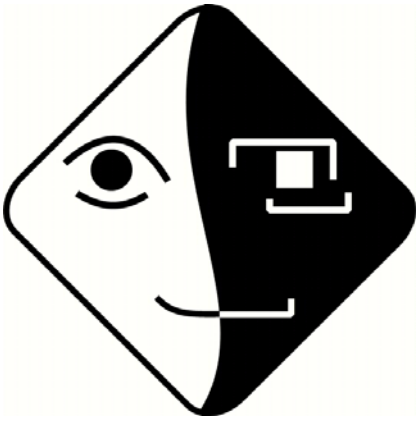
VIII. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2011

Szerkesztette:

Tanács Attila
Vincze Veronika

Szeged, 2011. december 1-2.
http://www.inf.u-szeged.hu/mszny2011

ISBN: 978-963-306-121-3

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

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

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

Szeged, 2011. november

Előszó

2011. december 1-2-án nyolcadik alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát. Nagy örömet jelent számomra, hogy a rendezvény fokozott érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A konferencia fő célja – a hagyományokhoz hűen – 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.

A konferenciafelhívásra szép számban beérkezett tudományos előadások közül a programbizottság 40-et fogadott el az idei évben, így 28 előadás és 12 poszter-, illetve laptopos bemutató gazdagítja a konferencia programját. A programban a magyar számítógépes nyelvészet teljes palettájáról találhatunk előadásokat a beszédtechnológiától kezdve a számítógépes szemantika és pragmatika területén át az információkinyerésig és gépi fordításig.

A korábbi évekhez hasonlóan idén is tervezzük a „Legjobb Ifjú Kutatói Díj” odaítélését, mellyel a fiatal korosztály tagjait kívánjuk ösztönözni arra, hogy kiemelkedő eredményekkel járuljanak hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz. A díj felajánlásáért az MTA Számítástechnikai és Automatizálási Kutatóintézetének tartozunk köszönettel.

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

Csirik János, a rendezőbizottság elnöke

Szeged, 2011. november

Tartalomjegyzék

I. Többnyelvűség

II. Korpusz, ontológia

III. Szintaxis, morfológia, névelem-felismerés

IV. Beszédtechnológia

V. Pszichológia, pragmatika, kognitív nyelvészet

VI. Szemantika

VII. Poszterek és laptopos bemutatók

I. Többnyelvűség

Többnyelvű dokumentum nyelvének megállapítása

Pataki Máté[1], Vajna Miklós[1]

[1] MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi utca 11.
{pataki.mate, vajna.miklos}@sztaki.hu

Kivonat: A cikkben egy olyan algoritmust ismertetünk, amely alkalmas arra, hogy gyorsan és hatékonyan megállapítsa egy szövegről nemcsak annak elsődleges természetes nyelvét, de többnyelvű szöveg esetén a második nyelvet is – mindezt szótár nélkül egy módosított n-gram algoritmus segítségével. Az algoritmus jól működik vegyes nyelvű, akár szótárként felépített, szavanként változó nyelvű dokumentumokon is.

1 Bevezetés

Egy digitális, természetes nyelven íródott dokumentum nyelvének megállapítására számos lehetőség van, és a szakma ezt a problémát nagyrészt megoldottnak tekinti [1][2][3], ugyanakkor a dokumentum nyelvének megállapítása nem mindig egyértelmű feladat.

A leggyakrabban használt algoritmusok igen jól működnek tesztdokumentumokon vagy jó minőségű, gondosan elkészített gyűjteményeken, ha lehet róluk tudni, hogy egy nyelven íródtak. Nekünk azonban szükségünk volt egy olyan algoritmusra, amely internetről letöltött dokumentumokon is jól – gyorsan és megbízhatóan – működik. A KOPI plágiumkereső programunk interneten talált, megbízhatatlan eredetű, gyakran hibás dokumentumokat dolgoz fel, és ennek során lényeges, hogy a dokumentum nyelvét, illetve főbb nyelveit megfelelően ismerje fel, azaz többnyelvű dokumentumok esetében is megbízhatóan működjön.

A jelenleg nyelvfelismerésre használt algoritmusok erre nem voltak képesek magukban, így az egyik algoritmust úgy módosítottuk, hogy amennyiben egy dokumentumban nagyobb mennyiségben található más nyelvű szöveg, akkor azt jelezze, és így a plágiumkereső rendszer ezt mint többnyelvű dokumentumot tudja kezelni.

Az algoritmussal szemben az alábbi elvárásokat támasztottuk:

1. Jelezze, ha a dokumentum több nyelven íródott, és nevezze meg a nyelveket
2. Az algoritmus gyors legyen
3. A szöveget csak egyszer kelljen végigolvasni
4. Ne szótár alapú legyen (kódolási és betanítási problémák miatt)

A legegyszerűbb megoldásnak az n-gram algoritmus tűnt [1][4], mivel ezen algoritmust használva csak egyszer kell végigolvasni a dokumentumot és az n-gram sta-

tisztikákból meg lehet állapítani, hogy a dokumentum milyen nyelven íródott, és – ha vannak megfelelő mintáink – még a kódolását is meg tudja határozni.

Az n-gram viszont nem teljesíti az első feltételt, miszerint a több nyelven íródott dokumentumokat is fel kell ismernie. Ugyan elméletileg elképzelhető lenne, hogy a dokumentumot szakaszokra osztjuk, és szakaszonként állapítjuk meg a dokumentum nyelvét, de ez a megoldás sajnos két esetben is hibás eredményre vezet. Gyakran találkoztunk olyan dokumentummal, amelyik úgy volt felépítve, mint egy szótár, azaz a két nyelv nem szakaszonként, hanem mondatonként – sőt egyes esetekben szavanként – váltakozott. A másik probléma akkor jelentkezett, amikor a dokumentum – például egy korábbi hibás konverzió miatt – tartalmazott HTML- vagy XML-elemeket, amelyek miatt rövid dokumentumok esetében hibásan angol nyelvűnek találta az algoritmus azokat.

Ezek kiküszöbölésére kezdtük el továbbfejleszteni az n-gram algoritmust, amely alapból csak arra alkalmas, hogy a dokumentumban leggyakrabban használt nyelvet megállapítsa, de a második leggyakoribb nyelv már nem a második a listában. Ennek oka, hogy a nyelvek hasonlítanak egymásra, és például egy nagyrészt olasz nyelvű dokumentum esetében a spanyol nyelv akkor is nagyobb értéket kap, mint a magyar, ha a dokumentum egy része magyar nyelven íródott.

Az új algoritmusunkba ezért beépítettünk egy nyelvek közötti hasonlósági metrikát, amelyet a hamis találatok értékének a csökkentésére használunk. A metrika segítségével meg lehet állapítani, hogy a második, harmadik... találatok valódiak-e, vagy csak két nyelv hasonlóságából fakadnak.

2 Az eredeti algoritmus

Az n-gram algoritmus működése igen egyszerű, legenerálja egy nyelvnek a leggyakoribb „betű n-gramjait”, azaz a például 1, 2, 3 betű hosszú részeit a szövegnek, majd ezeket az előfordulási gyakoriságuk szerint teszi sorba. A magyar nyelvben ez a 100 leggyakoribb n-gram az általunk használt tesztszövegben (_ a szóköz jele):

1. _	17. y	33. s_	49. er
2. e	18. _a	34. _m	50. f
3. a	19. b	35. _a_	51. ek
4. t	20. d	36. en	52. te
5. s	21. a_	37. ö	53. és
6. l	22. v	38. n_	54. _s
7. n	23. t_	39. _k	55. al
8. k	24. sz	40. j	56. ta
9. i	25. el	41. ._	57. í
10. r	26. ,	42. i_	58. _h
11. z	27. ,_	43. eg	59. _t
12. o	28. h	44. p	60. an
13. á	29. k_	45. _e	61. ze
14. é	30. .	46. u	62. me
15. g	31. et	47. le	63. at
16. m	32. gy	48. ó	64. l_

65. es	74. _é	83. ne	92. _A
66. ő	75. ny	84. os	93. _sz
67. y_	76. tá	85. ál	94. is
68. z_	77. c	86. _f	95. ve
69. tt	78. re	87. az	96. gy_
70. ke	79. to	88. zt	97. ít
71. _v	80. A	89. ár	98. _b
72. ás	81. e_	90. _n	99. ra
73. ak	82. ü	91. ko	100. or

Két szöveg összehasonlítása úgy történik, hogy a két n-gram listán összeadjuk az azonos n-gramok helyezéseinek a különbségét, és ez adja a két dokumentum közötti hasonlóság mértékét. Két azonos nyelven írt dokumentum között alig, míg különböző nyelvek között szignifikáns lesz a különbség. Ezért használható ez az algoritmus a dokumentum nyelvének megállapítására.

Példának nézzük meg az angol nyelvű példadokumentumunk első 10 n-gramját, és hasonlítsuk össze a magyarral.

1. _ (1-1)
2. e (2-2)
3. t (3-4)
4. o (4-12)
5. n (5-7)
6. i (6-9)
7. a (7-3)
8. s (8-5)
9. r (9-10)
10. h (10-28)

Az eredmény 0+0+1+8+2+3+4+3+1+18 = 40. Ez a különbség egyre nagyobb lesz, ahogy lejjebb megyünk a listában. Mivel nem lehet végtelen hosszú listát készíteni, így azokat az n-gramokat, amelyek az egyik listában szerepelnek, de a másikban nem, úgy vesszük figyelembe, mintha a lista utolsó helyén álltak volna. Mi egy 400-as listával dolgoztunk, azaz az első 400 n-gramot tároltuk el minden nyelvhez.

Ennek megfelelően a két nyelv elméleti minimális távolsága 0, maximális távolsága (r_{max}) pedig 400^2 azaz 160 000. Ebből a százalékos hasonlóságot a

$$h_{százalékos} = \left(r_{max} - r\right)/\left(r_{max}/100\right)$$

összefüggéssel kapjuk.

Példának nézzük meg, hogy mekkora hasonlóságot mutatnak különböző nyelvű dokumentumok a mintadokumentumainkhoz képest. Az egyszerűbb olvashatóság érdekében $h_{százalékos}$ értékekkel számolva a különböző nyelvű **Szeged Wikipédia-szócikkekre** [5][6][7][8][9].

A **magyar** nyelvű szócikk esetén az alábbi eredményt kapjuk, az első 5 találatot kérve:

1. magyar: 35.49
2. breton: 27.70
3. szlovák: 27.42
4. eszperantó: 26.98
5. közép-frízi: 26.79

Az **angol** nyelvű szócikk esetén az alábbi eredményt kapjuk:

1. angol: 44.37
2. skót: 35.67
3. romans: 35.34
4. német: 33.74
5. román: 33.73

A **német** nyelvű szócikk esetén az alábbi eredményt kapjuk:

1. német: 57.13
2. holland: 38.15
3. közép-fríz: 37.71
4. dán: 37.48
5. fríz: 36.58

Az **olasz** nyelvű szócikk esetén az alábbi eredményt kapjuk:

1. olasz: 35.21
2. román: 33.95
3. katalán: 33.46
4. spanyol: 32.18
5. romans: 31.78

Jól látható az eredményekből, hogy a barátságos nyelvek esetében magas hasonlóságot mutat a dokumentum a rokon nyelvekre, azaz egy olasz nyelvű dokumentum majdnem ugyanannyi pontot kap az olaszra, mint a spanyolra.

Most nézzük meg, hogy **kétnyelvű, 50-50 százalékban kevert dokumentumok**ra mit kapunk.

Egy **magyar-angol** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. angol: 40.80
2. magyar: 39.45
3. skót: 38.41
4. afrikaans: 34.69
5. közép-fríz: 34.19

Egy **magyar-olasz** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. olasz: 49.56
2. romans: 45.25
3. katalán: 41.60
4. latin: 41.26
5. román: 41.18

…

10. magyar: 38.02

Egy **magyar-francia** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. francia: 38.16
2. katalán: 36.74
3. eszperantó: 34.26
4. spanyol: 34.08
5. romans: 33.71

…

7. magyar: 33.2

Egy **angol-német** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. német: 53.47
2. angol: 44.14
3. fríz: 40.98
4. közép-fríz: 40.61
5. holland: 40.08

Látható, hogy a magyar-olasz, ill. magyar-francia kevert szövegben a magyar nyelv bele se került az első 5 találatba.

Végül nézzük meg, hogy egy háromnyelvű, harmadolt arányban kevert dokumentumra mit kapunk.

Egy **magyar-angol-olasz** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. angol: 46.55
2. olasz: 44.55
3. romans: 43.58
4. katalán: 42.41
5. román: 41.11

…

10. magyar: 38.26

Láthatjuk, hogy a háromnyelvű szövegben sem kerül be az első öt helyre a magyar nyelv.

3 Az új algoritmus

Mint láttuk, bizonyos nyelvek hasonlítanak egymásra az n-gram algoritmus szempontjából, így egy többnyelvű dokumentum esetén a második helyen nem minden esetben a dokumentum második nyelvét találjuk, ráadásul az se derül ki, hogy a második nyelv azért került oda, mert valóban szerepel a dokumentumban, vagy azért, mert hasonlít az első nyelvre. Ezért az új algoritmusunkban elkezdtük kiszámolni a **nyelvek közötti hasonlóság**ot, méghozzá a nyelvfelismeréshez használt n-gram minták közötti hasonlóságot. A távolságok tipikus értékeire nézzünk néhány esetet.

A **magyar** nyelvhez legközelebb álló nyelvek távolság-értékei:

1. breton: 104 541
2. közép-fríz: 104 751
3. svéd: 106 068

4. eszperantó: 106 469
5. afrikaans: 106 515

Az **angol** nyelvhez legközelebb állók:

1. skót: 85 793
2. francia: 88 953
3. katalán: 89 818
4. latin: 90 276
5. romans: 92 936

Végül az **olasz** nyelvhez legközelebb állók:

1. romans: 79 461
2. román: 85 232
3. katalán: 85 621
4. spanyol: 86 138
5. latin: 86 247

Számos algoritmussal próbálkoztunk, melyek közül az alább leírt bizonyult a legmegbízhatóbbnak.

Egy D dokumentumra kapott százalékos hasonlóságaink (hszázalékos), a százalékos hasonlóság mértékének növekvő sorrendjében legyen: h1, h2, h3 stb., a nyelveket jelölje L1, L2, L3, azaz a h1 a D dokumentum hasonlóságát mutatja az L1 nyelvű mintánkkal százalékban. A nyelvek közötti százalékos hasonlóságot pedig jelöljük hL1L2-vel. hi' legyen az új algoritmus által az Li nyelvre adott érték.

$$h_i' = h_i \quad ha \quad i = 1$$

$$h_i' = h_i - \frac{\sum_{k=1}^{i-1} h_k \times h_{LiLk}}{\sum_{k=1}^{i-1} h_i} \quad ha \quad i > 1$$

Az algoritmus tulajdonképpen minden nyelv valószínűségét csökkenti az előtte megtalált nyelvek valószínűségével, így kompenzálva a nyelvek közötti hasonlóságból adódó torzulást. Példának nézzük meg, hogy mekkora hasonlóságot mutatnak különböző nyelvű dokumentumok a mintadokumentumainkhoz képest ezzel az új algoritmussal számolva.

Egy **magyar** nyelvű dokumentum (Szeged Wikipédia-szócikke) esetén az alábbi eredményt kapjuk, az első 5 találatot kérve:

1. magyar: 35.49
2. kínai: 2.09
3. japán (euc jp): 1.81
4. koreai: 1.70
5. japán (shift jis): 1.58

Egy **angol** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. angol: 44.21
2. nepáli: 3.84
3. kínai: 2.53
4. vietnami: 2.08
5. japán: 1.14

Egy **német** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. német: 57.13
2. kínai: 2.55
3. japán (shift jis): 2.19
4. japán (euc jp): 1.93
5. nepáli: 1.27

Egy **olasz** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. olasz: 35.21
2. kínai: 1.07
3. perzsa: 0.68
4. japán: 0.57
5. jiddis: 0.55

Jól látható az eredményekből, hogy a barátságos nyelvek esetében a nyelvek hasonlóságából adódó hamis többletpontok kiszűrésre kerültek, azaz egy olasz nyelvű dokumentumnál a spanyol nyelv már meg se jelenik az első öt találatban. Most nézzük meg, hogy a **kétnyelvű, 50-50 százalékban kevert dokumentumok**ra mit kapunk.

