a programnak ilyen lehetőségeket kell felajánlania. A külvilággal (1a) relációban álló világocskához hozzárendelni egy infoncsoportot így értelmezhető: Jóska a külvilág pillanatnyi relációinak adott részét érzékeli és befogadja mint a világról való tudást. A külvilággal (1b) relációban álló világocskához hozzárendelni egy infoncsoportot így értelmezhető: Jóska úgy sejti, hogy Péter az adott információval akarja őt ellátni (függetlenül annak igazságtartalmától). Ha valakinek a pozitív hiedelemvilágocskájához (bel), valamint a negatív vágy- (des) és a 0 polaritásértékkel társított szándék- (int) világocskájához hozzárendeljük ugyanazt az infont, az ezt a tipikus helyzetet szimulálja: az illető észlel valamit, de arra vágyik, hogy az ne úgy legyen, ugyanakkor (esetleg átmenetileg) nem áll szándékában módosítani a helyzeten.

A polaritás paramétere úgy is módosítható, hogy egy valószínűségi változót adunk meg, amelyik az (1c) pontban megadott értékeket megadott eséllyel veszi fel.

Az imént definiált *pillanatfelvétel teljessé tételét* is kérheti a belső felhasználó, amin ezt értjük: ha egy k argumentumú reláció bizonyos entitásokra nincsen értelmezve, akkor negatív vagy „definiálatlan" (∅) polaritás paraméterű hiedelem-világocskához rendeljük, attól függően, hogy az érintett entitások mindegyikére igazak-e a megszorító relációk, vagy sem. Ha például Péter a külvilágban nincsen ott

a *nős* relációban, akkor negatív hiedelemvilágocskához társítandó a kérdéses infon, a „Pécs nős” voltát kimondó infon viszont „definiálatlan” polaritást kell, hogy kapjon.

2.4. A belső felhasználó a külvilágtól függetlenül is rendelhet információt a világocskákhoz, amit az alábbi módon kell a programnak levezényelnie.

Predikátumnevet és argumentumszámot kér, az argumentumhelyeket belső entitásokkal tölti fel, majd szorgalmazza ezek odahorgonyzását más (külső vagy belső) entitásokhoz (ami egyébként nem kötelező a felhasználó számára). Itt egészítjük ki a 2.3. pontot azzal, hogy a pillanatfelvétel során generált argumentumhelyeken álló külső entitásokat ki kell cserélni belső entitásokkal, amelyeket oda kell horgonyozni a csere előtti külső entitásokhoz.

2.5. A belső felhasználó kap egy maglexikont a 2.3. pontban előálló predikátumokról, a 2.4. pontban előálló predikátumokat pedig neki kell jelentésposztulátummal [5] ellátni a program felajánlotta lehetőségek alapján.

Az előbbi esetben triviálisan adódik a jelentés, ezért nem kell külön meghatározni. A jelentés ugyanis elsődlegesen a külvilágra való mintaillesztés sikerén múlik, a 2.3. pontban pedig éppen egy-egy „minta” átmásolása révén hoztunk létre egy-egy belső infont, így hát a mintaillesztés automatikusan sikeresnek tekintendő.

A 2.4. pontban előálló predikátumok a 2.6. pontban meghatározott módon látandóak el jelentéssel. A részletek előtt leszögezzük, hogy ez a ℜeALIS1.1 program meghatározó újdonsága, mivel ez az az eszköztár, ami a formális szemantika-elméletek, a diskurzusreprezentációs megközelítések és a kognitív nyelvészeti felismerések tapasztalatait egyaránt felhasználja.

2.5.1. A formális szemantikából [10] származik a mintaillesztési eljárás, az egymás alternatívájaként szolgáló interpretációs bázisok („lehetséges világok” → ℜeALIS-világocskák) alkalmazása, illetve a sikeres illesztési esetek arányának figyelembe vétele a lehetséges illesztési esetek teljes halmazához képest [11].

2.5.2. A diskurzusreprezentációs elméletekből [11] származik a világocskák részben rendezéses struktúráján való „mozgás” (2.2.).

2.5.3. A kognitív nyelvészetből [13] származik az olyan tényezők figyelembe vétele, amit a nyelv nagyjából ezekkel a szavakkal jelöl meg: *én, te ő, itt, ott, most, akkor, „ezek itt”* (a kontextusban)*, „azok ott”* (rámutatással).

2.6. A programnak folyamatos kérdésekkel kell a belső felhasználót arra késztetni, hogy minden predikátumnévhez jelentésposztulátumot társítson. Így az adatbázist gyarapítjuk, a külső felhasználó által elindított interpretációs feladat során azonban procedurális lépésekként használja majd fel a program a jelentésposztulátumokat.

2.6.1. Egy predikátumhoz mindenekelőtt hangalakot, verzió-megjelölést, változat-megjelölést és angol nyelvű kommentárt kell társítani (hogy később könnyű legyen alternatívákat kipróbálni).

2.6.2. Majd meg kell adni az egyes verziók egyes változatainak argumentum-számát. A program kínáljon változónevet minden argumentum számára, és kérjen szófaji besorolást, valamint specifikusabb alkategória-megjelölést azok számára.

2.6.3. Ezekről az argumentum-változókról és a 2.5.3. pontban említett objektumokról tehet a belső felhasználó újabb és újabb állításokat a külvilági infonokban használt predikátumok révén. Ezek a jelentésmeghatározó állítások konjunktív kapcsolatba lépnek egymással, míg azt az opciót nem választja a felhasználó, hogy az adott verzió adott változatában már nem kíván újabb állítást tenni.

2.6.4. A felhasználó bármely releváns ponton kérhet egy erősebb diszjunkciós lehetőséget is.

2.6.5. Egy definiáló állítás predikátumának kiválasztása után a program annak argumentumhelyeit tölteti fel [5], felajánlva a 2.6.3. pontban említett objektumokat, amelyekhez egy-egy relációt és arányszámot (ld. 2.5.1.) kér társítani. Egy adott argumentumhelyen tehát nem maga az előző mondatban említett objektum kerül majd ellenőrzésre, hanem az, hogy a vele bizonyos relációban álló objektumok milyen arányban elégítenek ki bizonyos követelményeket az összes ilyen objektum közül. A reláció persze default esetben az identikus reláció, az arány pedig a „minden", másodlagosan pedig a „létezik".

2.6.6. A program minden egyes definiáló állításhoz kér egy olyan világocskacímke-láncot, amilyet az (1b) pontban mutattunk be, valamint rákérdez, hogy azt a „bázishoz", az „én"-hez, vagy a „te" objektumhoz képest kell-e (2.5.3.) tekinteni, esetleg a kontextust szimuláló entitáshalmazból vagy az annál szűkebb rámutatási hatókörből kell kiválasztani. A „bázis" alapesetben a külvilág. Végül a program relációt és arányszámot (ld. 2.5.1.) kér társítani a 2.6.6. pontban eddig említett adatokhoz, hasonlóan a 2.6.5. pontban tárgyalt argumentumhelyekhez.

3. A külső felhasználók számára felkínált használati esetek

3.1. A külső felhasználó mondatokat állíthat össze, amelyeknek megkapja az igazságértékelését. A program ehhez kér(het)i a 2.5.3 pontban felsorolt adatok szükséges részhalmazát. Olyan összehasonlításokat is lehet kérni a külvilág és az interpretálói világocskák összevetésére támaszkodva, amelyek alapján [8] olyan kommunikációs „devianciákat" lehet kimutatni, mint példa a hazugság, a tévedés, a blöff.

3.2. Az interpretálandó mondattartalmak összeállítása úgy történik, hogy a külső felhasználó beírja a gép által felajánlott nyelvek egyikén, amelynek adatbázisából a gép a felismert karaktersorozatok alapján lexikai egységeket hoz elő, tipikusan alternatívákat felkínálva.

3.2.1. A belső felhasználó által betáplált grammatikai heurisztikák mennyiségén és minőségén múlik az alternatívák elburjánzásának a megfékezése. A ℜeALIS totálisan lexikalista eszköztár [6] garantálja a „hamis" alternatívák hatékony kiszűrését. A kiválasztott predikátumok argumentumhelyeikkel együtt jelennek meg.