Egy **magyar-angol** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. angol: 40.80
2. magyar: 9.40
3. thai: 1.54
4. armeniai: 1.39
5. koreai: 1.37

Egy **magyar-olasz** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. olasz: 49.56
2. magyar: 7.44
3. walesi: 2.31
4. breton: 1.92
5. ír: 1.68

Egy **magyar-francia** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. francia: 38.16
2. magyar: 2.11
3. thai: 1.42
4. koreai: 1.16
5. kínai: 0.70

Egy **angol-német** nyelvű dokumentum esetén az alábbi eredményt kapjuk:

1. német: 53.47
2. angol: 7.79
3. walesi: 2.08

4. fríz: 1.48
5. nepáli: 1.44

Látható például, hogy a magyar-olasz kevert szövegben a magyar nyelv immár a 2. helyre került, a korábbi – eredeti algoritmus által megadott – 10. helyről.

A kétnyelvű dokumentumok esetében nem mindegy, hogy a nyelvek milyen arányban keverednek, érthető módon egy bizonyos arány felett az egyik nyelv n-gramjai elnyomják a másikét. Ezt egy **angol-magyar** dokumentumsorozat segítségével nézzük meg. Az egyes részek aránya a 9 dokumentum során a 10% angol, 90% magyar összetételről 90% angol és 10% magyar összetételre változott:

10% angol, 90% magyar: 1. magyar: 38.01 2. koreai: 1.53 3. thai: 1.20 4. japán (euc): 1.14 5. japán (shift): 1.09	40% angol, 60% magyar: 1. angol: 37.62 2. magyar: 5.41 3. japán (euc): 1.47 4. thai: 1.46 5. japán (shift): 1.45	70% angol, 30% magyar: 1. angol: 44.92 2. vietnámi: 1.74 3. mingo: 1.67 4. kínai: 1.46 5. armén: 1.36
20% angol, 80% magyar: 1. magyar: 37.93 2. thai: 1.18 3. koreai: 1.17 4. japán: 1.16 5. armén: 1.11	50% angol, 50% magyar: 1. angol: 40.93 2. magyar: 5.30 3. thai: 1.49 4. japán (shift): 1.47 5. japán (euc): 1.37	80% angol, 20% magyar: 1. angol: 46.56 2. vietnámi: 2.07 3. mingo: 2.00 4. japán: 1.47 5. walesi: 1.43
30% angol, 70% magyar: 1. magyar: 37.47 2. angol: 4.91 3. thai: 1.22 4. armén: 1.18 5. japán: 1.16	60% angol, 40% magyar: 1. angol: 41.66 2. magyar: 3.43 3. kínai: 1.50 4. vietnámi: 1.48 5. mingo: 1.45	90% angol, 10% magyar: 1. angol: 48.1 2. vietnámi: 1.51 3. nepáli: 1.40 4. thai: 1.05 5. kínai: 1.05

A fenti táblázat csak egy példa, de a többi nyelvpárra is hasonló eredményeket kaptunk. Látható, hogy az algoritmus 30% körül kezd el hibázni, azaz akkor találja meg megbízhatóan a második nyelvet, ha az a szöveg több mint 30%-át teszi ki.

Hasonló eredményt kapunk egy **háromnyelvű**, harmadolt arányban kevert, **magyar-angol-olasz** nyelvű dokumentum esetén is:

1. angol: 46.55
2. magyar: 7.59
3. olasz: 6.18
4. breton: 3.11
5. skót: 2.85

Láthatjuk, hogy a háromnyelvű szövegben az első három helyen szerepelnek a valós nyelvek, de azért itt el kell mondani, hogy ez csak az egyenlő arányban kevert háromnyelvű dokumentumok esetén működik jól. Ha ez az arány eltolódik, akkor gyorsan kieshet egy-egy nyelv. Tapasztalatunk szerint az új algoritmus három nyelvet már nem talál meg megbízhatóan, így ilyen dokumentumok tömeges előfordulása esetén más algoritmust ajánlott választani.

5 Konklúzió

Ahhoz, hogy megállapítsuk, egy dokumentum egy vagy több nyelven íródott-e, kell választanunk egy olyan értéket, ami felett azt mondjuk, hogy a második nyelv is releváns, azaz a dokumentum többnyelvű. Ezt az értéket a tesztek alapján 4-nek választottuk, azaz 4-es érték felett jelezzük csak ki a nyelveket. Ez az érték a felhasználási igényeknek megfelelően választható. Akkor érdemes valamivel alacsonyabbra állítani, ha mindenképp észre szeretnénk venni, ha a dokumentum kétnyelvű, ha pedig csak igazán nagy idegen nyelvű részek érdekelnek, és nem okoz gondot a hibásan egynyelvűnek talált dokumentum, akkor állíthatjuk akár magasabbra is.

Ezzel a paraméterrel az algoritmust részletesen teszteltük a plágiumkeresőnkbe feltöltött dokumentumokon, és a vele szemben támasztott igényeknek messzemenőkig megfelelőnek találtuk. Ki tudtuk szűrni vele a rosszul konvertált és többnyelvű dokumentumok több mint 90%-át. A tesztek befejezése után az új algoritmust beépítettük a KOPI Plágiumkereső rendszerbe, ahol a korábbi, kevésbé pontos eredményt adó algoritmust váltotta ki.

Bibliográfia

1. Cavnar, W. B.; Trenkle, J. M.: N-Gram-Based Text Categorization. Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval. UNLV Publications/Reprographics, Las Vegas, NV, (1994) 161-175
2. Řehůřek, R.; Kolkus, M.: Language Identification on the Web: Extending the Dictionary Method. In: 10th International Conference on Intelligent Text Processing and Computational Linguistics (2009)
3. Benedetto, D.; Caglioti, E.; Loreto. V.: Language trees and zipping. Physical Review Letters Vol. 88, No. 4 (2002)
4. Dunning, T.: Statistical Identification of Language. Technical Report MCCS 94-273, New Mexico State University (1994)
5. Wikipedia: Szeged szócikk magyar nyelven, http://hu.wikipedia.org/wiki/Szeged (2011)
6. Wikipedia: Szeged szócikk angol nyelven, http://en.wikipedia.org/wiki/Szeged (2011)
7. Wikipedia: Szeged szócikk német nyelven, http://de.wikipedia.org/wiki/Szeged (2011)
8. Wikipedia: Szeged szócikk olasz nyelven, http://it.wikipedia.org/wiki/Seghedino (2011)
9. Wikipedia: Szeged szócikk francia nyelven, http://fr.wikipedia.org/wiki/Szeged (2011)

Statisztikai gépi fordítási módszereken alapuló egynyelvű szövegelemző rendszer és szótövesítő

Laki László János[1]

Pázmány Péter Katolikus Egyetem, ITK,
1083, Budapest, Práter u. 50/a,
e-mail: laki.laszlo@itk.ppke.hu

Kivonat Jelen munkában az SMT módszer alkalmazhatóságát vizsgáltam szófaji egyértelműsítő és szótövesítő feladat megoldására. Létrehoztam egy alaprendszert, illetve további lehetőségeket próbáltam ki a rendszer eredményeinek javítására. Megvizsgáltam, milyen hatást gyakorol a célnyelvi szótár méretének változtatása a rendszer minőségére, továbbá megoldást kerestem a tanító halmazban nem szereplő szavak elemzésének megoldására.

Kulcsszavak: Statisztikai Gépi Fordítás (SMT), szófaji egyértelműsítés (POS tagging), szótövesítés, Szeged Korpusz, OOV

1. Bevezetés

Az informatika fejlődése szinte az összes tudományág számára új lehetőségek halmazát nyitotta meg, és ez nincs másképp a nyelvészetben sem. Napjaink számítógépei segítségével képesek lettünk óriási méretű szöveges anyagok gyors és hatékony kezelésére, feldolgozására. A szövegek szintaktikai és/vagy szemantikai információval történő jelölése, valamint a szavak szófaji elemzése rendkívül fontos feladat a számítógépes nyelvészet számára. A szófaji egyértelműsítés problémája korántsem megoldott, annak ellenére, hogy sokféle rendszer létezik ennek implementálására. A legelterjedtebbek a gépi tanuláson alapulnak, melyek maguk ismerik fel a szabályokat a különböző nyelvi jellemzők segítségével. További nehézséget jelent azonban ezen jellemzők meghatározása, hiszen a különböző sajátosságok nehezen fogalmazhatók meg.

Ezzel szemben a statisztikai gépi fordító (SMT) rendszerek előzetes nyelvi ismeret nélkül képesek a fordításhoz szükséges szabályok felismerésére. Kézenfekvő megoldásnak tűnik SMT rendszerek alkalmazása szövegelemzésre. Munkám során az ebben rejlő lehetőségeket vizsgáltam a szófaji egyértelműsítés és szótövesítés feladatának megoldására.

2. A szófaji egyértelműsítés

Szófaji egyértelműsítés az a folyamat, amely a szövegben található szavakat általános lexikai jelentésük és kontextusuk alapján megjelöli a megfelelő POS cím-

kével. Egy helyesen címkézett mondatban minden szóhoz pontosan egy címke van rendelve. Ennek ellenére a szófaji egyértelműsítés sokkal komplexebb feladat egy szó és címkéjének listájából való kikereséshez képest, mivel egy szónak több szófaji alakja is lehet.

Erre a feladatra létrehozott első megoldások előre megírt szabályrendszerek segítségével elemezik a szöveget. A probléma ezekkel a rendszerekkel a szabályok létrehozásának magas költsége volt. Napjaink elterjedt rendszerei gépi tanuláson alapuló módszereket használnak, amelyek különböző nyelvi jellemzők segítségével maguk ismeri fel a szabályokat, ám a megfelelő jellemzők meghatározása szintén nehéz feladat. A különböző nyelvi sajátosságok nehezen fogalmazhatók meg és állíthatók össze olyan teljes, mindent magába foglaló szabályrendszerré, mely a számítógép számára feldolgozható. Ilyen nyelvi sajátosságok lehetnek például a nyelvek közötti fordítás szabályai, valamint a morfológiai elemzés.

A szófaji egyértelműsítők teljesítményének egyik nagyon fontos tényezője a tanítóhalmazban nem szereplő szavak (OOV: out-of-vocabulary) elemzése. Az OOV szavak elemzése nagyban függ az elemzendő nyelvtől. Például az angol nyelv esetében nagy valószínűséggel az OOV szavak tulajdonnevek lesznek. Ezzel szemben más nyelvek esetében – mint a magyar vagy a mandarin kínai – az OOV szavak főnevek és igék is lehetnek.[1]

2.1. A szótövesítés

Lemmatizálás számítógépes nyelvészeti szempontból az az algoritmikus folyamat, amelyik meghatározza egy szó szótári alakját. Napjainkban több megvalósítás is létezik ezen feladat megoldására (például: HUMOR [2]), de ezek általában bonyolult módszereket alkalmaznak. Ezzel szemben az SMT rendszeren alapuló szótövesítés előzetes nyelvtani ismeret nélkül végzi el ezt a feladatot.

2.2. Létező megvalósítások

Oravecz és Dienes 2002-ben készítették el az első magyar nyelvű sztochasztikus POS-taggert. A rendszer MSD-kódokat használ és 98.11%-os pontosságot ért el [3].

Halácsy et al. létrehoztak egy maxent modellen alapuló szófaji egyértelműsítőt. Csoportjával 2007-ben létrehozták a HunPOS nevű rendszert, ami napjaink legjobb magyar nyelvű POS-taggerjének számít. A rendszer MSD-kódokat használ és 98.24%-os pontosságot ért el [4].

3. Statisztikai gépi fordítás

A statisztikai nyelvfeldolgozás elterjedt alkalmazása a gépi fordítás. A statisztikai gépi fordító (SMT) módszer nagy előnye a szabályalapú fordítással szemben, hogy az architektúra létrehozásához nem szükséges a nyelvek grammatikájának ismerete. A rendszer tanításához csupán egy kétnyelvű korpuszra van szükség, amelyből statisztikai megfigyelésekkel nyerjük ki a szabályokat. A fordítás során

az egyetlen, amit biztosan tudunk, az a mondat, amit le szeretnénk fordítani (forrásnyelvi mondat). Ezért a fordítást úgy végezzük, mintha a célnyelvi mondatok halmazát egy zajos csatornán átengednénk, és a csatorna kimenetén összehasonlítanánk a forrásnyelvi mondattal.

$$\hat{E} = \underset{E}{argmax}\ p(E|F) = \underset{E}{argmax}\ p(F|E) * p(E) \tag{1}$$

Az a mondat lesz a rendszerünk kimenete ($\hat{E}$), amelyik a legjobban hasonlít a fordítandó (forrásnyelvi) mondatra. Ez a hasonlóság lényegében egy valószínűségi érték, amely a nyelvi modellből $p(E)$ és a fordítási modellből $p(F|E)$ számolható. Lásd az 1. egyenletben.

4. A POS-Tagging probléma mint SMT-probléma

Amint a bevezetőben már említettem, a szövegelemzés is megfogalmazható fordítási feladatként. Egy tetszőleges mondat (F) szófaji elemzése ($\hat{E}$) megfogalmazható a következő egyenlettel:

$$\hat{E} = \underset{E}{argmax}\ p(E|F) = \underset{E}{argmax}\ p(F|E) * p(E) \tag{2}$$

ahol $p(E)$ a címkék nyelvi modellje és $p(E|F)$ a fordítási/elemzési modell. A fordítási feladathoz hasonlóan a forrásnyelvi mondatot kifejezések halmazának tekintjük, ahol minden frázist a címkék egy halmazára „fordítunk”. Egy természetes nyelvek közti fordításhoz képest a szófaji egyértelműsítés egyszerűbb az SMT-rendszerek számára, hiszen nincs szükség a mondatban elhelyezkedő szavak sorrendjének megváltoztatására. A fordítás során a forrásnyelvi és célnyelvi oldal szavainak száma is megegyezik, azaz a rendszer nem végez elembeszúrást és törlést.[1,5] Ezen tulajdonságok miatt az SMT-rendszer jól alkalmazható megvalósításnak tűnik szófaji egyértelműsítésre.

5. Munkám során alkalmazott rendszerek

A következő fejezetben bemutatom a munkám során alkalmazott keretrendszereket.

5.1. MOSES

Több módszert is megvizsgáltam, melyek képesek párhuzamos korpuszból információt kinyerni. Végül az IBM modellek mellett döntöttem, mivel hatékony, viszonylag pontos, és a feladatnak nagyon jól megfelelő algoritmusnak bizonyultak. Ezért kezdtem használni a Moses keretrendszert [6,7,8], amely implementálja ezeket a modelleket. Ebben a rendszerben megtalálható a párhuzamos korpusz előfeldolgozása, a fordítási és nyelvi modellek létrehozása, a dekódolás, valamint a BLEU-metrikára való optimalizálás.

5.2. Joshua

Másfelől a Joshua keretrendszert [9] használtam, mely nem pusztán szó- vagy frázisszintű statisztikai valószínűségi modelleket használ, hanem bizonyos nyelvtani jellemzők előfordulását is figyelembe veszi. A Joshua rendszer további nagy előnye, hogy képes ezen generatív szabályok közti fordításra oly módon, hogy megadhatóak a szabályok mind a forrásnyelvre, mind a célnyelvre, valamint az is definiálható, hogy mekkora valószínűséggel transzformálhatók át a szabályok egymásba.

5.3. Korpusz

Az SMT-rendszer tanításához szükséges kétnyelvű párhuzamos korpuszt, a Szeged Korpusz 2.0-t használtam. A korpusz előnyei, hogy a szavak MSD-kódolású POS-címkéi mellett azok szótövei is szerepelnek benne, általános témájú, valamint készítői kézzel ellenőrizték annak helyességét. Hátránya, hogy viszonylag kis méretű. Mivel a szófaji címkék elemszáma korlátozott, ezért elvben kisebb méretű korpuszban is elég nagy gyakorisággal szerepelhetnek. [10,11]

5.4. Kiértékelő módszerek

A rendszer minőségének kiértékeléséhez a BiLingual Evaluation Understudy (BLEU) módszert használtam, amely egy gyakran alkalmazott módszer az SMT-rendszerek minőségének vizsgálata. Lényege, hogy a fordításokat referenciafordításokhoz hasonlítja, majd hozzájuk egy 0 és 1 közötti valós értéket rendel. Ezt BLEU-értéknek nevezzük. Tanulmányomban ennek százalékosított formáját használom. [12]

Másfelől egy Levenshtein távolságon alapuló automatikus módszer segítségével kiszámítottam az elemző rendszer pontosságát a mondatok és a tokenek szintjén egyaránt.

6. Eredmények

6.1. Az alaprendszer létrehozása

Az első betanítás. Mint már korábban említettem, az SMT-rendszer betanításához egy párhuzamos korpusz szükséges. A Szeged Korpusz 2.0-ből állítottam elő az általam használt forrásnyelvi és célnyelvi korpuszokat. Az előbbibe az eredeti, elemzetlen és tokenizált mondatokat tettem, míg az utóbbiba a mondatban szereplő szavak szótövei, valamint azok POS-címkéi kerültek. Az így kapott rendszer eredményei az 1. táblázatban szerepelnek.

A kiértékelésénél szembetűnt a rendszer egy súlyos hibája, miszerint az elemzett korpuszban egymás után szerepelnek a szavak szótövei, amikhez hozzákapcsolódnak az elemzést tartalmazó címkék, de a több tagból álló kifejezések esetében (pl.: többtagú tulajdonnevek, igei szerkezetek) a címke csak a kifejezés utolsó szaván, vagy utána helyezkedik el. Az egy szófaji egységbe tartozó kifejezések

1. táblázat. A 6.1. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.97%	90.29%	9.71%
JOSHUA	90.96%	91.02%	8.08%

jelölésének hiánya a statisztikai módszerben félrevezető fordítási modellt eredményez. Ennek köszönhetően a rendszer az elemzett szöveghez véletlenszerűen hozzáad címkéket, ezért gyengébb eredményt ért el.

Az önálló POS-címkék eltávolítása. Az eredmény javítása érdekében minden önálló címkét hozzácsatoltunk az előtte álló szóhoz, így kaptuk a 2. táblázatban látható eredményeket.

2. táblázat. A 6.1. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.97%	90.80%	9.20%
JOSHUA	90.96%	90.72%	9.28%

A 2. táblázatból látszik, hogy változatlan BLEU-értékek mellett a rendszer pontossága 0,5–0,6 százalékkal javult. Ezt annak köszönhetjük, hogy nem kerültek a fordításba felesleges elemek. Ennek ellenére a többtagú kifejezések fordítása továbbra sem megoldott.

A többtagú kifejezések kezelése. Többtagú kifejezések esetében a nehézség abból adódik, hogy mivel a rendszer szavakat elemez, így az összetett kifejezések részeit is külön-külön címkézi. Célom, hogy az elemző egy egységként kezelje a többtagú kifejezéseket. A probléma megoldásához elengedhetetlen ezeknek a kifejezéseknek az összekapcsolása például a tulajdonnevek felismerésével. Nem volt célom ilyen rendszer kifejlesztése, viszont az elmélet igazolása érdekében összekötöttem a korpuszban ezeket a kifejezéseket. A tanítás után a 3. táblázatban látható eredményt kaptam.

Az 1500 mondatos teszthalmazból számszerűsítve 506 mondat elemzése volt teljesen helyes és 994-ben volt valamilyen hiba. Első ránézésre ez rossznak tűnhet, de ha az eredményt címkék szintjén is megvizsgáljuk, sokkal jobb arányt kapunk, hiszen 24557 helyes és csak 2343 helytelen elemzést kaptam. Láthatjuk, hogy a 6.1 rendszerhez képest a többtagú kifejezések összekötése és egyként kezelése javított a rendszer pontosságán, annak ellenére, hogy rosszabb BLEU-eredményt kaptam.

3. táblázat. Az alaprendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.76%	91.29%	8.71%
JOSHUA	90.77%	91.07%	8.93%

Az eredmények mélyebb vizsgálatából kiderül, hogy a helytelen annotációnak két oka lehet. Az első, amikor a szó nem szerepel a tanító halmazban (out-of-vocabulary, OOV), ekkor a rendszer elemzetlenül adja vissza a forrásnyelvi kifejezést. Ez 1697 esetben fordult elő. A helytelen annotációk másik típusa, amikor az SMT rendszer helytelen címkét rendel az adott szóhoz (646 eset). Ennek további két csoportja lehet: egyrészt, amikor a megfelelő szófaji címkét megtalálja, viszont a mélyebb szintű elemzés során hibázik; másrészt amikor teljesen rosszul elemzi a szót.

A 4. táblázatban egy példamondat olvasható a 6.1. rendszer kimenetéből.

4. táblázat. Példamondat az alaprendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	ez_[pd3-sa] a_[tf] lobbyerőt_[x] és_[ccsw] képesség_[nc-sa] a_[tf] diplomáciai_[afp-sn] erőfeszítés_[nc-pp] kívül_[st] mindenekelőtt_[rx] a_[tf] magyarországi_[afp-sn] multinacionális_[afp-pn] adhat_[vmcp3p—y] ._[punct]
SMT elemző:	ez_[pd3-sa] a_[tf] lobbyerőt és_[ccsw] képesség_[nc-sa] a_[tf] diplomáciai_[afp-sn] erőfeszítéseken kívül_[st] mindenekelőtt_[rx] a_[tf] magyarországi_[afp-sn] multinacionális_[afp-pn] adhat_[vmcp3p—y] ._[punct]

Továbbiakban ezt a rendszert fogom alaprendszernek tekinteni. A továbbiakban vizsgált rendszereknél kikötés lesz, hogy a fent említett hibákat elhagyjam, vagyis ne álljanak önmagukban címkék, illetve a többtagú kifejezések össze legyenek kötve.

6.2. A célnyelvi szótár méretének csökkentése

Csak szófaji egyértelműsítés. Az SMT-rendszer tulajdonságaiból következik, hogy egy megfelelő korpuszból bármilyen szabály betanítható. Mivel az általam használt korpusz mérete korlátos, a rendszer minőségének javulása többek között elérhető az annotációs feladat komplexitásának csökkentésével. Ebben az esetben ezt úgy érhetem el, ha az elemzendő szöveget a POS-címkék „nyelvére" fordítom.

Ezt munkám során úgy valósítottam meg, hogy az elemző rendszeremből elhagytam a szótövesítést, és csak a szófaji egyértelműsítést alkalmaztam. Mivel ezáltal csak a szavak POS-tag-jeire fordítok, a célnyelvi oldal szótári elemeinek száma nagy mértékben csökken. Az alaprendszer esetében 152694 elemből állt a célnyelvi szótáram, ezt csökkentettem le 1128 elemre. Így a fordítási feladat bonyolultságát csökkentve egy relatíve pontos rendszer hozható létre kis korpuszból is. Másrészről a szótövek elhagyásával csak címkék halmazára fordítok, ezáltal az egyes címkék nagyobb súllyal szerepelnek, mind a fordítási, mind pedig a nyelvi modellben. A tanítás után az 5. táblázatban látható eredményt kaptam.

5. táblázat. A 6.2. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	89.01%	91.46%	8.54%
JOSHUA	88.57%	91.09%	8.91%

A rendszer eredményeit vizsgálva kiderült, hogy a BLEU-érték további csökkenésének ellenére a rendszer pontossága jobb lett. Itt már az 518 teljesen helyes mondat mellett 982 mondat volt helytelen (0.8%-os javulás az alaprendszerhez képest). Tokenek szintjén 24603 volt helyes és 2297 volt helytelen (0.17%-os javulás). Ebből a rendszer által nem elemzett szavak száma 1699, amely változatlan az alaprendszerhez képest. Ezekből az eredményekből világosan látszik, hogy a rendszer minőségének javulása abból adódik, hogy az alaprendszer által elrontott 646 elemzés az új rendszerben 598-ra csökkent. Az eredmények mélyebb vizsgálata során szembetűnt, hogy e mögött a 48 darabos javulás mellett több eddig helyes elemzés romlott el. Ilyen hiba például a határozószók és a kötőszók keverése, valamint a kötőszók és a mutató névmások tévesztése. A 6. táblázatban egy példamondat olvasható a 6.2. rendszer kimenetéből.

6. táblázat. Példamondat a 6.2 . rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] lobbyerőt [ccsw] [nc-sa] [tf] [afp-sn] erőfeszítéseken [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]

A POS címkék egyszerűsítése. Az előző (6.2) fejezet eredményeiből kiindulva megvizsgáltam, hogy a célnyelvi szótár további csökkentése milyen hatást gyakorol a rendszer minőségére. Annak érdekében, hogy megvizsgáljam a rendszer működését a lehető legegyszerűbb körülmények között, hogy az elemzési mélységet nagy mértékben csökkentettem.

Ezt a következő rendszer segítségével tanulmányoztam oly módon, hogy csak a fő szófaji címkéket (az MSD-kód első karaktereit) hagytam meg a célnyelvi szótárban. Ebben az esetben a célnyelvi szótár 14 elemből áll. A tanítás után a 7. táblázatban látható eredményt kaptam.

7. táblázat. A 6.2. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	79.57%	92.20%	7.80%

A rendszer kiértékeléséből kiderült, hogy az eddig megfigyelt tendencia folytatódik. Tehát amíg a BLEU-érték csökkent, a rendszer pontossága növekedett. Ebben az esetben a rendszer 553 mondatot elemzett helyesen, miközben 947-et rontott el. Ez a 6.2. rendszerhez képest 2.3%-os, míg az alaprendszer (6.1) esetében 3.1%-os növekedést jelent mondatok szintjén. Tokenek tekintetében 24803 volt helyes és 2097 volt helytelen elemzés, ami 0.74%-os javulás a 6.2. rendszerhez képest, illetve 0.88% az alaprendszerhez képest. A 8. táblázatban egy példamondat olvasható a 6.2. rendszer kimenetéből.

8. táblázat. Példamondat a 6.2. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	p t x c n t a n s r t a a v p
SMT elemző:	p t lobbyerőt c n t a erőfeszítéseken s r t a a v p

Konklúzió. A fent elért eredmények rendkívül bíztatóak, mivel egy viszonylag kisméretű korpusz esetén is az elemző rendszerek pontossága 90% feletti. Érdemes megfigyelni, hogy a 6.2. rendszer szótára két nagyságrenddel kevesebb elemet tartalmaz (1128 darab címke) az alaprendszeréhez képest (152 694 darab címke), ennek ellenére pontossága csupán 0.17%-al javult. Továbbá megfigyelhető, hogy a 6.2. rendszer csupán 14 címkéből álló szótára esetén (ami négy nagyságrend-

del való csökkentést jelent az alaprendszerhez képest) is csak 0.88%-os javulás mutatkozott.

Értékelésem szerint ez a 0.88%-os minőségjavulás nem áll arányban azzal a hatalmas információveszteséggel, amely a rendszerek célnyelvi szótárméretének csökkentésével jött létre. További tanulság, hogy a célnyelvi szótár méretének változtatásától függetlenül az OOV szavakat (1698 darab) egyik rendszernek sem sikerült elemeznie. Ebből arra a következtetésre jutottam, hogy a rendszer eredményének további javulása érdekében megoldást kell találnom a tanítóhalmazban nem szereplő szavak kezelésére.

6.3. Az OOV szavak kezelése

Az első, legkézenfekvőbb megoldás a korpusz növelése. A tanító halmazban minél több token fordul elő, annál pontosabb lesz a rendszer. A magyar nyelv agglutináló tulajdonságából adódóan, azért, hogy minden token megfelelő számban forduljon elő a korpuszban, nagyon nagy méretű korpuszra lenne szükség. A következő fejezetben egy olyan módszert vizsgálok, amely alkalmas lehet az OOV szavak kezelésére.

Sima szöveg esetén. Mivel az OOV szavak elemzéséhez a tanító halmazból semmilyen információt nem nyertünk ki, szükségünk van ezen szavak további vizsgálatára. Ebben segítségünkre lehet az ismeretlen szavak kontextusa. A nyelvi sajátosságok, valamint a zárt és nyílt szóosztályok miatt az OOV szavak nagy valószínűséggel csak egy-két szófaji osztályból kerülnek ki. Az előző rendszerek megfigyelése alapján elmondható, hogy a szótárban nem szereplő szavak túlnyomórészt főnevek.

Guillem és Joan Andreu módszere alapján [1] ezt a problémát úgy próbálom meg kiküszöbölni, hogy azokból a szavakból, melyek a tanító halmazban egy bizonyos küszöbértéknél gyakrabban fordulnak elő, egy szótárat hozok létre. Azokat a szavakat, amelyek nem kerülnek be ebbe a szótárba, egy tetszőleges (az esetemben „UNK") kifejezésre cserélem ki. Így ez a szimbólum nagy gyakorisággal kerül be az elemzendő szövegbe. Feltételezésem szerint, mivel az OOV szavak csak egy-két szófaji osztályból kerülnek ki, a környezetükben lévő szófaji szerkezetek nagyon hasonlóak lesznek. Mivel az SMT rendszer kifejezés alapú fordítást végez, figyelembe veszi mind az elemzendő szavak, mind a címkék környezetét is. Ennek segítségével tudja meghatározni az „UNK" szimbólum elemzését.

Kulcsfontosságú kérdés a megfelelő gyakorisági szint kiválasztása, hiszen ettől függ, hogy mennyi „UNK" szimbólum kerül a korpuszba. Egyrészt, ha túl nagy ez a szám, akkor túl sok token cserélődik ki az „UNK" szimbólumra, emiatt a környezet vizsgálatából sem kapunk megbízható elemzést, hiszen abban is előfordulhat nagy valószínűséggel „UNK". Másrészt viszont ha túl kicsi, akkor túl sok ritka szó marad a szótárban, ezzel nem tudjuk megfelelő mértékben kihasználni a módszer előnyét. Rendszeremben ezt a gyakorisági küszöböt 10-re választottam.

A fentiek alapján felépített rendszer betanítása után a 9. táblázatban látható eredményt kaptam.

9. táblázat. A 6.3. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	88.71%	85.74%	14.26%

Szembetűnő változás, hogy a rendszer eredménye nagymértékben romlott. Csupán 294 mondatot sikerült teljesen hibátlanul elemeznie a rendszernek, míg 1206-ban fordult elő valamilyen hiba. Tokenek szintjén 23064 volt helyes és 3836 volt helytelen. A 10. táblázatban egy példamondat olvasható a 6.3. rendszer kimenetéből.

10. táblázat. Példa mondat a 6.3. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a unk és unk a diplomáciai unk kívül mindenekelőtt a magyarországi unk unk .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct] [pd3-sa] [tf] [nc-sa] [ccsp] [vmis3p—y] [tf] [afp-sn] [nc-pn] [st] [rx] [tf] [afp-sn] [nc-pn] [nc-sa—s3] [punct]

A magyar nyelvű szövegben a főnevek és az igék különböző ragozott formái találhatók meg, melyek kis korpusz miatt nagy valószínűséggel az általam alkalmazott küszöb alá esnek. Ez magyarázza, hogy a korpuszban szereplő mondatok többségében a főnevek és az igék helyére is az „UNK” szimbólum kerül, ami a szóösszekötő munkáját nehezíti meg. Ez okozta, hogy a rendszer elrontotta az eddig helyes mondatelemzéseket is, ráadásul előfordult, hogy összekeverte a szavak sorrendjét az elemzés során.

Szótövek esetén. Az előző rendszer hibáinak kiküszöbölésére megvizsgáltam, hogyan befolyásolja a rendszer eredményét, ha a gyakoriságot nem a szövegben megtalálható szavakra, hanem azok szótöveire vizsgálom. Ettől azt vártam, hogy így csak azokat a szavakat/szótöveket cserélem „UNK”-ra, amelyek előfordulása tényleg nagyon alacsony. A két rendszer összehasonlításának érdekében ebben az esetben is 10-re választottam a küszöbértéket. A 11. táblázatban látható eredményt kaptam.

Az eredmények elemzése során az előző rendszer (6.3) eredményéhez képest viszonylag nagy javulás figyelhető meg, bár ez az alaprendszer (6.1) eredményét még mindig nem éri el. A rendszer 450 helyes mondat mellett 1050-et ront el. Tokenek szintjén 24190 volt helyes és 2710 volt helytelen.