3.2.2. A program sorban kéri az argumentumhelyek betöltését, újabb predikátumok kiválasztásával. Az eljárás akkor ér véget, amikor már nincsen kitöltetlen argumentumhely, mert az „utoljára" választott predikátumok nem kérnek

argumentum-megjelölést. Felhasználói kérésre a program megmutatja, hogy mely pontokon lehetne szabad mondatbővítést végrehajtani.

3.2.3. Bizonyos argumentumhelyeken a program determináns kiválasztását is megköveteli (pl. *minden, egy, a(z), a legtöbb, ez a*).

3.2.4. Bizonyos argumentumhelyekhez a program a „horgonyzó” címke társítását ajánlja fel (a 3.2.1. pontban meghatározott feladat teljesítését követően). A program fejlettebb változataiban ennek lehetnek alternatívái, nyelvészeti szempontokat érvényesítendő (pl. *fókusz* [1]).

4. Összefoglalás, példák

4.1. A külső felhasználó lényegében egy sajátosan megsokszorozott adatbázist kap, ami a való világ modellje mellett annak alternatíváit is felkínálja. A ℜeALIS alapállása szerint ezek a formális szemantikából ismerhető „lehetséges világok” mindig odaköthetőek a világmodellben jelen lévő humán ágensekhez mint azok (tév-) hiedelmei, vágyai, szándékai, álmai stb. (2.2. [2-3] [4]).

Ez a világocska-szerveződés teszi lehetővé, hogy ne csak a külvilág alapján végezzünk el igazságértékelést – ami például a (2a) mondat esetében szükséges és elégséges, hanem olyan mondatokat is tudjon értékelni a program, mint a (2b-c). Hogy a (2b) mondat igaz-e, az például egyáltalán nem múlik a külvilágon, hanem csakis a beszélő (3.1.) hiedelmeinek világocskáján. Ami pedig a (2c) mondat értékelését illeti, ezúttal több lépésben jutunk el ahhoz a világocskához, amely az igazságértékelés bázisát nyújtja; ilyen esetek miatt van szükségünk a világocskák (1b) példában bemutatott rekurzív lokalizálására. A modális attitűdöt kifejező igék (*gondolja, tudod, vágyik*) és egyéb nyelvi elemek (*szerintem*) a 2.6.6. pontban meghatározott eszköz segítségével láthatóak el jelentésposztulátummal: jelentésük lényege abban áll, hogy az állítást kifejező argumentumukban megjelenő állítás igazságértékeléséhez bázisként alkalmazandó világocskára rátaláljunk. Az ilyen nyelvi elemek tehát „irányjelzők” a világocskák részbenrendezett hierarchiájában.

2. példa. Igazságértékelés intenzionális tényezők figyelembe vételével

a. Havazott.
b. Szerintem havazott.
c. Petya úgy gondolja, hogy tudod, hogy Ili arra vágyik, hogy havazzon.
d. It was snowing.
e. It has snowed.
f. Ili éppen utazott haza.
g. Az a magas svéd lány csinos.

4.2. A belső felhasználó igényes grammatikát és szemantikát dolgozhat ki a ℜeALIS1.1 eszköztára révén. A (2a) példabeli magyar mondatnak például több jelentése van, amit a magyar múlt idő jel többféle változatának kidolgozásával ragadható meg (2.6.1.). Az egyik jelentés a (2d)-beli angol fordítással jellemezhető. Ennek elemzése során a program az „ott” és „akkor” értékek megadását kéri a külső

felhasználótól (3.1.) („akkor ott éppen havazott”). A (2e) jelentés igazság-értékeléséhez viszont az „itt” és „most” értékek bekérésére van szükség, amit pedig a külvilágban ellenőrizni kell, az a „havas” állapot. A havazik ige jelentés-posztulátumának részét képezi az eredményállapot („havas”) meghatározása is.

4.3. A fenti (2f) mondat értékelése szintén igényes jelentésleírást követel meg, ugyanis akkor is igaznak kell értékelnünk, ha Ili soha nem ért haza, de „utazott” az „akkor” pillanatában, szándékában állt „megérkezni”, és a beszélő valószínűsíti ezt a megérkezést. A progresszív aspektus megragadásáról van itt szó, ami tehát a külvilág ellenőrzésén kívül bizonyos humán ágensek bizonyos világocskáinak ellenőrzését is igényli (2.6.6.).

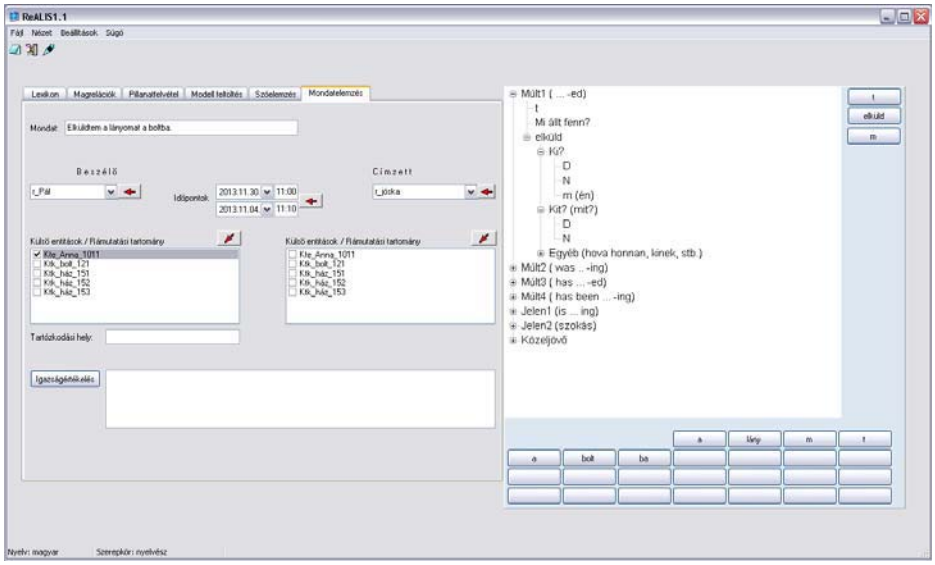
4.4. Ugyancsak a 2.6.6. pontban meghatározott eszközök teszik lehetővé a becenevek igényes pragmatikai kezelését. Ki az a *Petya* például a (2c) példában? Úgy ragadhatja meg a belső felhasználó a jelölet megtalálásának feladatát, hogy a becenevet predikátumként értelmezi, amelynek jelentés-értékelésébe nem (vagy nemcsak) a külvilág ellenőrzése tartozik bele („Péter nevű-e valaki?”), hanem hogy a beszélő Petyaként ismer-e valakit, a hallgató Petyaként ismer-e valakit, és hogy ezt tudják-e egymásról. Belső világocskák ellenőrzése szükséges tehát.

4.5. A (2g) példa azt hivatott illusztrálni, hogy a 2.6.5. pontban meghatározott eszközök is hasznosak az igényes pragmatikai-szemantika leírásban. Tekintsük ugyanis a *csinos* predikátumot definiálatlannak a külvilágban, mivel szubjektív ítéletet fejez ki. Mégis mást jelent azonban azt mondani valakiről, hogy „szerintem csinos”. *Szerintem* nélkül tehát azt jelenti: „a beszélőn kívül az általa mérvadónak tartottak többsége is csinosnak tartja”. A beszélőből kiindulva meghatározhatjuk a hasonló ízlésűek csoportját, és e csoporton belül kell vizsgálni azok arányát, akik csinosnak tartják a szóban forgó hölgyet.

4.6. A (2g) példa alanyi csoportja is a ℜeALIS1.1 eszköztár hasznosságát hivatott illusztrálni a pragmatikailag is „tudatos” igazságértékelésben. „Az a magas svéd lány”: a rámutatás miatt az „ott” értékét fogja kérni a program a külső felhasználótól. Elegáns megengedni, hogy az „ott” egy entitáshalmazt jelöljön ki, amelyből a programnak kell kiválasztania, hogy melyik entitásra igaz a leírás („magas”, „svéd” és „lány”). Az alany ideális esetben „horgonyzó” címkét visel (3.2.2.), amivel azt tudjuk kezelni, hogy a mondatot rosszul formáltnak, ám igaznak kell minősíteni akkor, ha a beszélő tévesen mond svédnek egy mondjuk norvég lányt. A horgonyzásnál tehát a beszélő hiedelemvilágocskái (is) számítanak.