11. táblázat. A 6.3. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.87%	89.93%	10.07%

A fent említett változtatások hatására valóban csak az igazán ritka szavak lettek lecserélve „UNK”-ra. Ezek többsége nagyrészt főnév, és már alig van köztük ige. Ezzel párhuzamosan viszont az igék esetében egyre gyakoribb jelenség, hogy az elemző OOV szóként elemezte őket. Ez abból adódik, hogy ragozott formájuk nem szerepel a tanító halmazban megfelelő súllyal. A 12. táblázatban egy példamondat olvasható a 6.3. rendszer kimenetéből.

12. táblázat. Példamondat a 6.3. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a unk és képességet a unk erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] [nc-sa] [ccsw] [nc-sa] [tf] [afp-sn] erőfeszítéseken [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]

7. Összefoglalás

Kutatásom során az SMT-rendszer lehetőségeit vizsgáltam a szófaji egyértelműsítés és a lemmatizálás feladatainak megvalósítására. Megfigyelésem szerint ezek a problémák megfogalmazhatók a sima szövegről elemzett szövegre való fordításként is. Az erre a célra használt rendszerek pontossága elérheti akár a 92%-ot is. Annak ellenére, hogy ez az eredmény nem éri el a napjaink legjobb POS-tagger rendszerének szintjét, az általam felépített rendszer teljesen automatikusan ismeri fel a szabályokat, és nincs szükség előzetes szövegfeldolgozásra. Másrészt ez a rendszer párhuzamosan végzi az annotálás és a lemmatizálás feladatát. Az itt elvégzett kísérletekkel bebizonyítottam, hogy a célnyelvi szótár méretének csökkentése csak minimális javulást okoz a rendszer pontosságában, viszont óriási információveszteséget eredményez.

Az eredmények azt is megmutatják, hogy tisztán statisztikai alapú módszerek nem elegek ezen feladatok megvalósítására, hanem szükség lenne valamiféle hibridizációra is. Az eredmények a jövőre nézve biztatóak, célom a további lehetőségek vizsgálata.

Hivatkozások

1. 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
2. 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
3. Oravecz, C., Dienes, P.: Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: Proc. of the Third LREC, pages 710–717, Las Palmas, Espanha. (2002) ELRA.
4. Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC 2006. (2006) 2245–2248
5. Laki, L.J., Prószéky, G.: Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Egyetem (2010) 69–79
6. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
7. Koehn, P.: Moses - A Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models. (2009)
8. 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
9. Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W.N.G., Weese, J., Zaidan, O.F.: Joshua: an open source toolkit for parsing-based machine translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. StatMT '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 135–139
10. Csendes, D., Hatvani, C., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Egyetem (2003) 238–247
11. Farkas, R., Szeredi, D., Varga, D., Vincze, V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Egyetem (2010) 349–353
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 311–318

Fordítási plágiumok keresése

Pataki Máté

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi utca 11.
pataki.mate@sztaki.hu

Kivonat: Napjainkban egyre több diák beszél idegen nyelveken, ami előny, hiszen fel tudják dolgozni az idegen nyelvű szakirodalmat és tudományos eredményeket, hátrány azonban, ha ezt hivatkozás nélkül teszik, azaz plagizálnak. Az elmúlt egy év alatt egy kutatás keretében arra kerestük a választ, hogy meg lehet-e találni, fel lehet-e ismerni a fordítási plágiumokat. Ennek során egy olyan algoritmust fejlesztettünk ki, amely képes egy nagyméretű, idegennyelvű adatbázisból kikeresni egy magyar nyelvű dokumentumban idézett, lefordított szövegrészeket.

1 Bevezetés

Természetes nyelvű szövegek fordításának megtalálása nemzetközi szinten is megoldatlan, még a sokak által beszélt angol és német nyelvek között is, ugyanakkor megoldása számos területen jelentene nagy előrelépést. A kutatási eredmények nemcsak plágiumok felkutatásában, hanem a párhuzamos korpuszok építésében, a hírek, cikkek, szövegek terjedésének a vizsgálatában, hasonló témákkal dolgozó emberek, kutatócsoportok megkeresésében is alkalmazhatók.

A párhuzamos korpuszok nagy jelentősége nemcsak az oktatásban rejlik, e korpuszok számos kutatás alapjaként, algoritmusok tanító adatbázisaként is szolgálnak. Használják őket az alkalmazott nyelvészetben: szótárkészítők, gépi fordítók számára, valamint kontrasztív nyelvészeti kutatásokhoz is elengedhetetlenek.

Európában fontos téma a plágiumkeresés, de még nemzetközi szinten is csak kutatási terület a fordítási plágiumok keresése. [1] Az irodalomban ismertetett legtöbb algoritmus nyelvpárfüggő, azaz egymáshoz nyelvtanban hasonló nyelvek esetén – barátságos nyelvpárok – jól működik, de jelentősen eltérő nyelvtanú nyelvek esetén rossz eredményt mutat. Angol-német nyelvpárra például egész szép eredményeket értek már el, míg az angol-lengyel nyelvpárra ugyanaz az algoritmus használhatatlannak bizonyult. A magyar nyelvben három fő akadály van: a) nem kötött szórend, b) ragozás, c) jelentős nyelvtani különbözőség az angol nyelvtől.

Dr. Debora Weber-Wulff kétévente teszteli az összes elérhető plágiumkeresőt, 2010-ben 48 plágiumkeresőt tesztelt, és azt állapította meg, hogy:

> „The biggest gap in all the plagiarism checkers was the inability to locate translated plagiarism.” [2]

Azaz a jelenleg elérhető plágiumkeresők egyáltalán nem foglalkoznak a fordítási plágiumok problémájával. Az első publikus eredmények többnyelvű plágiumkeresési algoritmusokról a CLEF 2010 konferencián [3] jelentek meg, de itt is csak barátságos nyelvpárokkal (angol, német, spanyol) próbálkoztak, és automatikus fordítót használtak a plágiumok megtalálására:

> „After analyzing all 17 reports, certain algorithmic patterns became apparent to which many participants followed independently. ... In order to simplify the detection of cross-language plagiarism, non-English documents in D are translated to English using machine translation (services).” [4]

2 Az algoritmus

A legtöbb szakirodalomban és kezdeti kutatásokban olyan algoritmusokat láthatunk a fordítási plágiumok keresésére, amelyek a jelenlegi egynyelvű keresések adaptálásai egy adott nyelvpárra. A legjobb plágiumkeresők átlapolódó szavas darabolást (n-gramokat) használnak a szövegek összehasonlítására, a plágiumkeresésre. [4] Ez az algoritmus szó szerinti egyezést keres, amelyet számos más algoritmussal igyekeznek javítani, hogy kisebb átírásokat, eltéréseket ne vegyen figyelembe, ezek közül a legygyakrabbak az alábbiak: a) stopszavak szűrése, b) szótövezés, c) bizonyos szavak kicserélése egy szinonimára, d) szavak sorrendezése az n-gramon belül. Ezek a változtatások sokkal nehezebbé teszik a plágiumok elrejtését, és jelentősen megnövelik a lebukás kockázatát, ugyanakkor különböző nyelven írt szövegek között még mindig nem teszik lehetővé az összehasonlítást.

Többen is próbálkoztak automatikus, gépi fordítók alkalmazásával, hogy két szöveget azonos nyelvre hozzanak, ugyanakkor ezen fordítók eredményei ma még nagyon megbízhatatlanok, nagyban függnek az adott nyelvpártól, a szöveg témájától, a mondatok összetettségétől. Összefoglalva elmondhatjuk, és ez nem csak a gépi fordítókra igaz – habár azokra kiemelten az –, hogy egy fordítás komoly változtatást eredményez a szövegen, hibákat visz be, és a szavak mondaton belüli sorrendjén is nagymértékben változtat, főleg az olyan nem kötött szórendű nyelvek esetében, mint amilyen a magyar.

A gépi fordítókat alkalmazó algoritmus tulajdonképpen két – különböző algoritmussal történő – fordítási lépésnek veti alá a szöveget (egy kézi a plagizáló által és egy gépi az ellenőrzéskor), majd az ezek után kapott, visszafordított szöveget hasonlítja össze az eredeti szöveggel. Esetleg egy adott szöveget kétszer fordít le egy másik nyelvre (egyszer kézzel, egyszer géppel), majd ezeket hasonlítja össze. Mivel a legtöbb mondatnak nincsen egy adott jó fordítása, hanem számos lehetséges fordítása van, így majdnem teljesen biztosak lehetünk benne, hogy komoly különbségek lesznek a mondatok között, nemcsak a szórendben, hanem a használt szavakban, kifejezésekben is. Fischer Márta ezt így fogalmazza meg:

„A nyelvészeti fordítástudomány eredményei – amelynek fontos területe az ekvivalencia kutatása – eloszlathatják azt a téves elképzelést, mely szerint a fordítás automatikus és teljes megfeleltetést (ekvivalenciát) feltételez a két nyelv között. A kutatók különböző megközelítései és a számtalan ekvivalencia-elmélet éppen arra világítanak rá, hogy az ekvivalencia több szinten, több szempont szerint értelmezhető Ezek ismerete tehát éppen abban erősítheti meg a tanulót, hogy nincs egyetlen helyes (ekvivalens) válasz.” [5]

Magyar nyelv esetében további hátrány, hogy a gépi fordítók igen rosszak, a legjobb angol-magyar nyelvpár esetében is tulajdonképpen majdnem minden mondatban hibáznak, és minél összetettebb a mondat, annál valószínűbb, hogy teljesen félre is fordítanak valamit.

Angol-német nyelvpár esetén már el lehet talán gondolkodni, hogy egy automatikus fordító alapján készítsünk egy algoritmust, de még ott is számos hiba adódik. Emellett komoly hátrány, hogy egy külső programra vagy algoritmusra kell hagyatkozni, hiszen a jó minőségű algoritmusok mind fizetősek, így nagyobb mennyiségű szöveg rendszeres lefordítása komoly költségekkel is járna. A Google Translate meghívható egy API-n keresztül, és korábban lehetett is nagyobb mennyiségű szöveget fordítani rajta, de pár hónapja a Google úgy döntött, hogy még fizetség ellenében sem engedi napi 100 000 karakternél nagyobb szöveg lefordítását. Ez még egy rövidebb diploma ellenőrzéséhez is kevés.

„The Google Translate API has been officially deprecated as of May 26, 2011. We are not currently able to offer additional quota.”

2.1. Az algoritmus kialakítása

Két nyelv között a legkisebb egyezés egy **szó** egyezése lehet. Természetesen, ha egy angol szövegben az *eleven* szót olvashatjuk, akkor annak magyarul nem az *eleven* szó fog megfelelni, hanem a *tizenegy* vagy a *11*, de ennek ellenére beszélhetünk egyezésről. Ugyanakkor érdemes megjegyezni, hogy számos szónak nem lesz megfelelője a másik nyelvben, vagy egyáltalán nem is lesz megfelelője, vagy nem szóként jelentkezik. Most a teljesség igénye nélkül vegyünk sorra pár lehetséges eltérést.

- Összetett szavak: elképzelhető, hogy míg az egyik nyelvben egy gondolatot egy szóval, addig a másikban több szóval fejezünk ki, mint például *tavaly és last year*. Fordítva pedig, míg magyarul *szabadlábra helyeznek* valakit, angolul ezt a jelentést a *liberated* adja vissza.

- Ragozás: a magyar nyelv (akárcsak például a török) számos dolgot ragokkal, a szóval egybe írva fejez ki, míg más nyelvek erre elöljárót használnak. Ami magyarul az *álmomban*, az angolul *in my dream* történt.
- Antoníma: gyakran egy kifejezést jobb antonímával fordítani, nem önmagával. Míg magyarul valami *nem felel meg a célnak*, addig ugyanez angolul *inadequate*.
- Ismétlések elkerülése: bizonyos nyelvek, mint például a magyar, kevésbé szeretik az ismétlést, és inkább utalnak az ismétlődő dolgokra, illetve szinonimákat használnak. A „80 nap alatt a föld körül" magyar fordításában találkozunk a *gentleman* szóval, ahol az angolban a *Mr. Fogg* szerepel.
- Teljes átalakítás: kifejezések és a forrás- valamint célnyelv különbözőségén, illetve a két olvasótábor kulturális ismeretének a különbözőségéből adódóan. A *Queen's pudding*-ból *rakott palacsinta* lesz, az *egg and spoon races* pedig *ügyességi gyerekjáték*. [6]

Azaz számos eset képzelhető el, amikor egy adott szó nem felel meg egyértelműen a másik nyelv egy szavának, ugyanakkor a szavak jelentős része megtalálható lesz mindkét nyelvben. Ugyan a szavakat jól fel lehet használni arra, hogy fordításokat keressünk, de önmagában két szöveg még nem lesz azonos pusztán azért, mert sok közös szavuk van.

Ha eggyel magasabb szintre lépünk, a **tagmondat**ok szintjére, akkor azt látjuk, hogy bár gyakran előfordul a tagmondatok egyezése, de míg a magyarban igen sok vesszőt használunk, és legtöbbször egyértelműen jelöljük a tagmondatok határát, addig az angol nyelvben alig vannak vesszők, és kimondottan nehéz feladat a tagmondatok határának megkeresése. Emiatt ezzel a lehetőséggel most itt nem is foglalkozunk.

A következő szint a **mondat**ok szintje. Ha valaki nekiáll egy szöveg fordításának, akkor azt az esetek túlnyomó részében mondatonként fordítja le. Egy irodalmi fordítás esetén gyakrabban találkozunk azzal, hogy egy mondatot kettőbe szed a fordító, vagy két mondatot összevon, de még itt is viszonylag ritkán fordul elő ez a gyakorlat.

Az ennél magasabb szintekkel, **bekezdés**ekkel, **fejezet**ekkel ugyanaz a legnagyobb gond, mint a tagmondatokkal: nem egyértelmű a jelölésük, elhagyhatóak, összevonhatóak, így ezek egyezésének a vizsgálatára úgyszintén nem térünk most ki.

Mint láttuk, fordítások esetében a legértelmesebb szint a szavak vagy a mondatok szintje. A szavak esetében viszont lényeges a szó többi szóhoz viszonyított pozíciója, a szövegkörnyezet, hiszen bármely két azonos nyelven íródott szövegben vannak azonos szavak, még akár ezek mértéke is magas lehet, azonban ekkor sem biztos, hogy a két szövegnek ugyanaz a jelentése, vagy esetleg csak a témája egyezik. Mint azt a webes keresők esetében látjuk – ahol adott szavakat tartalmazó szövegekre keresünk – nagyon nagy az olyan találatok száma, amelyek ugyan megfelelnek a keresőkérdésnek, de semmi közük sincs ahhoz, amit kerestünk. Azaz önmagában a szavak egyezősége nem tesz két szöveget egymás másolatává, nem lehet általa megállapítani a plagizálás tényét. Ez két különböző nyelv esetében még inkább így lesz, hiszen egy adott szónak a másik nyelvben számos másik felel, vagy felelhet meg, így még ez is komoly bizonytalanságot eredményez.

Természetesen ez nem azt jelenti, hogy a szavak nem használhatók két szöveg közti egyezés megtalálására, de önmagában ez nem elég: hiszen ha valaki lefordít egy egyoldalas szöveget angolról, és beteszi a 120 oldalas magyar diplomájába, akkor ennek a megtalálása csak a szavak használatával lehetetlen. Mindenképpen definiálnunk kell egy szövegkörnyezetet, ahol a szavakat keressük. Ezért a kutatáshoz a legjobb kiindulási pontnak a mondatalapú keresés tűnt, ahol a szavaknak van szövegkörnyezetük (egy mondat), ráadásul a mondat már elég egyedi ahhoz, hogy két dokumentumban – még ha azonos témában íródtak is – nagyon kicsi annak az esélye, hogy két azonos mondat lesz (rövid, egy-, két-, háromszavas mondatokat és közös idézeteket nem számítva). Könnyen beláthatjuk ezt, ha belegondolunk, hogy a legtöbb nyelvnek több százezer szava van [7], a nyelvtani szabályokat most figyelmen kívül hagyva, százezer szóval számolva az adott nyelven egy *n* szóból álló mondat (S_n) összes lehetséges változata:

$$|S_n| = (2 \cdot 10^5)^n$$

Ez egy még hosszúnak sem mondható 10 szavas mondat esetében:

$$|S_{10}| \approx 10^{53}$$

Természetesen ennek a jelentős része értelmetlen mondatot eredményezne, de ennek a hatalmas számnak még az egy tízezreléke is hatalmas. Ha hozzávesszük, hogy például a magyar nyelvben a legtöbb szónak számos alakja van, akkor ez a szám még jelentősen növekedne, de az angol nyelv esetében is a többesszám és egyéb alakok miatt az alapszókincs többszöröse a ténylegesen előforduló szóalakok száma. Ezért tekinthetünk úgy egy mondatra, mint egyedi alkotásra. Sokak szerint egy mondatnál kezdődik a plagizálás, azaz egy (tartalmas, hosszabb) mondat már rendelkezik annyi egyedi tulajdonsággal, hogy lemásolása esetén lehet plagizálásról beszélni.