4.7. A ℜeALIS1.1 program tehát a belső felhasználó számára azt teszi lehetővé, hogy tetszőleges természetes nyelvhez nyelvtant írjon és olyan lexikont építsen fel emellé, amelyben minden egységhez (szóhoz vagy morfémához) tetszőlegesen árnyalható pragmatikai-szemantikai leírás tartozik, szervesen ötvözve a legkülönbözőbb nyelvészeti megközelítések er(edm)ényeit (2.5.1-3.). A belső felhasználó egyik célja az lehet, hogy minél teljesebb nyelvleírást adjon, bemenetét nyújtva ezzel például egy fordítást segítő programnak. A másik cél pedig az lehet, hogy bizonyos külső felhasználók igényei szerint építse fel a nyelvtant és a lexikont, például egy nyomozást segítve. Az igazságértékelés ebben a kontextusban nem a nyelvleírás

„próbára tételét” szolgálja, hanem az adatgyűjtést, ami ebben az esetben igen hasznos melléktermék annak.



Hivatkozások

1. Alberti G.: ℜeALIS, avagy a szintaxis dekompozíciója. Általános Nyelvészeti Tanulmányok XXIII. (szerk. Bartos H.) (2011) 51–98
2. Alberti G.: ℜeALIS. Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
3. Alberti G.: Az intenzionalitás számítógépes nyelvészeti kezelése – avagy a ℜeALIS λ szintfüggvénye. MSzNy 2011. Szeged. SzTE Informatikai Tanszékcsoport (2011) 263–275
4. Alberti, G., Károly, M.: Multiple Level of Referents in Information State. Gelbukh, A. (ed.): Computational Linguistics and Intelligent Text Processing (CICLing2012, New Delhi, India), Lecture Notes in Computer Science, Berlin–Heidelberg (2012) 349–362
5. Alberti G., Kilián I.: Vonzatkeretlisták helyett polaritásos hatásláncccsaládok – avagy a ℜeALIS σ függvénye. MSZNY 2010. SzTE Informatikai Tanszékcsoport (2010) 113–126
6. Alberti, G., Kleiber J.: The Grammar of ℜeALIS and the Implementation of its Dynamic Interpretation. Informatica 34/2. (2010) 103–110
7. Alberti, G., Kleiber, J.: Where are Possible Worlds? (Arguments for ℜeALIS). Acta Linguistica Hungarica 59 (1-2) (ed. Katalin É. Kiss) (2012) 3–26
8. Alberti, G., Vadász, N., Kleiber, J.: Ideal and Deviant Interlocutors in a Formal Interpretation System. To appear in A. Zuczkowski (ed.): The communication of certainty and uncertainty. Benjamins (2014)
9. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge Univ. Press (2003)
10. Dowty, D. R., Wall, R. E., Peters, S.: Introduction to Montague Semantics. D. Reidel Publishing Company, Dordrecht (1981)

11. Kamp, H., van Genabith, J., Reyle, U.: Discourse Representation Theory. In Gabbay, D., Guenthner, F. (eds.): Handbook of Philosophical Logic, Springer-Verlag, Berlin, vol. 15, (2011) 125–394
12. Kilián I., Alberti G., Szabó V.: Metamodell vezérelt felépítmény modális világszerkezetek létrehozására, feltöltésére és lekérdezésére. SzámOkt 2013 (Nagyszeben, 2013. október 10-13.). XXIII. Nemzetközi Számítástechnika és Oktatás Konferencia. Szerk. Bíró K. Á., Sebestyén-Pál Gy. Erdélyi Magyar Műszaki Tudományos Társaság (2013) 225–232
13. Kiss Szabolcs: Elmeolvasás. Budapest, Új mandátum (2005)
14. Kleiber, J., Alberti, G.: Uncertainty in Polar Questions and Certainty in Answers? To appear in S. Cantarini, W. Abraham, E. Leiss (eds.): Certainty-uncertainty – and the attitudinal space in between. Benjamins (2004)
15. Seligman, J., Moss, L. S: Situation Theory. van Benthem, J., ter Meulen, A.: Handbook of Logic and Language. Amsterdam / Cambridge (1997) 239–309

PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer

Orosz György, Novák Attila

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

1. Bevezetés

A szófaji egyértelműsítés és a lemmatizálás jól ismert problémái a nyelvtechnológiának. A fenti feladatokat gyakran különálló komponensek végzik egy szövegfeldolgozási láncban, ami forrása lehet a lánc teljesítményének romlásának. Az utóbbi évtizedben számos olyan eszköz jött létre, mely képes magyar nyelvű szövegek szófaji vagy morfológiai egyértelműsítésére, ilyenek pl.: a Hunpos [2], az OpenNLP[1], a `magyarlanc` [5] és a PurePos [4]. Ezek közül csak néhány [5,4] képes teljes morfológiai elemzések közti egyértelműsítésre, továbbá egy olyan hibrid láncban, ahol szabályalapú modulok is fontos szerepet töltenek be, ezek egyike sem tud maradéktalanul együttműködni társaival. A szófaji és morfológiai egyértelműsítés napjaink egyik aktuális problémája a doménadaptáció kérdése: hogyan alkalmazható egy általános nyelvi modell egy új doménen?

Írásunkban ismertetjük a PurePos rendszer továbbfejlesztett változatát, melyben az eredeti algoritmus több ponton megváltoztattuk, úgy, hogy az egy hibrid elemző lánc hasznos tagja legyen. Továbbá az eszközt olyan jellemzőkkel láttuk el, melyek lehetővé teszik használatát azon doménadaptációs feladatokban is, amikor szabályok alkalmazásával növelni lehet az elemzőlánc teljesítményét. Cikkünk végén ismertetjük a módosított rendszer megnövekedett teljesítményét.

2. A továbbfejlesztett algoritmus

A PurePos alapja olyan rejtett Markov-modellezésen alapuló algoritmusok, melyeket már számos alkalommal sikerrel használtak szófaji egyértelműsítő rendszerekben (pl. HunPos[2], TnT[1]). A tagger a címkézéshez egy lexikális és kontextuális modellt használ, melyek együttese (1) formalizálja az egyértelműsítés feladatát[2].

$$\arg\max_T P(T|W) = \arg\max_T P(W|T)P(T) \tag{1}$$

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

[2] T címkesorozatot, t címkét, W mondatot, míg w egy szót jelöl.

$$P(t_k|t_{k-1,k-n}) = \sum_{i=1}^{n} \lambda_i \hat{P}(t_k|t_{k-1,k-i}) \tag{2}$$

$$P(w_k|t_{k,k-m+1}) = \sum_{i=1}^{m} \lambda_i \hat{P}(w_k|t_{k,k-i+1}) \tag{3}$$

Az egyértelműsítő a lexikai és kontextuális modellek becslésére simított n-gram modelleket használ ((2) és (3)), melyek paraméterei változtathatóak, de alapesetben $n = 2$ és $m = 2$. Az interpoláció környezetfüggő módon deleted interpolation módszert használva történik (l. [1]). A PurePos a tanítóanyagban nem látott (OOV) szavak taggeléséhez az integrált morfológiai elemző analízisein túl egy szóvég alapú javasló rendszert is tartalmaz, melynek működése a [1]-ben részletezett hasonló moduljára épül. Az eredeti egyértelműsítő a lemmatizáláshoz egy maximum likelihood becslésen alapuló unigram modellt használ, míg a dekódolást a Viterbi algoritmus végezte.

2.1. A morfológiai tudás fejlettebb használata

A PurePos korábbi verziói is használtak morfológiai elemzőt az ismeretlen szavak jobb taggelése céljából, viszont nem voltak képesek teljesen kihasználni ezt az értékes tudást. Az egyik ilyen eset, amikor egy ismeretlen szóhoz az egyetlen morfológiai elemzés olyan, hogy annak címkéje a tanítóanyagban még nem fordult elő. Ekkor a tag valószínűsége 0 volt, ami vagy a jó megoldást kizárásával járt, vagy pedig azt eredményezte, hogy a tagger – a logaritmikus reprezentáció tulajdonságai miatt – a mondat elemzéssorozataihoz egyforma valószínűséget rendelt. Ezt a hibát úgy javítottuk, hogy valószínűségi értékként 1-et használunk.

Egy ettől összetettebb jelenség, amikor a helyes elemzés továbbra is ismeretlen a statisztikai rendszer számára, viszont mellette több más lehetséges annotáció is feltűnik, amikhez a trigram modell már képes gyakorisági értékeket rendelni. Ilyenkor is számtalan esetben a 0 gyakorisági érték miatt elveszett a helyes elemzés, amit az új PurePosban címkék megfeleltetésével küszöböltünk ki. A módszer alapja, hogy a tagger indításakor egy konfigurációs fájl használatával lehetősége van a címkék megfeleltetésére, ami az egyértelműsítés folyamán azt eredményezi, hogy az így megadott, tanítóanyagban nem látott elemzésekhez is a leképezett annotáció gyakorisági értékei számolódnak.

2.2. Fejlettebb szótövezés

$$\arg\max_{l} P(l|t, w) \tag{4}$$

Egyes w szavak t címkéjéhez tartozó l optimális lemmát (4) segítségével határozzuk meg. Ennek becslésére a szoftver korábbi változata a tanítóanyag alapján számolt maximum likelihood becsléssel végezte a szótövek rangsorolását. Jelen

munkánkban módosítjuk ezt az eljárást, hogy az ismeretlen szavakhoz használt guesser valószínűségi becsléseit is figyelembe vegye a szoftver.

$$P(l|t,w) = \frac{P(l,t|w)}{P(t|w)} \tag{5}$$

Ehhez (4) átírásából kapjuk (5)-öt, amiből a nevező konstans volta miatt elhagyható. A számláló eloszlására, mint korábbi munkánkban azt megmutattuk, jól alkalmazható a suffix guesser interpolált modellje. Hogy mind az unigram valószínűségek, mind pedig az utóbbi erősségeit alkalmazhassa az egyértelműsítő, ezek log-lineárisan interpolált kombinációját számoljuk (6).

$$P(l|w,t) = P(l)^{\lambda_1} P(l,t|w)^{\lambda_2} \tag{6}$$

Algoritmus 1 Az interpolált modell paramétereinek számítása

```
 1: for all (w, t, l) do
 2:     candidates ← generateLemmaCandidates(w, t)
 3:     maxUnigramProb ← getMaxProb(candidates, w, t, unigramModel)
 4:     maxSuffixProb ← getMaxProb(candidates, w, t, suffixModel)
 5:     actUnigramProb ← getProb(w, t, l, unigramModel)
 6:     actSuffixProb ← getProb(w, t, l, suffixModel)
 7:     unigramProbDistance ← maxUnigramProb − actUnigramProb
 8:     suffixProbDistance ← maxSuffixProb − actSuffixProb
 9:     if unigramProbDistance > suffixProbDistance then
10:         λ2 ← λ2 + unigramProbDistance − suffixProbDistance
11:     else
12:         λ1 ← λ1 + suffixProbDistance − unigramProbDistance
13:     end if
14:     normalize(λ1, λ2)
15: end for
```

Az interpoláció paraméterinek kalkulálásához Brants [1] ötletét használjuk, miszerint a tanítóanyagon jobban teljesítő modell nagyobb súlyt kap (vö. 1. algoritmus). Ehhez az egyes komponensek tanítása után, a korpusz összes szavára kiértékeljük a szótövező modulokat (3-8. sor), és negatív súlyokat adunk a rosszabbul teljesítőnek (9-13. sor). A tanítás végén a $\lambda_{1,2}$ paraméterek értékei normalizálásra kerülnek.

2.3. *k*-legjobb kimenet

Számos esetben igény van a tagger kimenetén a legjobbnak vélt elemzési szekvencián túl a lehetséges annotációk egy halmazára is. Ennek érdekében a PurePos a Viterbi algoritmus használatán túl támogatja a Beam-search dekódolást is, mely paraméterei futtatási opciókként állíthatóak. Az eszköz az egyes szekvenciákhoz nyilvántartja még azok rangsorolásához használt logaritmikus

valószínűségi értékeket is, amik szintén megjelenhetnek az outputon. Ezek a (7) alapján számolódnak.

$$Score(w_{1,m}, t_{1,m}) = log \prod_{i=1}^{m} P(w_k|t_{k,k-m+1})P(t_k|t_{k-1,k-n}) \tag{7}$$

2.4. Hibrid komponensek használata

A morfológiai elemző használatán túl, a rendszer lehetővé teszi még a felhasználó számára hogy további nyelvi tudással segítse a taggelés eredményességét. Így a PurePos inputján az egyes tokenekhez lehetséges elemzések és azokhoz tartozó valószínűségek is megadhatóak. Ez a képessége jól használható pl. olyan doménadaptációs feladatok esetén, amikor a céldomén egyes, különleges módon használt szavai jól körülhatárolhatóak. Ezeken túl az adaptálható input és egy morfológiai elemző használatának segítségével további, nagyobb hatótávolságú szabályok is megfogalmazhatóak. A k-legjobb elemzési opciót használva az elemző lánc építőjének további lehetősége nyílik a teljesítmény további javítására, vagy ún. self-training használatára is.

3. Eredmények

A fejezetben bemutatunk egy olyan esetet, amikor a PurePos 2.0 fent részletezett tulajdonságai használatával jelentősen sikerült javítani az elemzőlánc teljesítményén. Az Ó- és Középmagyar Korpusz [3] morfológiai annotációjának készítése során 200 dokumentum mintegy 75000 tokenjéhez kellett egyértelműsített morfológiai elemzést rendelni. Munkánk során a korpusz 80%-át tanításra használtuk, míg 10-10%-ot az algoritmus paramétereinek beállítására, illetve annak kiértékelésére.

1. táblázat. Az egyértelműsítő pontossága a teszthalmazon

	Szófaji címkézés	Teljes egyértelműsítés
PurePos 1.0	91,09%	51,32%
PurePos 2.0	96,72%	96,48%
Címkeleképezésekkel	96,75%	96,51%
Előfeldolgozó szabályokkal	96,86%	96,66%
A teljes lánc	96,89%	96,67%

A 1. táblázatban bemutatjuk a PurePos első verziójának teljesítményét, ezen túl ismertetjük még az új komponensek használatával elért teljesítményjavulást is. A bemutatott szótövező algoritmus használatával jelentős mértékben sikerült csökkenteni a hibák számát, míg a többi modul is javított a pontosságon. A címkék megfeleltetéséhez mindössze egy szabályt alkalmaztunk, mely igekötős igék eloszlását köti az igekötő nélküliekhez. A hibrid komponensben használt

szabályok mindössze két megfigyelés formalizálásával történtek. Ezek közül az egyik a mondat eleji *a* szó névelő voltát határozza meg, míg a másik gyakori foglalkozást jelentő tulajdonnevek elemzéseit egyértelműsíti. Végül fontos még megemlíteni a *k*-legjobb elemzési szekvencia használatát. Ezzel az opcióval a bemutatott legjobb teljesítményű konfiguráció hibái további csökkenthetőek akár 98,65% teljes egyértelműsítési pontosságot megközelítve.

4. Összefoglalás

Munkánkban bemutattuk a nagy pontosságú PurePos rendszer egy továbbfejlesztett változatát, mely a jobb szótövezési teljesítményen túl immár hasznos eleme lehet egy hibrid elemző láncnak is. A tagger jól használható olyan környezetekben, ahol egyszerű szabályok bevezetésével lehetséges a teljesítmény javítása. Dolgozatunkban egy használati eseten keresztül megmutattuk, hogy akár kis méretű tanítóanyag esetén is nagy pontosságú morfológiai egyértelműsítő hozható létre.

Az alkalmazás JAVA nyelven íródott, nyílt forráskódú[3] és Python nyelvhez is tartalmaz illesztést. A részletezett tulajdonságok és a megengedő felhasználási feltételek miatt is a PurePos megfelelő választás lehet elemzési feladatok egyértelműsítő komponensének.

Köszönetnyilvánítás

Ez a projekt a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014. támogatásával készült.