Érdemes megnézni a Wikipédia ide vonatkozó oldalán található összefoglaló táblázatot, amelyből itt csak egy kivonatot mutatunk be. [8]

Dokumentum, bemeneti adat, szövegkörnyezet	**Szavak száma**	$\lvert S_{10} \rvert$
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 25%-át.	15	5,8E+11
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 60%-át.	100	1,0E+20
Kb. egy 2 éves gyerek szókincse	300	5,9E+24
Az Ogden-féle egyszerű angol nyelv (Basic English) szókincse	850	2,0E+29
Ennyi szót használnak az első osztályosok olvasástanításában.	1000	1,0E+30
Kb. egy 6 éves gyerek szókincse	2500	9,5E+33
Arany János Toldi c. művében felhasznált szókincse	3000	5,9E+34
Az átlagember aktív szókincse (élő-aktív és szunnyadó-aktív)	3 000-5 000	5,9E+34

Középfokú nyelvtudásnak megfelelő szókincs	3 500-3 900	2,8E+35
Kb. egy 11 éves gyerek szókincse	5 000	9,8E+36
Az átlagember passzív szókincse	5 000-10 000	5,6E+38
Ennyi szóval a Shreket 95%-ban megértjük.	6 000	6,0E+37
Ennyi szó szükséges a 20. századi angol próza megértéséhez.	8-9 000	1,1E+39
Ennyi szóval a tankönyveket 95%-ban megértjük.	10-12 000	1,0E+40
Egy kétnyelvű kisszótár terjedelme (címszavak)	10-30 000	1,0E+43
Shakespeare (műveiben felhasznált) szókincsét ennyire becsülik	18-25 000	1,7E+43
Petőfi Sándor verseiből kimutatható szókincse	22 719	3,7E+43
Egy átlag értelmiségi egyévi beszédét gondolatban rögzítve kb. ennyiféle szó fordulna elő.	25-30 000	3,0E+44
Igen művelt embereknél a passzív szókincs nagysága	50-60 000	2,5E+47
Kb. ennyi mai magyar szót tartanak számon.	60-100 000	1,1E+49
Egy kétnyelvű nagyszótár terjedelme (címszavak)	120 000	6,2E+50
A 20 kötetes Oxford English Dictionary 2. (nyomtatott) kiadásából (1989) a ma is használt szavak száma	171 476	2,2E+52
A 20 kötetes Oxford English Dictionary 2. (nyomtatott) kiadásának (1989) terjedelme (címszavak)	291 500	4,4E+54
A 33 kötetes Deutsches Wörterbuch terjedelme (1960-as kiadás, címszavak)	350 000	2,8E+55
A Webster's Third New International Dictionary, Unabridged terjedelme (címszavak)	>450 000	3,4E+56
A magyar nyelvben kb. ennyi szó (lexéma!) van (túlnyomórészt elavult vagy rendkívül speciális szavak)	1 000 000	1,0E+60
Az 1,48 milliárd szövegszót (v. szóelőfordulást) tartalmazó magyar webkorpusz 4%-os hibatűréssel készült metszetéből kinyert szókincs mérete (lexémák, ill. szótári szavak), kézi ellenőrzés nélkül	7 200 000	3,7E+68

Jól látható a táblázatból, hogy már egy kétéves gyerek is több száz szót ismer, és ha csak a rövidebb mondatokat vesszük, akkor is több tízezer mondatot tud elméletileg összetenni.

Összefoglalva az előzőeket, láthatólag a mondat egy értelmes egységnek tűnik ahhoz, hogy plágiumot, illetve szövegek közötti egyezéseket keressünk. Ennek az alábbi előnyei vannak:

- Egy értelmes gondolati egységet képvisel
- A mondathatárok nagy pontossággal meghatározhatóak
- A mondat elég egyedi ahhoz, hogy két szöveg között több mondat egyezésekor már valami közös forrást feltételezzünk
- Fordítások esetén a mondat a fordítás egysége, amely mint egység legtöbbször megmarad a különböző nyelvek között [9]

- Egy mondat és fordítása között ekvivalencia van, amely biztosítja, hogy a két mondat jelentése minél közelebb legyen egymáshoz

Miután beláttuk, érdemes a mondatok közötti hasonlóságot vizsgálnunk ahhoz, hogy a fordítási plágiumot megtaláljuk, definiálnunk kell egy metrikát, amely a különböző nyelven íródott mondatok közötti hasonlóság mértékét határozza meg.

2.2 A hasonlósági metrika

Mint korábban említettük, egy angol és egy magyar nyelvű mondat szavai – ha nem is teljes mértékben –, de megfeleltethetők egymásnak. A két nyelv nyelvtanának különbségéből és a magyar nyelv kötetlen szórendjéből adódóan a szavak sorrendje teljesen lényegtelen ebben a megfeleltetésben, azaz az angol nyelvű mondat első, második, harmadik… szava bárhol lehet a magyar mondatban, és fordítva.

A sorrendet figyelembe nem vevő, egy szöveg szavait reprezentáló modell a szózsák (bag of words) [10] – egy adott szöveg összes szavát tartalmazó, de a sorrendet figyelembe nem vevő halmaz –, amelyet számos helyen használnak a szakirodalomban például dokumentumok csoportosítására, spamszűrésre, de még érzelmek felismerésére is [11]. Mi most sokkal kisebb egységben, a mondatok szintjén fogjuk a szózsákot alkalmazni.

Egy *n* szóból álló mondatot (*S*) képviseljenek a benne lévő szavak (*w*).

$$w_x \in S_x \text{ és } w_y \in S_y$$

Természetesen ez egy egyszerűsítés, hiszen elméletileg ugyanazokból a szavakból más mondatokat is össze lehet rakni. Azonban, mivel az esetek túlnyomó részében elég egyértelműen visszaállítható a mondat értelme a szavak ismeretében, túl sok hibát ez az átalakítás nem fog eredményezni.

$$S_x = \{w_{x1}, w_{x2}, w_{x3}, \ldots w_{xn}\}$$

Most definiáljuk két mondat hasonlóságának a mértékét (*Sim*) a bennük levő közös szavak számával.

$$Sim(x,y) = | S_y \cap S_z |$$

Ez már egy jó megközelítés, de számos dolgot nem vesz figyelembe. Például egy hosszú és egy rövid mondat hasonlósága így maximum akkora lehet, amekkora a rövid mondat hossza. Ez helyes is, ugyanakkor például ha a hosszú mondatban megtalálható a rövid mondat összes szava, akkor ez a két mondat ugyanannyira hasonló lesz, mintha a rövid mondatot önmagával hasonlítottam volna össze, ami viszont egyértelműen rossz: ezért figyelembe kell venni nemcsak a közös szavakat, hanem a hiányzó szavakat is. Ezeket érdemes súlyozni is, most legyen a megtalált szavak súlya α , a nem megtaláltaké β.

$$Sim(x,y) = \alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y |$$

Amennyiben α értékét 3-nak, β értékét pedig 1-nek vesszük, akkor az azt jelenti, hogy minden olyan szót, amelyik megvan a másik mondatban, háromszoros súllyal veszünk figyelembe a hiányzó szavakhoz képest.

Ez a képlet már majdnem tökéletes, de nem szimmetrikus $S_x \setminus S_y$ miatt, azaz: $Sim(x,y) \neq Sim(y,x)$. Ez nem jó így, hiszen annak az esélye, hogy S_x S_y-nak a fordítása elvileg ugyanannyi kell legyen, mint annak esélye, hogy S_y S_x-nek a fordítása. Ezt a hibát úgy lehet kiküszöbölni, hogy például kiszámoljuk mindkét értéket, majd ennek vesszük az összegét. Ugyanakkor azért vezettük be az egyenlet második tagját ($S_x \setminus S_y$), mert azok a szavak, amelyek csak az egyik mondatban találhatóak meg, csökkentik annak valószínűségét, hogy a két mondat egymás fordítása. Ha annak az esélye, hogy S_x fordítása S_y-nak kisebb, mint a fordítottja azaz $Sim(x,y) < Sim(y,x)$, akkor ez a legtöbb esetben azt jelenti, hogy S_x hosszabb, azaz több olyan szó van benne, aminek nincs fordítása a másik mondatban. Ez lényeges: hiába kapunk Sim(y,x)-re egy nagyon magas értéket, ha Sim(x,y) alacsony, hiszen akkor majdnem biztos, hogy a két mondat nem fordítása egymásnak, esetleg az egyik a másik része. Ezért a továbbiakban úgy számoljuk ki Sim(x,y) értékét, hogy a korábban definiált értékek közül az alacsonyabbat vesszük. Ezzel az új képlet:

$$Sim(x,y) = \min (\alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y | ,$$
$$\alpha \cdot | S_y \cap S_x | - \beta \cdot | S_y \setminus S_x |)$$

Ez a definíció már eleget tesz a szimmetria (ekvivalencia) követelményének, azaz most már

$$Sim(x,y) = Sim(y,x)$$

A továbbiakban még néhány lényeges dolgot figyelembe kell vennünk ahhoz, hogy a szózsák algoritmus fordítások esetében is jól működjön. Mivel S_x és S_y nyelve nem azonos, ezért definiálnunk kell, hogy mit jelent két szó azonossága, illetve különbözősége: azaz mikor mondjuk, hogy $w_x \equiv w_y$ és mikor mondjuk, hogy $w_x \not\equiv w_y$ Ahhoz, hogy ezt meghatározzuk, definiálnunk kell még egy műveletet, a fordítás műveletét, azaz egy fordítási függvényt, amely egy szónak, illetve annak összes szótövének az összes fordítását adja vissza a másik nyelven.

$$trans(w_x) = W_y \text{ ahol } w_y \in W_y$$

$$trans(w_y) = W_x \text{ ahol } w_x \in W_x$$

mivel a fordítás egy szimmetrikus művelet, ezért ha

$$w_x \in trans(w_y) \text{ akkor } w_y \in trans(w_x)$$

ezek alapján definiáljuk, ha

$$w_y \in trans(w_x) \text{ akkor } w_x \equiv w_y$$

illetve ha

$$w_x \in \text{trans}(w_y) \text{ akkor } w_x \equiv w_y$$

hasonló módon ha

$$w_y \notin \text{trans}(w_x) \text{ akkor } w_x \not\equiv w_y$$

illetve ha

$$w_x \notin \text{trans}(w_y) \text{ akkor } w_x \not\equiv w_y$$

A fent leírt algoritmusnak számos előnye van: először is nem kell szóegyértelműsítést használni, hiszen az azonossági függvényünk – amelynek pontos működésének leírásától eltekintünk, csak a definícióját adtuk meg – ezt feleslegessé teszi azzal, hogy minden lehetséges jelentést figyelembe vesz. Az egynyelvű plágiumkeresésekben használt szinonima-egyértelműsítést, illetve -szűrést sem kell alkalmazni, hiszen egy szónak a lehetséges fordításai a másik nyelven egy vagy több szinonimahalmazba rendezhetőek, és ezeket az algoritmus transzparensen kezeli. Az algoritmus nem érzékeny a szavak sorrendjére, mint az n-gram algoritmus, azaz nem függ a fordítástól és nem működik nagyon eltérően barátságos és nem barátságos nyelvpárok esetében. Az algoritmus hátránya viszont a hatalmas keresési tér és a lineáris keresési idő, azaz a keresés ideje lineárisan függ az adatbázis méretétől. Nagy adatbázisok esetén ez gyorsan elfogadhatatlan keresési időket eredményez. Ez utóbbi problémát az implementációs fázisban egy indexált kereséssel meg tudtuk oldani, de most a részletek ismertetésétől – helyszűke miatt – eltekintünk.

2.3. Tesztkörnyezet kialakítása

Az algoritmus teszteléséhez szükségünk van olyan szövegekre, amelyeknek ismerjük a fordítását, valamint egy olyan hatalmas korpuszra, amely lehetővé teszi a hamis pozitív találatok tesztelését is, azaz egy olyan korpuszra, amely már biztos tartalmaz hasonló mondatokat, hiszen 10 mondatból kiválasztani egy adott mondat fordítását egy igen rosszul teljesítő algoritmusnak se lenne gond. Nagyméretű korpusznak a Wikipédiát választottuk, abból is az angol nyelvűt. [12] Amennyiben egy algoritmus képes egy Wikipédia méretű adatbázisból kiválasztani a megfelelő mondato(ka)t, akkor elmondhatjuk, hogy jól működik. Utóbbira azért is esett a választás, mert sokan idéznek, illetve sokan plagizálnak is sajnos a Wikipédiából, így gyakorlati haszna is van egy olyan keresőnek, amely kiemeli a Wikipédiából átvett részeket egy dolgozatban. Szótövezésre a MOKK által fejlesztett, ingyenesen elérhető Hunspellt alkalmaztuk [13]. Számos eszköz létezik, amely képes szövegeket mondatokra bontani, de mi három okból döntöttünk a saját algoritmus használata mellett: a) Először is a Wikipédia szövege – még szöveges formátumra alakítás után is – tartalmazott hibákat, például mondatok rendszeresen egybeíródnak a következővel (hiányzik a szóköz a mondatot lezáró írásjel után). b) Másodszor pedig egy olyan algoritmusra volt szük-

ségünk, ami gyors, és segítségével elkerülhetjük az újabb köztes fájlok létrehozását. c) Mivel ekkor már látszott, hogy a teljes folyamat igen erőforrás-igényes, ezért szerettünk volna minél kevesebb külső programot használni, hogy a plágiumkereső program minél több gépen legyen képes futni.

Több okból kifolyólag is elengedhetetlennek bizonyult egy automatikus fordító használata a tesztekhez. Az első és legfontosabb, hogy nem rendelkezünk annyi Wikipédiából – vagyis tulajdonképpen bárhonnan – származó angol-magyar párhuzamos korpusszal, amely elegendő lenne az algoritmus tesztelésére. Természetesen össze kell vetni az automatikus fordítóval és egy személy által fordított szövegen elért eredményeket, hogy megbizonyosodjunk arról, hasonló eredményt kapunk a két esetben. A könnyű elérhetőség és az API felület miatt esett a választás a Google fordítójára. [14]

Ahhoz, hogy egy angol és egy magyar szó azonosságát meg tudjuk állapítani, szükségünk van egy szószedetre, egy lapos szótárra. Ehhez kitűnő alapot nyújtott a SZTAKI online szótára. [15] Mivel azt is szükséges tesztelni, hogy a szótár mérete, illetve a hiányzó fordítások mennyire befolyásolják az algoritmust, ezért más, online elérhető szótárakkal illetve szószedetekkel is végeztünk kísérleteket. A kutatás jelentős részét az összes szótár uniójával végeztük.

3 Konklúzió

Az algoritmus teszteléséhez a teljes feldolgozott angol Wikipédiát feltöltöttük egy adatbázisba, és ebben kerestünk, mind a kézzel magyarra fordított, mind a géppel fordított Wikipédia cikkeket. A két keresés között statisztikai különbséget nem találtunk, így most a sokkal nagyobb mennyiségű, géppel fordított korpuszon elért eredményeket ismertetjük.

A magyar mondatokra keresve 0,67 recall értéket kaptunk, azaz ennyi volt az aránya azon mondatoknak, ahol a teljes Wikipédiából sikerült kiválasztanunk azt a mondatot, amelyiknek ez a magyar mondat a fordítása. Ez annyit jelent, hogy egyenletes valószínűséget feltételezve a mondatoknál annak az esélye, hogy egy 10 mondatból álló szakaszból egy hasonlót se találunk meg, 0,000016; és csak az esetek 2%-ban fogunk kevesebb mint 4 mondatot hasonlónak találni.

A recall értéke könnyedén mérhető, amennyiben tudjuk, hogy mit fordítottunk le a másik nyelvre. Ugyanakkor a pontosság meghatározása sokkal körülményesebb, hiszen kézzel kell ellenőrizni, hogy a visszaadott találatok közül melyek tényleges lehetséges fordítások, és melyek nem. Egy véletlen kiválasztott, kézzel fordított, és kézzel ellenőrzött korpusz esetében, ahol α értékét 2-nek, β-t pedig 1-nek választottuk, a hasonlósági metrika (*Sim*) minimumát pedig 8-nak, a pontosságra 0,92-t kaptunk, a recall értéke pedig 0,85 lett. Ebből F_1=0,88 adódik.

Az algoritmus kutatása már befejeződött, jelenleg az algoritmus finomhangolásán és a KOPI Plágiumkereső Portálba való integrálásán dolgozunk. A konferenciára már mindkettő elkészül és reményeink szerint be tudunk számolni az első publikus tesztek eredményéről is.

4 További tervek

Az algoritmust kézzel ellenőriztük más nyelvpárok esetében is, és az eredmények bíztatóak, de célunk, hogy pontosan kiszámoljuk a recall és pontosság értékeket legalább 10 további nyelvpár esetében is.

A szószedet mérete lineáris összefüggést mutat a futási idővel, azaz minél több lehetséges fordítása van egy szónak, annál nagyobb a keresési tér, és annál lassabb lesz a keresés. A pontosságot ugyanakkor sokkal kisebb mértékben javítja egy adott mérethatár felett, így meg kell határozni, hogy mi az ideális szószedet mérete, amely még gyors algoritmust eredményez, de már a találati pontossága is megfelel egy adott alkalmazáshoz. Ez a méret valószínűleg nyelvpárfüggő lesz.

Az algoritmus működik egynyelvű keresések esetében is, amennyiben a fordítási azonosság (*trans*) helyett szinonimákat, antonimákat, hiper- és hiponimákat használunk. Össze szeretnénk hasonlítani az egynyelvű keresést a jelenleg legtöbb plágiumkereső által használt n-gram algoritmus eredményével is.

Bibliográfia

1. Bailey, J: The Problem with Detecting Translated Plagiarism, http://www.plagiarismtoday.com/2011/02/24/the-problem-with-detecting-translated-plagiarism/ (2011)
2. Dr. Weber-Wulff, D.: Results of the Plagiarism Detection System Test 2010, http://plagiat.htw-berlin.de/software-en/2010-2/ (2010)
3. PAN 2010 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse http://www.uni-weimar.de/medien/webis/research/events/pan-10/ (2010)
4. Potthast, M.; Barrón-Cedeño, A.; Eiselt, A.; Stein, B.; Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection, http://www.clef2010.org/resources/proceedings/clef2010labs_submission_125.pdf (2010)
5. Fischer, M.: Fordítás és közvetítés a nyelvoktatásban – mit nyújthat a nyelvoktatásnak a fordítástudomány? , http://ecml.opkm.hu/files/FischerM.doc (2008)
6. Tóth, P.: Fordításelmélet, http://dettk.ucoz.com/load/0-0-0-93-20 (2005)
7. How many words are there in the English language?, Oxford University Press, http://oxforddictionaries.com/page/93 (2011)
8. Wikipedia, Szókincsméretek összehasonlító listája, http://hu.wikipedia.org/wiki/Szókincsméretek_összehasonlító_listája (2011)
9. Nida, E. A.: Toward a Science of Translating. E. J. Brill, Leiden (1964)
10. Wikipedia: Bag of words model, http://en.wikipedia.org/wiki/Bag_of_words_model (2011)
11. Miháltz, M.: OpinHu: online szövegek többnyelvű véleményelemzése. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2010)
12. Wikipedia the free encyclopedia, http://en.wikipedia.org/ (2011)
13. BME MOKK: Hunspell szótövező, helyesírás ellenőrző, morfológiai elemző, http://hunspell.sourceforge.net/ (2011)
14. Google: Google Translate, http://translate.google.com/ (2011)
15. MTA SZTAKI: SZTAKI Szótár, http://szotar.sztaki.hu/ (2011)

Soknyelvpáros gépi fordítás hatékony és megbízható kiértékelése

Oravecz Csaba, Sass Bálint, Tihanyi László

MTA Nyelvtudományi Intézet
e-mail:{oravecz.csaba,sass.balint,tihanyi.laszlo}@nytud.hu

Kivonat Gépi fordítások kiértékelésére a legmegbízhatóbb módszer az emberi szakértői kiértékelés, mely egyértelműen elsődleges mindenfajta egyéb megközelítéssel szemben. A dolgozat arra keresi a választ, hogy milyen elfogadható alternatívákkal váltható ki a szakértői kiértékelés abban az esetben, amikor ez a preferált, ugyanakkor rendkívül erőforrásigényes módszer a kiértékelendő szövegek nagy mennyisége, illetve a kiértékelési feladat sajátos paraméterei miatt nem alkalmazható. A javasolt megoldás a rendelkezésre álló többféle típusú kiértékelési információt rugalmasan kombináló és ennek alapján minőségi klasztereket képző eljárás, ahol az egyes klasztereken belül minden fordítási kimenethez véletlenszerűen generálódik az aktuális rangsor.

Kulcsszavak: gépi fordítás, fordításkiértékelés, korreláció, fordítóportál

1. Bevezetés

A kutatás hátterét az iTranslate4.eu nemzetközi projektum adja, melynek keretében elkészült egy 63 nyelvpár közötti automatikus gépi fordítást és egyéb fordításon alapuló szolgáltatást kínáló webportál. A weboldalon a fordítást 14 szolgáltató által kifejlesztett szabályalapú, illetve statisztikus fordítómotorok végzik. A 63 nyelvpár összesen $63 \times 62 = 3906$ nyelvpár közötti fordítást tenne szükségessé. Bár a portál számára valójában csak 233 nyelvi motor áll rendelkezésre, megfelelő közvetítő nyelvek megválasztásával a portál kiszolgálja valamennyi nyelvi irányt, így tetszőleges nyelvről tetszőleges másikra fordít.

A portál egyedi sajátossága hasonló online fordítókkal szemben, hogy egy-egy kérésre több megoldással is tud szolgálni. Mind a különböző programok gyártóinak, mind a felhasználóknak természetes igénye, hogy ezek az alternatívák minőségi sorrendben jelenjenek meg. Ehhez szükség van az egyes fordítók kérdéses nyelvpárok szerinti teljesítményének a kiértékelésére, hatékony és megismételhető, a fordítómotorok minőségi változását követni képes módon. A feladat volumenének következtében a szakértői emberi kiértékelés nem vehető számításba, más módszereket kell kidolgozni. A kiértékelési feladat célja tehát alapvetően bekezdés hosszúságú szövegek sorrendbe rendezése, amelynél figyelembe kell venni, hogy

– a minősítés nem lassíthatja a fordítási folyamatot,
– a szövegek megjelenítésének célja a megértés és nem az újrafelhasználás, ezért olyan offline kiértékelési eljárások preferálandók, amelyek inkább a felhasználói vélemény, mintsem az esetleges utószerkesztéshez szükséges költségmetrika alapján rangsorolnak.

Az offline megoldással természetesen nem az éppen megjelenő fordításokat rangsoroljuk, hanem az azokat létrehozó fordítóprogramokat. A rangsor a fordítóprogramok szempontjából releváns, hiszen a következő kiértékelésig meghatározza azok sorrendjét. A minősítések a fordításokkal együtt nem jeleníthetők meg, hiszen a felhasználó a konkrét megoldás minősítését várná el, a fordítók általános minősítése ezt pedig csak közelítheti.

2. Gépi fordítások kiértékelése

A gépi fordítások kiértékelése közismerten körülményes és bonyolult feladat, melyre hosszú ideje keresnek hatékony és könnyen kivitelezhető megoldást. Az automatikus kiértékelő metrikák legismertebbje, a Bleu-mérték [17] mellett mára további módszerek sokaságát fejlesztettek ki (lásd pl. a [7] kiadványt, illetve a [4] tanulmányban található összefoglalót). Széles körben elfogadott ugyanakkor, hogy az automatikus módszerek megbízhatósága jelentősen elmarad a (szakértői) humán kiértékeléstől [4], ezért gyakorlati hasznuk leginkább a fordítómotorok fejlesztése során van [6]. A legjobb eredményeket adó eljárások ezen túl olyan nyelvi előkészítést és adott nyelvi erőforrások (pl. WordNet) meglétét igénylik, melyek a jelen feladat kontextusában nyilvánvalóan a kérdéses nyelvek nagy részében nem állnak rendelkezésre. További probléma, hogy a statisztikai alapú fordítórendszerek, melyek egyre inkább dominánsak a szabályalapú rendszerek felett, egyre több, gyakorlatilag minden elérhető adatot igyekeznek felhasználni betanításuk érdekében. Ezért lehetetlen, de legalábbis bizonytalan kimenetelű egy elfogulatlan, fenntartható és folyamatos nagy léptékű kiértékelő környezetet kifejleszteni, hiszen a tesztadatok függetlensége nem biztosítható.

A fentiek fényében egyértelmű a humán kiértékelés elsődlegessége akkor, amikor a feladat a többféle fordítómotor által szolgáltatott fordítások valamilyen rangsorba állítása. A legjobb megoldás természetesen a szakértői kiértékelés, ám az így kapott eredmények objektív értelmezése sem problémamentes [2]. Kézenfekvő persze, hogy jelen esetben ez a rendkívül erőforrásigényes módszer a kiértékelendő szövegek nagy mennyisége, illetve a kiértékelési feladat sajátos paraméterei miatt eleve szóba sem jön, a végső megoldásban fenntartható módon nem alkalmazható.

3. Módszerek és vizsgálatok

3.1. A kiértékelendő nyelvek, nyelvpárok és fordítómotorok

Bár 63 nyelv esetén a nyelvpárok elvi kombinációjának száma 3096, ennél jóval kevesebb nyelvpár kiértékelésével kellett foglalkoznunk. Ennek több oka is volt:

egyrészt a valójában nyelvi motorral is támogatott nyelvpárok száma csak 233, a többi esetben pedig közvetítő nyelven keresztül két lépésben fordít a rendszer. A portálunkhoz hasonlóan a Google és a Microsoft fordítóprogramjai is közvetítő nyelvet használnak, azaz az általuk támogatott nyelvpárok száma ezek esetén is csak a nyelveik számának a kétszerese. A többi 12 fordítóprogram a minőségi normák betartása érdekében nem végez közvetítő nyelves fordítást, itt a nyelvpárok száma közvetlenül ismert. Mivel a kiértékelési feladatunk célja rangsorolás volt, ezért nem kellett figyelembe venni azokat a nyelvpárokat sem, amelyeken csak egy versenyző indult, ezzel a nyelvpárok száma 106-ra csökkent.

A weboldalon fordító programok két nagy kategóriába csoportosíthatók. Az egyikbe a szerződéses partnerek, a másikba pedig a Google és a Microsoft tartoznak. Az utóbbiak szabadon elérhető programozói felület (API) segítségével integrálhatók. Mivel azonban mind a Google, mind a Microsoft fordítók ilyen jellegű felhasználása hamarosan fizetős szolgáltatássá válik, ezért ezeknek a nyelvpároknak üzemeltetése és kiértékelése csupán tájékoztató jellegű eredménnyel szolgálhat, a végleges megoldásban nem játszik szerepet. A 12 partnerfordítóból a legalább kettő által támogatott nyelvpárok száma 58 volt. Mivel a kiértékelési eljárások költségét alapvetően a kiértékeléshez szükséges nyelvi erőforrások (párhuzamos szövegek gyűjtése, tesztek összeállítása) teszik ki, ezek csak egy-egy új nyelvpár esetén jelentenek többletköltséget. Vagyis a partnerek miatt kiértékelendő nyelvpárok esetén a kiértékelés további költség nélkül kiterjeszthető a Google és Microsoft fordítókra is.

A kiértékelési feladat során a versenyzők számának alakulása és a különböző nyelvpárok (nyelvek ISO kód szerinti rövidítésével) az alábbiak voltak:

8: fr-de, en-de, de-fr, de-en
7: fr-en, en-fr
6: it-en, es-en, en-it, en-es
5: ru-en, pt-en, pl-en, fr-es, es-fr, es-de, en-ru, en-pt, en-pl, de-es
4: zh-en, uk-en, tr-en, sv-en, sl-en, ru-pl, ru-fr, ru-de, pl-ru, pl-fr, pl-de, no-en, lv-en, it-fr, it-es, it-de, hu-en, fr-ru, fr-it, fi-en, es-it, en-zh, en-tr, en-sv, en-lv, en-hu, en-fi, en-da, de-ru, de-pl, de-it, da-en, bg-en

A fenntartható kiértékeléshez kétféle kivitelezhető megközelítés választható, ám mindegyik felvet számos olyan kérdést, melyet a hatékony módszer kidolgozása érdekében meg kell válaszolni:

A. Valamilyen sztenderd mérték(ek) szerinti automatikus, gépi kiértékelés.
B. Emberi, de nem szakértői kiértékelés, amely nagy léptékben is alkalmazható.

3.2. Automatikus kiértékelés

Az automatikus kiértékelés (a továbbiakban AU) során az IQMT [12] keretrendszer által szolgáltatott 5 féle sztenderd mérték normalizált átlagát használtuk: BLEU [17], NIST [9], GTM [16], METEOR [1] és ROUGE [13]. Ideális esetben 3 humán referenciafordítás szükséges a kiértékeléshez, tekintve azonban a projektben szereplő nyelvek széles skáláját, ilyen mennyiségű fordítás beszerzése,

előállítása reménytelen, így egy referenciafordítást alkalmaztunk, és a felhasznált szövegek műfajának és forrásának variábilitásával próbáltuk kiegyensúlyozottabbá tenni az automatikus kiértékelést. A kívánt nyelvi erőforrások az EU párhuzamos hírkorpuszból származnak, 13 különböző témakategóriából, mintegy 80 ezer szövegszó méretben. Természetesen, hiába saját gyűjtésről van szó, itt is felmerül a források függetlenségének kérdése: vajon ezek a szövegek nem alkották-e a részét a statisztikus fordítóprogramok tanítókorpuszának.

3.3. Emberi, nem szakértői kiértékelés a Mechanical Turk rendszerben

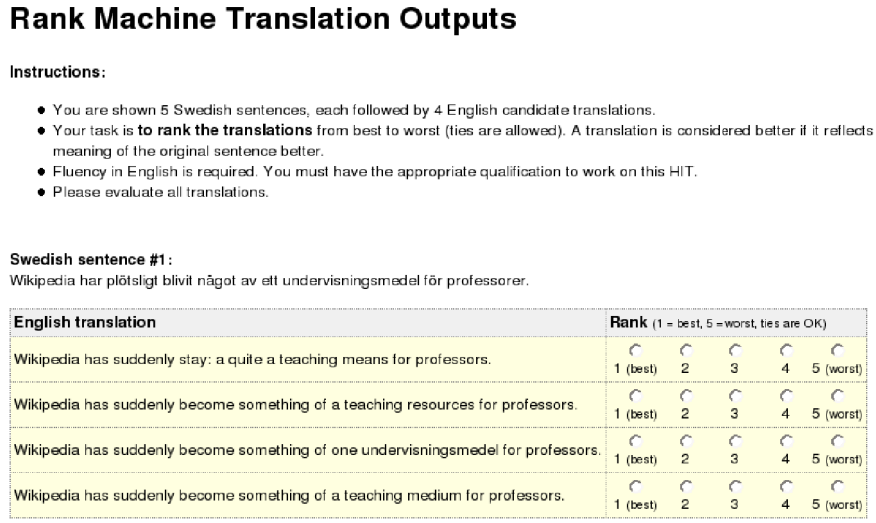
A nagyobb volumenű emberi, nem szakértői fordításértékelés megvalósítására lehetőséget adnak az utóbbi években létrejött, online elérhető *crowdsourcing* rendszerek. Ezekben a rendszerekben internetes űrlap formájában megfogalmazható, emberi intelligenciát igénylő feladatok (HIT, human intelligence task) tehetők közzé. A feladatokat a regisztrált dolgozók (worker) meghatározott fizetség ellenében végzik el. Lehetőség van a dolgozók előzetes szűrésére, például megtehetjük, hogy csak olyan dolgozók jelentkezését fogadjuk, akik már korábban adott számú HIT-et sikeresen megoldottak. A nem megfelelő minőségűnek ítélt munkavégzés esetén a fizetség visszatartható. Ezek az eszközök segítenek a munkavégzés általános minőségi szintjét magasan tartani. A *crowdsourcing* rendszerekkel tehát olcsón és gyorsan lehet megbízható minőségű megoldást találni emberi intelligenciát igénylő feladatokra [3], ugyanakkor legújabban már az ilyen rendszerek esetleges kockázataira is felhívják a figyelmet [11].