Hivatkozások

1. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the sixth conference on Applied Natural Language Processing. pp. 224–231. Universität des Saarlandes, Computational Linguistics, Association for Computational Linguistics (2000)
2. Halácsy, P., Kornai, A., Oravecz, C.: HunPos: an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL. pp. 209–212. Prague, Czech Republic (2007)
3. Novák, A., Orosz, G., Wenszky, N.: Morphological annotation of Old and Middle Hungarian corpora. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 43–48. Sofia, Bulgaria (2013)
4. Orosz, G., Novák, A.: PurePos – an open source morphological disambiguator. In: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science. pp. 53–63. Wroclaw (2012)
5. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of Recent Advances in Natural Language Processing 2013. Association for Computational Linguistics (2013)

[3] http://nlpg.itk.ppke.hu/software/purepos

Online nganaszan történeti-etimológiai szótár

Szeverényi Sándor[1], Tóth Attila[2]

[1] Szegedi Tudományegyetem, Finnugor Nyelvtudományi Tanszék,
6722 Szeged, Egyetem u. 2.
szevers@hung.u-szeged.hu

[2] Szegedi Tudományegyetem, JGYPK Informatika Alkalmazásai Tanszék,
6725 Szeged, Boldogasszony sgt. 6.
attila@jgypk.u-szeged.hu

Kivonat: A bemutatóban a nganaszan nyelv online diakrón kognitív onomasziológiai szótár munkálatairól számolunk be. A szótár diakrón, mert a szókészlet történeti-etimológiai hátterét tárja fel, kognitív, mert az egymással összefüggő alakok közötti szemantikai kapcsolatokat is meghatározza, és onomasziológiai, mivel fogalmak felőli keresést, rendszerezést is lehetővé tesz. Mindezt úgy, hogy nem egy kész szótárat digitalizál, hanem olyan webes felületet hozunk létre, amely egyben a kutatás eszköze is.

1 A projekt célja

Projektünk újszerűsége egy történeti lexikográfiai probléma új típusú számítógépes feldolgozása. A szótár alapja már létezik, nyilvánossá a projekt végén, 2015 tavaszán fog válni. A projekt az OTKA támogatásával valósul meg.[1]

A munkálat nyelvészeti célja a nganaszan nyelv kognitív diakrón onomasziológiai szótárának kialakítása, a nganaszan szókincs rendszerezése szinkrón és diakrón szempontból (erről részletesebben [13]), olyan módon, hogy a későbbiekben a szótár más nyelvek adataival is ki tudjanak egészülni. Éppen emiatt a célkitűzések között szerepel a folyamatos javíthatóság és bővíthetőség biztosítása.A megvalósításhoz kapcsolódó technikai elvárások a következőkben foglalhatók össze: egy olyan szabad felhasználású, weben elérhető online felület, „eszköz” létrehozása, amely egyszerű módon jeleníti meg egy-egy lexéma formai, szemantikai tulajdonságait, történeti hátterét, valamint kapcsolatait más lexémákkal, és a megjelenített információk között összetett keresési kombinációkat tesz lehetővé.

2 A nganaszan nyelv

A nganaszan nyelv szókincsének és annak történetének dokumentáltsága tipikusnak mondható – a világ nyelveinek jelentős részéhez hasonlóan kevesen beszélik, hiányosan dokumentált és a beszélői kompetencia gyorsan tűnik el.

[1] A nganaszan nyelv diakrón kognitív onomasziológiai szótára (K100854).

Már az első lejegyzett nyelvi adatok is viszonylag késői időkből, a 18. század végéről származnak, s a módszeres nyelvi gyűjtés csak a 20. század utolsó évtizedeire vált általánossá. M. A. Castrén 19. századi gyűjtései ugyan történeti szempontból is jelentősek, ám a mennyisége nem teszi lehetővé, hogy külön történeti rétegként jelenítsük meg. Nganaszan írásbeliség nem alakult ki, mindössze egy gyakorlati szótár [10] és egy iskolás könyv [14] jelent meg. A kilencvenes években elsősorban Eugene Helimski, majd később tanítványa Valentin Guszev vezetésével történt szisztematikus nyelvi gyűjtés, melynek révén a nganaszan anyag mennyisége megsokszorozódott, ehhez magyar kutatók gyűjtései is hozzájárultak. Jelenleg a számunkra elérhető anyag mennyisége kb. 40-50 000 mondat.

Nem meglepő, hogy a nganaszan nyelv nagyon gyorsan halad az eltűnés felé. Beszélőinek száma a 2010-es oroszországi népszámlálási adatok szerint 125, a nyelvet anyanyelvi szinten beszélőké viszont ennek csak a töredéke lehet. Ez azt jelenti, hogy anyanyelvi kompetencia a projekthez nem áll rendelkezésre, jelentős mennyiségű, normalizált írásos korpusz pedig nincsen.

A nganaszan szókincs történeti háttere is csak részben feltérképezett, ez elsősorban a szókészlet szamojéd, uráli eredetű részére vonatkozik. Nincsen olyan korábban megjelent munka, amely a teljes nganaszan szókincs történetét, sajátosságait bemutatná, azaz a mára általánossá váló eljárás – egy nyelv vagy nyelvcsalád történeti-etimológiai szótárának digitalizálása, majd annak átdolgozása, frissítése, kiegészítése – a mi esetünkben nem lehetséges. Ugyanakkor annak sem látjuk értelmét, hogy napjainkban (csak) papíralapú szótárat készítsünk (noha az elmúlt időszakban a nemzetközi irodalomban van ilyenre példa, például [2, 3, 11]), illetve annak sem, hogy először elkészítsünk egy szótárat, s utána végezzük el a digitalizálást.

Mi megfordítottuk a sorrendet: előbb készítjük el a digitális verziót, s onnan lehet majd letölteni – a kívánt keresési eredményekkel – a nyomtatottat. Ehhez viszont olyan szerkezetet kellett kialakítani, amelyet lehetőség szerint a későbbiekben ne kelljen módosítani, csak finomítani, még akkor sem, amikor új nyelvek adatait dolgozzuk fel. Ennek megfelelően nemcsak az a feladat, hogy a nganaszan nyelvhez „passzoló” paraméterlistákat dolgozzunk ki, hanem a tipológiai szempontok is érvényesülni tudjanak.

2.2 A nganaszan korpusz

A nganaszan nyelvi anyagot zárt korpuszként kezeljük, ennek törzsanyagát az említett szótár adja (kb. 3500 címszó), illetve az azon alapuló angol változat [1]. Ezt az anyagot egészítjük ki olyan szócikkekkel, amelyek más forrásokban fordulnak elő. A történeti tárgyú munkák anyagát is külön-külön dolgozzuk fel, ezek legfontosabb forrásai: Janhunen 1976, Janhunen 1981, Helimskij 1997, [5, 6, 7]. Ezért gondoltuk, hogy célszerű lenne egy olyan szótár kialakítása, amelybe folyamatosan lehet „pakolni” az információkat, ha új közlések, publikációk jelennek meg, akkor azok anyagát rögtön be lehessen építeni az adatbázisba.

3 A szótár szerkezete

A szótár sajátos vonása, hogy a hangtörténeti jellemzők helyett a lexikológiai hátteret vizsgálja: definiálja a szóalakok közötti kapcsolatot, és a hozzájuk rendelhető jelentések közötti kapcsolatokat. Ennek megfelelően a szótárnak három fontos felülete van: a paraméterlisták („data") felülete, a „form-concept" felület, és a „process-relation" felület.

3.1 A paraméterlisták (data)

A következő információcsoportok szerkeszthető rendszere található itt:

- nyelv / nyelvjárások: a rekonstruált (proto) nyelvek és az adatbázisban előforduló természetes nyelvek és nyelvjárások együttes listája;
- a szófaji rendszer: a jelentéssel együtt tárolt információ, jelenleg a nganaszan szófaji rendszerét tükrözi;
- irodalomlista: egyfelől az elsődleges adatokat tartalmazó munkákat, másfelől a szekunder hivatkozásokat tartalmazza;
- a szóalakok közötti kapcsolatok rendszere: a szóalkotási módok és azok alcsoportjai (összetétel, képzés, reduplikáció, kölcsönzés, folytonosság);
- opacitás: a motivációra vonatkozik, azaz átlátszó vagy átlátszatlan-e egy kifejezés;
- bizonyosság: a megállapított kapcsolat bizonyossága (biztos vs. bizonytalan);
- a szemantikai kapcsolatok rendszere: a rendszer nagyrészt a tübingeni kutatók által kidolgozott felosztást követi (például [4, 8], lásd lejjebb);
- jelentéscsoportok rendszere: a jelentéscsoportok rendszerét a Rapid Word Collection módszerét – amelyet kifejezetten dokumentációs nyelvészek számára dolgoztak ki a SIL munkatársai [12] – követve alakítottuk, illetve alakítjuk ki. Azért döntöttünk e felosztás mellett, mivel egyfelől az anyag szabadon felhasználható és adaptálható, másfelől a kategorizálás során hasonló kérdések merülnek fel, mint amikor terepmunkát végzünk, azaz egy gyakorlati szótári anyagot leginkább ez követ.
- speciális karakterek: egy újabb nyelv bekapcsolása azt is jelentheti, hogy új karakterre van szükség, itt könnyedén tudjuk előállítani a megfelelő karakterek, amelyek rögtön megjelennek a virtuális „billentyűzeten".

Mindegyik csoport egyszerűen módosítható (bővíthető, ill. törölhető). Természetesen arra figyelemmel kell lenni, hogy például egy adott paraméter törlése (pl. nyelvjárás) milyen kapcsolatokban okoz változást (pl. az adott nyelvjárásba tartozó lexémák).

3.2 Szóalakok és jelentések (form & concept)

Ez a rész szolgál a szóalakok és jelentések bevitelére, katalogizálására és a lexéma-jelentés kapcsolatok létrehozására. Ez azt jelenti, hogy egy szóalakot csak egyszer tárolunk el, homonímia esetén sem szükséges az alakot újra rögzíteni. A jelentéseknél

hasonló a helyzet, azzal a különbséggel, hogy a jelentéseket minden esetben úgy kell megadnunk, ahogyan a forrásban szerepelnek, így például a ’mountain’ jelentés háromszor szerepel jelenleg a szótárban:

’mountain ridge, mountain range’
’mountain, rock’
’mountain, hill, ridge’

Egy ’mountain’ részleges egyezéses keresés kiadja mindhárom találatot, s ha teljesen biztosak akarunk lenni abban, hogy minden találat megjelent-e, akkor a ’mountain’ jelentéscsoportját (jelenleg LAND) is lehet használni.

3.3 Lexémák és szemantikai kapcsolatok

Saját szerkesztői felülete van az egyes lexéma+jelentés párok közötti alaki és szemantikai kapcsolatoknak (process – relation), ugyanitt lehet a változás irányát is meghatározni (source – target). Ez felveti azt a kérdést, hogy a jelentésváltozás és a szinonímia között megállapítható-e a határ.

A szóalkotási eljárások (process) jelenleg a nganaszan szóalkotási módokat tartalmazza (képzés, átvétel, összetétel, lexikai folytonosság stb.), illetve ezek alcsoportjait. A jelentések közötti kapcsolatokat két nagy csoportja a metaforikus (hasonlóságon alapuló), illetve a metonimikus (kontiguitáson alapuló) kapcsolatok.

Természetesen egy kapcsolatot több minősítéssel is el lehet látni. Amit pedig a minősítésekkel nem lehet megadni, azt a „comment” részben lehet megmagyarázni. Fontos, hogy a rendszer a formai és a jelentésbeli változásokat, kapcsolatokat együtt láttatja, a diakrón kognitív onomasziológiai munkálatoknak ez az egyik alapvető célja.

Mivel a kapcsolatok meghatározása gyakran nem egyértelmű, vagy csak nagyon „leegyszerűsítve” adja vissza a tényleges relációkat, ezért a „comment” résznél lehetőség van szöveges kiegészítésre.

Ezáltal gyakorlatilag szóláncokat tudunk létrehozni, be tudjuk mutatni egy adott szótő eredetét, más nyelvekben való megjelenését, származékait, jelentéseit, s azok viszonyait.

3.4. Keresés

Az elmondottakat az *ə̂ntə̂j* ’boat’ > *ŋənduj* ‘a kind of boat’ > *tuu ŋənduj* ’steamboat, steamer, steamship’ szólánccal szemléltetjük. A nganaszan *ŋənduj* ‘a kind of boat’ szóra keresünk rá. Elsődleges forrása az említett Kosterkina et al. (2003) szótár [10]. A jelentést besoroltuk a TRAVEL és a FISHING kategóriákba. Ha rákeresünk a *ŋənduj* szó, akkor a következő lényeges információkat kapjuk:

- a *ŋənduj* forrása a proto-szamojéd rekonstruált *ə̂ntə̂j* ’boat’. Ennek forrása Janhunen etimológiai szótára;
- a *ŋənduj* és a *ə̂ntə̂j* szóalakok között kapcsolat lexikai folytonosság (azaz a nganaszanban egy korábbi nyelvállapotra rekonstruálható alak a hangváltozásokat leszámítva változatlanul meg);

- a *ŋənduj* és a *ânntâj* szóalakok közötti kapcsolat leginkább a konceptuális/fogalmi azonosság kategóriájába tartozik, mivel mindkettő csónakot jelent;
- a *ŋənduj* 'boat' szóalak + jelentés kapcsolat részleges forrása újabb elemeknek, így például a *tuu ŋənduj* ' steamboat, steamer, steamship' szókapcsolatnak;
- a *tuu ŋənduj* 'steamboat, steamer, steamship' szóalak + jelentés forrásai között megjelenik a *tuj* 'fire' szó is. A *tuu ŋənduj* összetételt szóalkotási szempontból összetételnek minősítjük. A *tuu* a *tuj* szóalak genitívuszi alakja (ezt az információt a comment részben tudjuk tárolni). Természetesen a *tuu ŋənduj* forrásai között a *tuj* is megjelenik;
- A *ŋənduj* 'a kind of boat' és a *tuu ŋənduj* 'steamboat, steamer, steamship' közötti szemantikai kapcsolat egyfajta fogalmi hasonlóságon alapuló specializáció, a csónak jármű egy speciális fajtájára utal, ezért a metaforikus kapcsolatok közül a fogalmi hasonlóság mellett a taxonomikus alárendelés is szerepel a minősítések között.

4 A technikai háttér

Mivel a cél olyan online rendszer kifejlesztése volt, amely adattartalma folyamatosan fejleszthető és felhasználása minél szélesebb kör számára elérhető, így a webes alkalmazás a legkézenfekvőbb megoldás. Ezáltal a felhasználói és az adminisztrátori funkciók elvégzéséhez is elég egy böngésző. Ez jelentősen megkönnyíti a bővítési, további nyelvekkel való kiegészítési munkafolyamatot.

Alapvető elvárás a rendszerrel szemben, hogy az adattartalom dinamikusan változtatható, bővíthető legyen úgy, hogy az adatok redundanciáját elkerüljük. Így a rendszer alapját egy olyan SQL adatbázis képezi, amely központi magját a szóalak és jelentés párok alkotják, illetve az ezekből képezett formális és szemantikai kapcsolatok. Azaz külön egységként tároljuk a szóalakokat és a jelentéseket, az ezek közötti kapcsolatot, valamint az így képzett párok közötti átmeneteket. Ez a modell alkalmas arra, hogy bizonyos szóalakok (illetve jelentések) több jelentéssel (illetve szóalakkal) is párt alkossanak, így a poliszém és a homonim alakok redundanciamentesen jól ábrázolhatók. Továbbá az ezeket jellemző attribútumok lehetséges értékei szintén külön tároltak, így ezek bővítése könnyen elvégezhető.

Egy ilyen rendszerben elemi elvárás, hogy az alkalmazás képes legyen a tartalmazott nyelvek speciális karaktereinek a kezelésére, illetve olyan felhasználói felületet nyújtani, ahol az ilyen karakterek könnyen beilleszthetőek. Mivel a szerzők célja a rendszert további nyelvekre is kibővíteni, így ennek kezelését rugalmasan kell megoldani. Emiatt egyrészt az adattárolás UTF-8 kódolással történik, valamint az adatbázisban külön tárolásra kerülnek a speciális karakterek és azok kódjai is. Másrészt a speciális karakterek bevitelét a felhasználói felületen egy virtuális billentyűzet segíti, amelyen szereplő karakterek dinamikusan állnak össze az adatbázis ilyen karaktereit tartalmazó tábla tartalma alapján.

5 Tervek

Szótárunkkal azokhoz a kutatásokhoz kívánunk a jövőben kapcsolódni, amely leginkább a lexikális tipológia, s annak különösen a diakrón ágához tartozik. Koch és Marzo [9] szerint a lexikalizáció formai és kognitív motivációjának diakrón tipológiai rendszerezése a következők miatt fontos:

(i) lehetővé teszi az egyes nyelvek motivációs „profiljának" megalkotását;
(ii) lehetővé teszi nyelveken átívelő tendenciák és idioszinkráziák megállapítását (Vannak-e „transzparensebb" vagy kevésbé transzparens nyelvek? Vannak-e „metaforikusabb" nyelvek?);
(iii) lehetővé teszi nyelveken átívelő és nyelvspecifikus motivációs preferenciák megállapítását.