Eljárásunk. A gépi fordítások emberi, nem szakértői kiértékelésére (a továbbiakban MT) a Mechanical Turk (`http://www.mturk.com`) internetes rendszert alkalmaztuk.

Forrásnyelvenként 30 darab, téma szerint minél változatosabb közepes hosszúságú (legnagyobbrészt 10–30 szavas) mondatot gyűjtöttünk. Ezeket a mondatokat a rendelkezésre álló fordítóprogramok mindegyikével lefordítottuk. Hogy egy kiértékelési feladat ne legyen túl időigényes, egy feladatba (HIT-be) 5 mondatot tettünk, azaz a 30 mondatot 6 db 5-ös csoportra osztottuk. Egy kiértékelőnek tehát egy feladat keretében 5 db mondat fordításait kellett értékelnie.

A kiértékelőknek az a feladata, hogy 1-től 5-ig terjedő skálán minőség szerint *pontozzák* a fordításokat. Az instrukciók és egy mintafeladat – svéd–angol nyelvpárra, ahol 4 különböző automatikus fordító van – a 1. ábrán látható. A feladat a fordítások sorba rendezése, 1-től (legjobb) 5-ig (legrosszabb) skálán adott pontszám segítségével. Több mondatnak adható azonos pontszám, és a fordítások számától függetlenül 1-től 5-ig terjedő skálát használunk.

A rendszer működéséből adódóan egy kiértékelő tetszőleges számú mondat kiértékelését elvégezhette (azaz akár az összes 30 mondatét is). Ezért – hogy semmiképp se csak egy dolgozó véleményére támaszkodjunk – minden mondatot 3 különböző kiértékelővel értékeltettünk ki. Itt a különbözőséget szintén a rendszer biztosítja. Végeredményben tehát fordítónként $3 \times 30 = 90$ kiértékelési pontszámot kaptunk, ami minimum három különböző kiértékelőtől származott.



Rank Machine Translation Outputs

Instructions:

- You are shown 5 Swedish sentences, each followed by 4 English candidate translations.
- Your task is **to rank the translations** from best to worst (ties are allowed). A translation is considered better if it reflects the meaning of the original sentence better.
- Fluency in English is required. You must have the appropriate qualification to work on this HIT.
- Please evaluate all translations.

Swedish sentence #1:
Wikipedia har plötsligt blivit något av ett undervisningsmedel för professorer.

English translation	Rank (1 = best, 5 = worst, ties are OK)
Wikipedia has suddenly stay: a quite a teaching means for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of a teaching resources for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of one undervisningsmedel for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of a teaching medium for professors.	1 (best) 2 3 4 5 (worst)

1. ábra. Egy a Mechanical Turk rendszerben megvalósított fordításkiértékelési feladat dolgozóknak szóló felülete a svéd–angol nyelvpár esetén.

A kapott 90 db érték összesítésére kétféle mérőszámot alkalmaztunk. Egyrészt egyszerűen átlagot számoltunk, másrészt az EuroMatrix projektben [5, 3.1 rész] alkalmazott mértéket használtuk, miszerint egy fordítórendszer minden olyan esetben kap egy pontot, ha egy kiértékelő szerint egy másik rendszernél jobb (vagy vele egyforma), és végül pontszám szerint rendeztük a fordítórendszereket. A két mérőszám lényegében minden esetben ugyanazt az értéket adta, ezért a pontszámok átlagával dolgoztunk a továbbiakban.

Minőségbiztosítás. A fordításértékelési feladat megoldásához nyilván szükséges mindkét nyelv megfelelő ismerete, magasszintű ismeret főként a célnyelv esetében kívánatos. Annak érdekében, hogy valóban jó minőségű értékeléseket kapjunk, bevezettük azt, hogy a dolgozóknak először ki kell tölteniük egy rövid tesztet az adott nyelvpárra vonatkozóan, és csak akkor dolgozhatnak a kiértékelésben, ha ez jó eredményű. A Mechanical Turk terminológiájával egy megfelelő minősítés (qualification) meglétét követeljük meg, mielőtt a dolgozó hozzákezd a munkához.

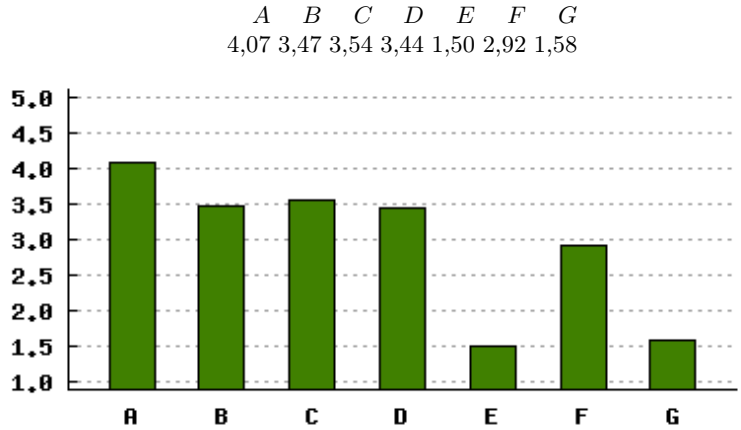
A célnyelvre fordítás képességét egy négy kérdésből álló teszttel mértük, négy darab forrásnyelvi mondat esetében kellett megmondani, hogy a felkínált fordítások közül melyik a legjobb. A szándékosan hibás fordításokban morfológiai, szintaktikai és szemantikai, szókincsbeli hibák egyaránt előfordultak.

Sorrendkeverés. Kutatásunk első szakaszában a fordítások mindig fix sorrendben jelentek meg. Ez a sorrendből adódó nem kívánt torzító hatáshoz vezetett.

E hatást és kiküszöbölését a német–angol nyelvpáron mutatjuk be, ahol 7 fordítórendszert teszteltünk.

A pszichológiában ismert az a jelenség, hogy ha több azonos típusú entitást kell értékelnünk, akkor jelentősége van annak, hogy ezek a bizonyos értékelendő dolgok milyen sorrendben kerülnek elénk. Megfigyelték, hogy bizonyos esetekben hajlamosak vagyunk az elsőként látottat előnyben részesíteni (*primáciahatás*, vö. [15]), más feltételek mellett pedig az utolsót (*recenciahatás*, vö. [8]). Ezek a jelenségek főként akkor figyelhetők meg, mikor az adott jelölt megfigyelése után azonnal értékelni kell, nem várhatjuk meg a pontszámokkal az összes versenyzőt (ilyen például a műkorcsolya-zsűrizés struktúrája). Esetünkben lehetőség volt a jelöltek (fordítások) többszöri vizsgálatára, összevetésére, és csak az összes jelölt vizsgálata után kellett döntést hozni, mégis határozott primáciahatást találtunk, amit torzította az eredményeket.

A német–angol nyelvpáron végzett első kísérletekben tehát a 7 angol fordítás mindig fix sorrendben, a fordítórendszerek neve szerinti betűrendben jelent meg az eredeti német mondat után. A fordítónként 90 értékből adódó átlagos pontszámok a 2. ábrán láthatók.

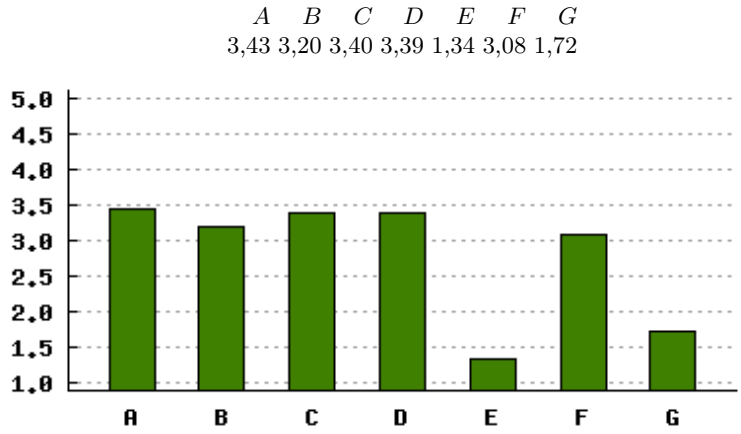


2. ábra. Fordítókénti átlagos pontszámok. Itt a 7 angol fordítás mindig *a fordítók neve szerinti betűrendben* következett az eredeti német mondat után. (Az osztályzás itt eredetileg 1-től 7-ig történt, utólag normáltuk ezt az összehasonlíthatóság kedvéért az 1..5 skálára a következő módon: normált = eredeti $\times \frac{2}{3} + \frac{1}{3}$.)

A sorrendi hatások kiegyenlítése nem mindig könnyű [8], esetünkben azonban egy egyszerű, determinisztikus *sorrendkeverő* algoritmus segítségével biztosítani lehetett azt, hogy minden pozíció esetében igaz legyen az a feltétel, hogy minden fordító ugyanannyiszor fordul elő az adott helyen.

A sorrendkeverő algoritmus alkalmazásával a fordítások determinisztikus módon változó, a keverőalgoritmus által meghatározott sorrendben követték egy-

mást. A német–angol nyelvpár esetében a fordítókénti 90 értékből így adódó átlagos pontszámokat a 3. ábrán láthatjuk.



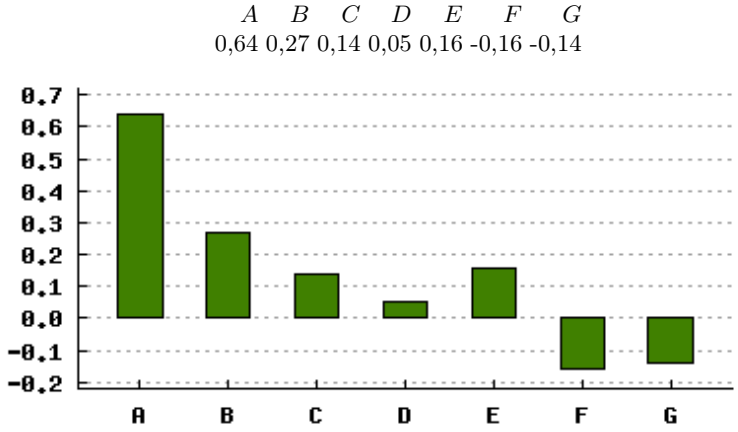
3. ábra. Fordítókénti átlagos pontszámok. Itt a 7 angol fordítás mindig *változó, a keverőalgoritmus által meghatározott sorrendben* következett az eredeti német mondat után.

A 2. és a 3. ábrát összevetve látjuk, hogy egy helyen maga a sorrend is megváltozott (*B*-*D* helyett *D*-*B*), de ennél lényegesebb annak feltárása, hogy milyen mértékben változtak a pontszámok a két elrendezés között. A különbségeket ábrázoltuk a 4. ábrán. Az ábra tanúsága szerint egyértelmű primáciahatást tapasztalunk („a fix első hely jogtalan előnnyel jár; aki előrébb van, az érdemtelenül több pontot kap”), egyfajta fordított recenciahatással erősítve („aki hátrébb van, az igazságtalanul kevesebb pontot kap”). A torzító hatás arányos az eredeti pozícióval.

Az eredmény arra hívja fel a figyelmet, hogy az ilyenfajta többszöri értékeléses feladatokban egyáltalán nem mindegy, hogy milyen sorrendben szerepelnek az értékelendő entitások, a sorrend nagyban befolyásolja az eredményt. Az igazságos értékeléshez fontos a sorrendi hatások kiküszöbölése, különben torzul az eredmény.

3.4. Felhasználói visszajelzések

A harmadik kiértékelő komponenst a felhasználói visszajelzések (továbbiakban FV) alkotják. Ezek valójában az egyes fordításokra érkezett szavazatok, amelyeket a portálon adhatnak le a felhasználók. Egy fordítás esetén több megoldás is megjelölhető. A szavazatokat a portál megnyitása óta gyűjtjük. Bár a szavazati hajlandóság viszonylag magas (5%-os), az induló weboldal látogatóinak alacsony száma miatt az adatok mennyisége csak lassan nő. A szavazás során



4. ábra. Fordítókénti átlagos pontszámok *különbsége* az első – sorrendi hatásnak kitett (vö. 2. ábra) –, és a második – sorrendi hatásra semleges (vö. 3. ábra) – elrendezés között. Bár az eltérés csak A esetében szignifikáns (kétmintás Welch-próba: $p \ll 0.05$), jól látható egy trend, miszerint a sorrendi hatásnak kitett esetben az előrébb lévők jogtalan előnyhöz jutnak, a hátrébb lévők pedig hátrányt szenvednek.

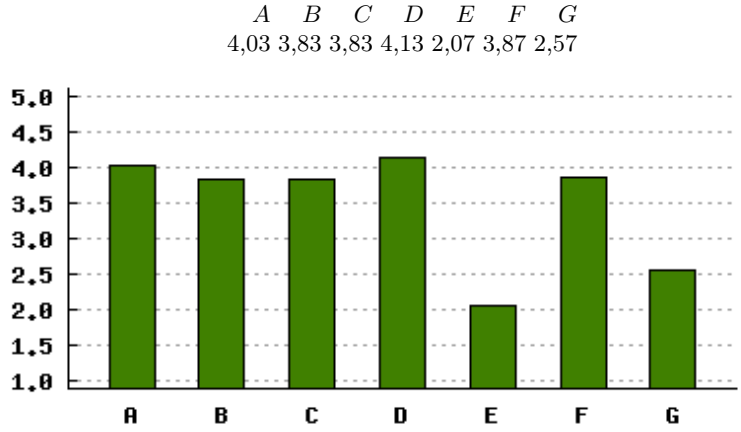
kiderült, hogy a közvetítőnyelves megoldások is használatban vannak, és szavazatokat tudnak gyűjteni. Ezekre sem az automatikus, sem az MT kiértékelések erőforrás hiányában nem tudtak adatokkal szolgálni. A partnerek között elvi egyetértés alakult ki arról, hogy a jövőben, megfelelő mennyiségű adat birtokában az FV kiértékelés legyen elsődleges, hiszen ez elvileg valódi fordítási kérdésekre valódi felhasználók által adott értékelést képvisel. Vizsgálni kell azonban ennek a kiértékelési módszernek a megbízhatóságát is.

4. Eredmények

4.1. A kiértékelések összevetése

Fontos kérdés, hogy a 3.3. részben leírt módszer segítségével a Mechanical Turk rendszerrel valóban lehetséges-e magas megbízhatóságú kiértékelést végezni. Ezt úgy vizsgálhatjuk meg, hogy a szakértő véleményét vetjük össze a nem szakértő dolgozók véleményével. Ennek érdekében kiértékeltettük a már említett német–angol nyelvpárt egy szakértővel. A szakértő által adott 30 darab pontszám átlagos értéke a 5. ábrán látható.

Annak ellenére, hogy a kis eltérések miatt a fordítók sorrendjében lényeges különbségek vannak, megfigyelhető, hogy a nem szakértői kiértékelők (3) és a szakértő (5) meglehetősen hasonlóan értékelte a fordításokat, ahogy a két ábrán látható grafikon lefutásán is látható. Célszerű ezért a rangsorok összehasonlítására szokásosan használt Spearman-féle rangkorrelációs együttható helyett más



5. ábra. A szakértő átlagos pontszámai német–angol nyelvpárra. A grafikon lefutása lényegében megegyezik a 3. ábrán láthatóval.

megközelítést alkalmazni a hasonlóság mértékére. Kolmogorov–Szmirnov próbával vizsgáltuk meg, hogy mennyire valószínű, hogy a két grafikon ugyanazt írja le. A p értékre 0,05-nek adódott, azaz 5% hiba mellett mondhatjuk, hogy igaz az, hogy a nem szakértők és a szakértő gyakorlatilag ugyanúgy értékelte a fordításokat. Emiatt a Mechanical Turk rendszerben kapott kiértékeléseket is megbízhatónak tarthatjuk, azaz általánosságban támaszkodhatunk erre a sokkal olcsóbb és egyszerűbben kivitelezhető emberi kiértékelési metódusra. Korábban úgy gondolták [3], hogy a *crowdsourcing* megbízható kiértékelési eredményeket ad, ez később megkérdőjeleződött [4], jelen eredményeink azt mutatják, hogy ha az alkalmas dolgozókat a 3.3. részben bemutatott eljárás segítségével választjuk ki, a megbízhatóság megfelelő szintű lesz.