Ezért célunk, hogy az adatbázis további nyelvekkel, s adatokkal bővüljön, s a munka a projekt lejárta után is folytatódjon.

Hivatkozások

1. Bradley, J., Wagner-Nagy, B.: Nganasan–English Dictionary. Ms. Wien: Hamburg. (2013)
2. Fortescue, M., Jacobson, S, Kaplan, L.: Comparative Eskimo Dictionary . Alaska Native Language Press, Fairbanks (1994, 20122)
3. Fortescue, M.: Comparative Chukotko-Kamchatkan Dictionary. Trends in Linguistics. Documentation. Mouton de Gruyter, Berlin: New York (2005)
4. Gévaudan, P.: Typologie des lexikalischen Wandels. Stauffenburg, Tübingen. (2007)
5. Helimski, E.: Die matorische Sprache. SUA 41. JATE, Szeged (1997)
6. Janhunen, J.: Samojedischer Wortschatz. Castrenianumin toimitteita 17, Helsinki (1977)
7. Janhunen, J.: Uralilaisen kantakielen sanastosta. JSFOu 77. (1981) 219–274
8. Koch, P.: Lexical typology from a cognitive and linguistic point of view. In Haspelmath, Martin, König, Ekkehard, Oesterreicher, Wulf, Raible, Wolfgang (Hrsg.): Linguistic Typology and Language Universals = Handbook of Linguistics and Communication Science 20/2. Mouton de Gruyter, Berlin. (2001) 1142–1176
9. Koch, P., Marzo, D.: A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. Studies in Language 31:2 (2007) 259–291.
10. Kosterkina, N. T., Momde, A. Č., Ždanova, T. Ju. [Костеркина, Н. Т., Момде, А. Ч., Жданова, Т. Ю.]: Словарь нганасанского-русский и русско-нганасанский, Филиал издательство «Просвещение», Санкт-Петербург (2001)
11. Nikolaeva, I.: A Historical Dictionary of Yukaghir. Trends in Linguistics. Documentation. Mouton de Gruyter, Berlin: New York (2006)
12. Rapid Word Collection http://www.rapidwords.net/ (2013. november 28.)
13. Szeverényi S.: Mire jó egy nganaszan online diakrón kognitív onomasziológiai szótár? Nyelvtudományi Közlemények 108 (2012) 197–218
14. Žovnickaya, S. N. [Жовницкая, С. Н.]: Букварь, Санкт-Петербург, Просвещение (2001)

IX. Angol nyelvű absztraktok

Deep cases in the 4lang concept lexicon

Márton Makrai

Hungarian Academy of Sciences, Institute for Computer Science and Control
e-mail: makrai@sztaki.mta.hu

4lang is a multilingual lexicon for general human language understanding containing formal representations of word meaning in the monosemic approach to lexical semantics, which means that items are language independent concepts covering different uses of the same word, uses in different sentence patters and even in different parts of speech with the same meaning representation.[1] Multilinguality and abstractness of items have the effect that a simple deep case (or thematic) frame captures uses with different arity (i.e. transitive and intransitive). Deep cases denote the nodes in the graph representing the meaning of a predicate where the representation of the argument (single word, entity or phrase) has to be inserted.

4lang makes no clear cut between complements and adjuncts. Basically an argument is represented by a deep case whenever its needed for building of the representation of the verb. As uses of the same verb with different arities are handled in the same item, deep cases are used consequently in different verb patterns, and all possible arguments are included in the representation. However, as verbs can be defined as special cases of other verbs (biting is cutting with teeth), arguements are inherited, so not every argument is listed directly in the definition of some verb. An other source of inplicite arguments are constructions providing verbs with outer arguments e.g. *pain a picture* ***for somebody***.

Most frequent verbal deep cases are agents (denoted by AGT), patients (PAT), and datives (DAT). Patient plays the role of the neutral case it seems to play in many systems (Somers 1987)[2]. Following the unaccusative hypothesis, arguments of intransitive verbs split to agents and patients. The label "dative" is taken from Fillmore (1968), but our understanding is narrower as we mainly restrict dative to recipients in ditransitives (verbs of communication (e.g. *tell*) and transfer (e.g. *give*)). There are three locative cases in 4lang (TO, FROM, and AT), the latter being used for the abstract goal of relational nouns such as *occasion* and *need* as well. A greater group of relational nouns require the possessive (POSS) such as *absence* and *duty*. Quirky cases can be marked in a language dependent module.

Deep cases in 4lang are not restricted to verbs. Some grammatical features such as plural contribute to meaning, so morphemes expressing them have deep cases. Representations of productive derivational suffixes and adpositions also refer to the conceptual element they attach to with deep cases (REL).

[1] The lexicon, automatically collected word forms in 50 languages, a verctor space language model (embedding) computed from 4lang, and articles can be found at http://hlt.sztaki.hu/resources/4lang/

[2] References can be found in the full version of the article that is in Hungarian.

4FX: Automatic Detection of Light Verb Constructions in a Multilingual Corpus

Anita Rácz[1], István Nagy T[1], Veronika Vincze[2]

[1]University of Szeged, Department of Informatics
raczanita89@gmail.com, nistvan@inf.u-szeged.hu
[2]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

In this paper we describe the 4FX corpus, the first English, Hungarian, Spanish and German parallel corpus, which is manually annotated for light verb constructions (LVCs). For corpus construction, legal texts from the JRC-Acquis legal parallel corpus were selected. Annotation principles and statistical data on the corpus are also provided, and data for the four different languages are contrasted. We also present the results of a machine learning-based approach that allows us to identify light verb constructions in free texts. The tool was originally implemented to automatically detect Hungarian and English LVCs. However, we were able to easily adapt this data-driven machine learning-based approach to the other languages, since manually annotated corpora are also available in Spanish and German in the 4FX corpus. Moreover, we were able to define language-specific features, like the gender of the noun in Spanish and German, for the machine learning-based method to detect LVCs in free texts in these different languages. Our applied method proved to be sufficiently robust, since it outperformed our dictionary labeling baseline method in the case of all the four different languages.

Multi-level Syntactic Representation in the Szeged FC Treebank

Katalin Ilona Simkó[1], Veronika Vincze[2], Richárd Farkas[1]

1 University of Szeged, Department of Informatics
kata.simko@gmail.com
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

The two most widely used syntactic theories among the existing ones are constituent and dependency syntactic theories. The Szeged Treebank contains manually annotated syntactic trees in both constituent and dependency formats. Both analyses have their advantages and disadvantages as well. The constituent representation groups words that are part of the same unit of meaning into phrases, while dependency grammars connect the words of the sentence directly to each other without the use of abstract nodes.

It is undecided whether either of these grammars can be considered superior for the analysis of Hungarian and other morphologically rich languages, as both representations contain important information on their syntax. We have therefore decided to create a syntactic representation in which the information encoded in both of these structures is preserved.

In order to make use of the benefits of both, we are currently working on a complex syntactic representation for the sentences of the Szeged Treebank that utilizes the constituent and the dependency trees as well as the morphological analysis of the words. The new structure analyses different types of syntactic information at different levels, similar to Lexical-Functional Grammar. This multi-level syntactic representation is created by automatic conversion of the already existing constituent and dependency trees and the words' morphological analyses available for the sentences of the Szeged Treebank.

The phrase structures of the constituent analysis are represented here in a c-structure reflecting the surface structure of the sentences. These are converted directly from the constituent trees of the Szeged Treebank.

The sentences' argument structure is represented at a different level, in the f-structure. We convert these using the dependency trees and the morphological information on the words of the sentence.

The new database enables the training and evaluation of statistical syntactic parsers with a new approach, as well as testing these in real-world natural language processing tasks. Thus the usefulness of this multi-level syntactic representation can be empirically compared to that of the classical constituent and dependency analyses as well.

Analyzing Hungarian webtext

Viktor Varga[1], Vilmos Wieszner[1], Hangya Viktor[1],
Veronika Vincze[2], Richárd Farkas[1]

[1] University of Szeged, Department of Informatics
{viktor.varga.1991,vilmos.wieszner,hangyav}@gmail.com,
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

The Internet's role in people's lives is becoming more and more significant, especially due to its importance in modern communication. A large amount of data is generated by the users' communication through this medium, and this could be useful for a number of natural language processing applications, for example in information extraction and sentiment analysis. Thus analyzing webtext is gaining importance. Non-standard language use is the biggest difficulty in this context, which decreases the efficiency of language processing tools developed for standard texts.

In this paper, we focus on Hungarian webtexts. As Hungarian is the prototype of morphologically rich languages, we investigate the question whether the required adaptation techniques from standard texts to webtexts are similar to the ones introduced for English. We identified the most frequent error types of our linguistic analyzing toolchain for Hungarian (magyarlanc) and our Named Entity Recogniser on public facebook messages along with their comments and tweets. These tools were developed on the Szeged Treebank (i.e. on standard texts).

Imitating spoken language and therefore focusing on speed and the expression of emotions are part of the fundamental nature of social media texts. Speed is increased by quicker typing: diacritics, punctuations, whitespaces and capitals often disappear, abbreviations are used and typos are often made. Emotions may be expressed through the overuse of capitals and punctuations, or by emoticons. Explicit expression of hesitation, inventing words, and the use of English words and abbreviations are also frequent stylistic means. All these depend on the individual language use, registers and contexts.

Capitalization and punctuations cannot be used as guidelines in the segmentations of sentences, and the lack of whitespaces make word tokenization difficult. NER systems cannot handle lowercase names, while uppercase words are automatically detected as named entities. The morphological parser cannot analyze or assigns the wrong code to misspelt or unknown words, which affects the syntactic analysis as well. The differences between English and Hungarian make modifications based solely on English chat language insufficient, different solutions are required, e.g. phonetic transcription (*thru* instead of *through*) is more problematic for English texts due to the complexity of English orthography but the lack of accents (*kerek* vs. *kerék* vs. *kérek*) is only relevant for Hungarian. We propose the normalization of the input text, expansion of the lexica and domain-adaptation of current processing modules. We believe that the combination of all these methods could significantly increase performance.

Uncertainty Detection in Hungarian Texts

Veronika Vincze[1,2]

[1] MTA-SZTE, Research Group on Artificial Intelligence
[2] University of Szeged, Department of Informatics
vinczev@inf.u-szeged.hu

Distinguishing between factual (i.e. true or false) and uncertain propositions is essential both in linguistics and natural language processing applications. For instance, in information extraction (IE) many applications seek to extract factual information from text, and they should handle detected modified parts in a different manner. Due to this, uncertainty detection has received a considerable amount of attention in the last few years in the natural language processing community.

In this paper, we report on a Hungarian corpus – hUnCertainty – manually annotated for several types of linguistic uncertainty, which is – to the best of our knowledge – is the first one developed for Hungarian.

The hUnCertainty corpus contains paragraphs from the Hungarian Wikipedia. Hungarian equivalents of typical uncertainty cues in English were collected and paragraphs containing them were randomly sampled from the Hungarian Wikipedia dump. Besides, paragraphs which did not contain such words were also included in the corpus so as to avoid biased data.

The corpus is manually annotated for linguistic cues denoting several types of uncertainty. A sentence is epistemically uncertain if on the basis of our world knowledge we cannot decide at the moment whether it is true or false. As for hypothetical uncertainty, the truth value of the propositions cannot be determined either. This class contains conditionals and investigations, which is frequent in science papers where research questions are often stated in the form of this linguistic tool. Non-epistemic types of modality (such as doxastic modality – related to beliefs – or dynamic modality – related to e.g. necessities) also belong to this group.

Concerning discourse-level uncertainty, we annotated three classes. First, weasels are sourceless propositions or propositions with any underspecified argument that would be relevant or is not common knowledge in the situation. Second, hedges blur the exact meaning of some qualities or quantities. Third, peacocks express unprovable qualifications or exaggerations.

This corpus served as the training and test database for our CRF-based approach, which makes use of a rich feature set including orthographic, lexical, morphological, syntactic and semantic features as well. The results of our experiments show that uncertainty detection can be successfully carried out on Hungarian texts as well.

Morphological Modifications in Szeged Corpus 2.5

Veronika Vincze[1], Viktor Varga[2], Katalin Ilona Simkó[2],
János Zsibrita[2], Ágoston Nagy[2], Richárd Farkas[2]

[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
[2]University of Szeged, Department of Informatics
{vinczev,zsibrita,nagyagoston,rfarkas}@inf.u-szeged.hu
{viktor.varga.1991,kata.simko}@gmail.com

In this work, we present Szeged Corpus 2.5, in which we applied some morphological modifications which we believe will benefit real-world NLP applications. The modifications involve the introduction of new codes in the coding system as well as the correction of some morphological codes, with special emphasis on misspelled words.

Recently, there has been a successful attempt to harmonize the coding systems MSD and KR. The two coding systems cannot be mapped in a one-to-one way, so if we want to exploit both resources in a statistical language parser (POS tagger, constituency parser, dependency parser etc.), we have to employ conversion rules, which leads to the loss of information. In order to prevent this, the two coding systems (MSD and KR) were harmonized and their basic principles were also made compatible.

Here, we applied the principles of the harmonized morphology in the annotation of Szeged Corpus 2.5. For instance, only those pieces of derivational information are explicitly marked that are expressed with syntactic tools in other languages. We applied this approach to verbs with frequentative, modal and causative suffixes and the lemma became the word form without any of the above mentioned suffixes.

As for the treatment of adverbial pronouns, we decide to derive them from personal pronouns and thus inserted them into the pronominal system of morphological codes.

Present, past and future participles were also given a new code since in the earlier version of the corpus, they could not be distinguished on the basis of their codes, what is more, their code coincided with that of adjectives.

We also eliminated the differentiation between proper nouns and common nouns at the level of morphology. In addition to the morphological modifications described above, we also paid attention to the correction of misspelled words. All in all, changes involved about 4.36% of the tokens in the corpus.

These modifications also made it possible to train and evaluate the morphological analyzer and POS-tagger modules of magyarlanc on the new version of the corpus. According to our results, the accuracy of POS-tagging does not change significantly as compared to that achieved by training magyarlanc on Szeged Corpus 2.0.

Automatic Error Detection concerning the Definite and Indefinite Conjugation in Texts by Learners of Hungarian

Veronika Vincze[1], János Zsibrita[2], Péter Durst[3], Martina Katalin Szabó[4]

[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]University of Szeged, Department of Informatics
zsibrita@inf.u-szeged.hu
[3]University of Szeged, Hungarian Studies Center
durst.peter@gmail.com
[4]University of Szeged, Hungarian Linguistics PhD Programme
szabomartinakatalin@gmail.com

In this paper we focus on automatic error detection concerning the definite and indefinite conjugation in Hungarian, based on data from the HunLearner corpus.

The texts of HunLearner were POS-tagged and dependency parsed by magyarlanc, a linguistic preprocessing toolkit of Hungarian. On the basis of the syntactic and morphological analysis we were able to define rules for the object-verb agreement, which made it possible to collect those sentences where there was a mismatch between the definiteness of the object and the verbal conjugational pattern.

Here we just focused on cases where the object is phonologically present in the sentence, so we neglected cases when the presence of the pronominal object could be only deduced from the verbal form. We also neglected cases when the object was a subordinate clause.

Our results reveal grammatical structures that might pose problems for learners of Hungarian. The most frequent source of errors was when the object is a common noun with a definite article: it triggers definite conjugation but in 17% of the errors, it co-occurred with an indefinite verb. Other frequent errors are a demonstrative pronoun as the object and a bare common noun (i.e. without an article) as the object: in 13-13% of the errors, they do not co-occur with the required type of conjugation. Together with the errors induced by possessive forms, these types altogether are responsible for 50% of the mismatches in conjugation.

It is also shown that the definite object + indefinite conjugation (59%) is a much more frequent phenomenon than the opposite, i.e. indefinite object + definite conjugation.

Our results may be fruitfully applied in language teaching on the one hand as the statistical analysis makes it possible for the students to concentrate on grammatical structures that seem to give rise to more difficulties. On the other hand, from a natural language processing point of view, definiteness errors in conjugation may be automatically corrected as the automatic detection of the type of the object triggers the type of conjugation. If the sentence does not contain the required form, a grammar checker may automatically propose some corrections concerning the word form of the verb.

Névmutató