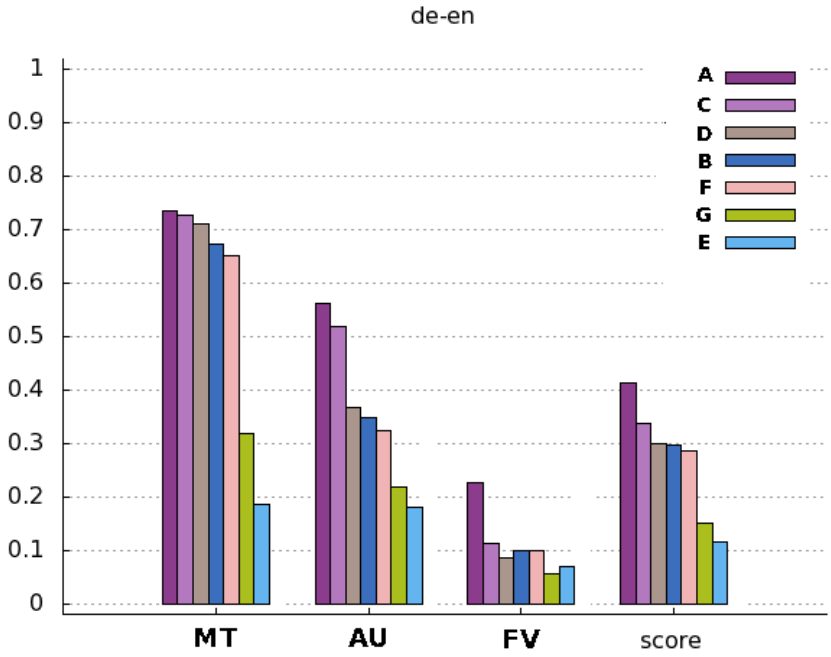
A további komponensek összehasonlítása során beigazolódott, hogy a szakértői kiértékeléshez legközelebb álló MT módszer után a felhasználói visszajelzések a legmegbízhatóbbak, az automatikus kiértékelés pedig, különösen a statisztikai fordítókkal szembeni elfogultság miatt a legkevésbé megbízható. Azokon a nyelvpárokon, ahol közvetett és közvetlen fordítások is elérhetők voltak, egyértelműen megmutatkozott az utóbbiak minőségi fölénye.

4.2. Javasolt kiértékelési módszer

A gyakorlati alkalmazásban nehezen védhető egy, a kiértékelések alapján rögzített rangsorba rendezés a fordítómotorok között, és a fordítások e szerinti megjelenítése. A 6. ábra illusztrál egy olyan összevont rangsort, ahol az egyes fordítómotorokhoz rendelt mérték (*score*) a három komponens (s) súlyozott átlaga ($w_1 = 0.1, w_2 = 0.3, w_3 = 0.6$):

$$score = \frac{w_1 s_{\mathrm{AU}} + w_2 s_{\mathrm{MT}} + w_3 s_{\mathrm{FV}}}{3} \tag{1}$$

A kis minőségi különbséggel hátrább sorolt partner jogosan tiltakozik, hogy a



6. ábra. Az egyes komponensek eredményei és az összevont rangsor.

sohasem 100%-osan megbízható értékelés(ek) alapján *véglegesen* rosszabb helyre kerül. Ezért a rögzített rangsor helyett az alábbi javasolt módszerrel próbáljuk kiküszöbölni ezt a problémát.

Képezzünk az értékelés során kapott eredmények alapján a fordítómotorok között minőségi klasztereket. A klaszterek számát az értékeléskor kapott adatok alapján kell automatikusan meghatározni (a 3., 5. és 6. ábrán látható adatok alapján például két minőségi klasztert célszerű képezni, ha eltekintünk az AU módszer elfogultságától a statisztikus fordítók felé). Erre kétféle megközelítés alkalmazható: a klaszterek számát előre megkívánó algoritmus (pl. k-means) esetében valamilyen segédalgoritmus (lásd pl. [14,18]), illetve a klaszterek számát is meghatározó klaszterező algoritmus [10]. Az egyes klasztereken belül alapesetben véletlen rendezés szerint jelennek meg a fordítások. A klaszterek képzéséhez szükséges bemenő adatot az adott nyelvpárra kétféleképpen állíthatjuk elő. Egyrészt a rendelkezésre álló kiértékelő komponensek eredményeinek például (1) szerinti összevonásával, vagy az éppen legmegbízhatóbbnak tekinthető és elegendő adatot szolgáltató komponens kizárólagos figyelembevételével (ahol a megbízhatósági sorrend a következő MT $\rightarrow$ FV $\rightarrow$ AU). A legjobb megoldás kiválasztásához

további értékelési adatok és vizsgálatok szükségesek, ahol természetesen azt is meg kell határozni, mit fogadunk el elegendő adatnak.

Ez a módszer feltétlen igazságosabb és a partnerek által is elfogadhatóbb, mint a kötött rangsor alapján történő rendezés, megvalósítása azonban technikai okok miatt csak részleges lehet. A fordítómotorok eltérő sebessége miatt portál felületen definiált meghatározott maximális válaszidő (jelenleg 1mp) már eleve kialakít egy sorrendet. A portál szolgáltatásait közvetítő API alkalmazásokban pedig a hívó fél állítja be a kért megoldásokat, az általa tapasztalt sebességi és minőségi eredmények alapján.

5. Összefoglalás és további feladatok

A tanulmányban megvizsgáltuk, hogy egy konkrét alkalmazásban hogyan valósítható meg gépi fordítások kiértékelése olyan környezetben, ahol számos gyakorlati paramétert kell figyelembe venni. Javaslatot tettünk olyan kiértékelési módszerre, amely választ ad a felmerülő problémákra: megbízható, fenntartható és soknyelvpáras fordítás esetén is alkalmazható, ezzel együtt védhető és igazságos minősítést eredményez. A portál működése során gyűjtött adatok mennyiségének növekedése további részletes vizsgálatok elvégzésére ad lehetőséget, melyek kiértékelése még megalapozottabban kimutathatja az egyes fordítók közötti minőségi különbségeket.

Hivatkozások

1. Banerjee, Satanjeev és Lavie, Alon. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005, 65–72.
2. Bojar, Ondřej, Ercegovčević, Miloš, Popel, Martin és Zaidan, Omar. A Grain of Salt for the WMT Manual Evaluation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland. Association for Computational Linguistics, July, 2011, 1–11.
3. Callison-Burch, Chris. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, 2009, 286–295.
4. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Peterson, Kay, Przybocki, Mark és Zaidan, Omar. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden. Association for Computational Linguistics, July, 2010, 17–53.
5. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Schroeder, Josh. Findings of the 2009 Workshop on Statistical Machine Translation. In: *Proceedings of the EACL Workshop on Statistical Machine Translation*, 2009, 1–28.
6. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Zaidan, Omar. Findings of the 2011 Workshop on Statistical Machine Translation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland. Association for Computational Linguistics, July, 2011, 22–64.

7. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Zaidan, Omar F. szerk. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, July, 2011.
8. de Bruin, Wändi Bruine. Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations. *Acta Psychologica*, 2005, 118:245–260.
9. Doddington, George. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *HLT-01*, 2002.
10. Ester, Martin, Peter Kriegel, Hans, S, Jörg és Xu, Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, 1996, 226–231.
11. Fort, Karën, Adda, Gilles és Cohen, K. Bretonnel. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 2011, 37(2):413–420.
12. Giménez, Jésus. *IQMT. A Framework for Automatic Machine Translation Evaluation based on Human Likeness*. TALP Research Center, 2007.
13. Lin, Chin-Yew és Och, Franz Josef. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics, 2004.
14. Lleti, R., Ortiz, M.C., Sarabia, L.A. és Sánchez, M.S. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 2004, 515(1):87 – 100. Papers presented at the 5th COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM.
15. Mantonakis, Antonia, Rodero, Pauline, Lesschaeve, Isabelle és Hastie, Reid. Order In Choice: Effects of Serial Position on Preferences. *Psychological Science*, 2009, 20(11):1309–1312.
16. Melamed, I. Dan, Green, Ryan és Turian, Joseph P. Precision and recall of machine translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003, 61–63.
17. Papineni, Kishore, Roukos, Salim, Ward, Todd és Zhu, Wei-Jing. Bleu: A method for automatic evaluation of machine translation. In: *ACL-02*, Philadelphia, PA. 2002.
18. Sugar, Catherine A. és James, Gareth M. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 2003, (98):750–763.

Igei bővítménykeretek fordítási ekvivalenseinek kinyerése mélyen elemzett párhuzamos korpuszból

Héja Enikő[1], Takács Dávid[1], Sass Bálint[1]

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

Kivonat: Jelen cikk célja annak vizsgálata, hogy a mély szintaktikai elemzés növeli-e a fedést és a pontosságot igei szerkezetek fordítási megfelelőinek teljesen automatikus kinyerése során. Első lépésként a párhuzamos korpusz forrásnyelvi és célnyelvi oldalát külön-külön elemeztük, majd ebből nyertük ki az igei szerkezeteket egy felügyelet nélküli tanuló algoritmussal. Az így előállt igeiszerkezet-listát gyakorisági alapon szűrtük. A következő lépésben az igei szerkezeteket egytagú kifejezésekké vontuk össze a párhuzamos korpuszban, hogy az egytokenes igei szerkezetek az illesztési algoritmus bemeneteként szolgálhassanak. Eredményeink azt mutatják, hogy az alkalmazott módszer jól használható igei szerkezetek fordítási ekvivalenseinek detekciójára.

1 Bevezetés

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

A statisztikai gépi fordítás térhódításával jelentősen megnőtt a párhuzamos korpuszok szerepe a nyelvtechnológiában. Érdekes módon a lexikográfusok között nem eldöntött kérdés, hogy használhatóak-e a párhuzamos korpuszok emberi felhasználásra készülő szótárak előállítására (l. pl. [1]). Eddigi kisérleteink azt mutatták, hogy ha előfeldolgozásként szóillesztést végzünk, akkor az általunk javasolt módszer számos előnnyel rendelkezik a hagyományos lexikográfiai módszertannal szemben [5]. A javasolt módszer hátránya, hogy nem kezeli a többszavas kifejezéseket, így önmagában alkalmatlan a több szóból álló fordítási ekvivalensek kiszűrésére. Ennek a feladatnak a megoldása kiemelten fontos, hiszen egy szótárnak tartalmaznia kell azokat a többszavas kifejezéseket is, amelyek fordítása nem kompozicionális.

[6], illetve [9] alátámasztották, hogy egy előfeldolgozó modul hozzáadása elvileg lehetővé teszi a többszavas *ige* + *bővítmény* szerkezetek fordítási megfelelőinek automatikus kinyerését. Eredményként olyan összetett igei szerkezeteket kapunk, mint a

francia *faire partie de...* vagy holland megfelelője, a *deel uitmaken van...* (*részét képezi vminek*).

Feladatunk a módszert továbbfejleszteni úgy, hogy a kinyert párhuzamos igei szerkezetek felvehetőek legyenek a szótárba: vagyis a pontosság és a fedés növelésére egyaránt szükség van. Ennek érdekében a kutatás jelen szakaszában a [6]-ban, illetve [9]-ben leírtakat az alábbiak szerint módosítottuk. (1) Előre meghatározott igék helyett minden elegendően gyakori igét figyelembe vettünk, (2) minden igei szerkezet a vizsgálat tárgyát képezi, nemcsak azok a szerkezetek, amelyek főnévi lemmát is tartalmaznak, (3) részlegesen elemzett párhuzamos korpusz helyett mély szintaktikai annotációval rendelkező párhuzamos korpuszt használtunk az igei szerkezetek kinyeréséhez.

Azt várjuk, hogy a javasolt módszer az ige+bővítmény szerkezetek fordítási ekvivalenseinek teljesen automatikus meghatározásával hozzájárul a szótári tételek mikrostruktútájának kialakításához.

A következő szakaszban vázoljuk a munkafolyamatot (2), amely három fő lépésből áll: a párhuzamos korpusz szintaktikai elemzése (2.1), az igei szerkezetek automatikus kinyerése (2.2), valamint a protoszótár létrehozása (2.3). Majd eredményeinket mutatjuk be (3), végül pedig a konklúziókat és a további teendőket (4).

2 A munkafolyamat

A munkafolyamat három fő szakaszból áll. Az első lépésben elvégezzük a párhuzamos korpusz francia és holland részének mély szintaktikai elemzését, majd az így előállt frázisstruktúra-szerkezeteket az igei szerkezet kinyerő algoritmus által megkövetelt részleges függőségi elemzésekké konvertáljuk (2.1). A második lépésben a francia és holland igei szerkezetek egymástól független automatikus kinyerésével létrehozzuk a vizsgálandó igei szerkezetek listáját (2.2). A harmadik lépésben a kiválasztott többszavas igei szerkezeteket egytokenes kifejezésekké vonjuk össze, így ezek az illesztés bemenetül szolgálhatnak. Eredményül egy többszavas igei szerkezeteket tartalmazó protoszótárat kapunk (2.3).

2.1 A holland-francia párhuzamos korpusz szintaktikai elemzése

A kísérlethez a TLT-Centrale által fejlesztett Holland Párhuzamos Korpusz (DPC – Dutch Parallel Corpus) francia-holland alkorpuszát használtuk [7]. Az összesen 6,820,547 tokenes párhuzamos korpusz 186,945 illesztett egységet tartalmaz.

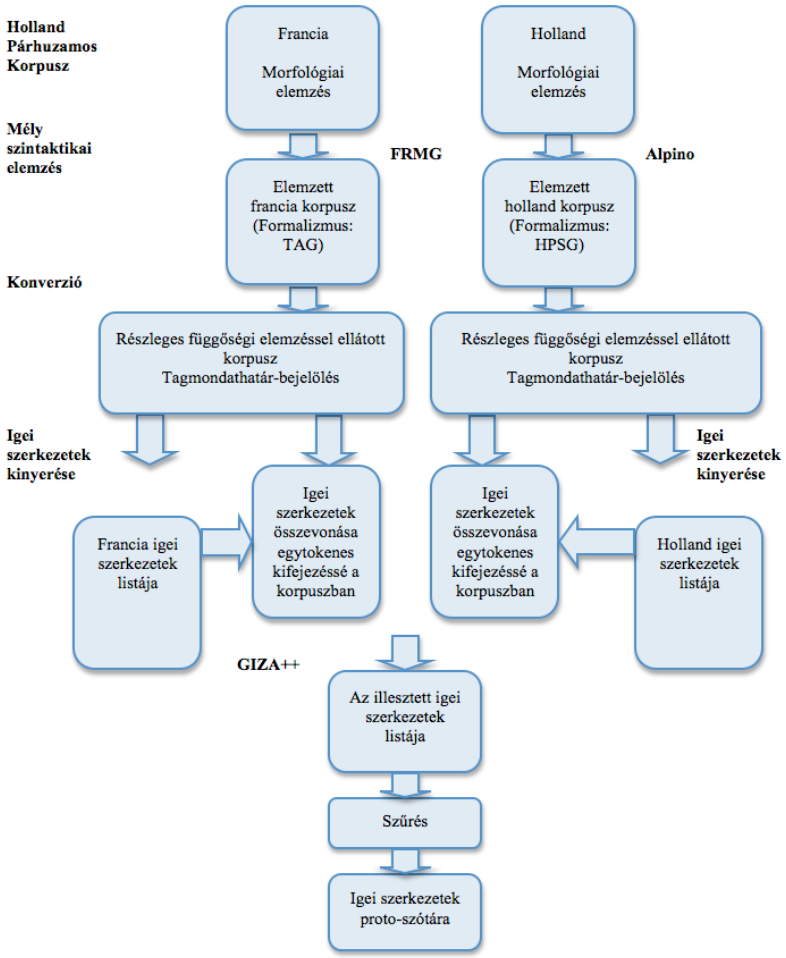
A holland esetében a HPSG elemzést végző Alpinot [2] használtuk, míg a francia korpuszt az FRMG hibrid TIG/TAG-parszerrel elemeztük[1] [11].

Az Alpino szabályalapú szintaktikai elemző a párhuzamos korpusz holland részkorpuszát részletetes annotációval látja el: megjelöli a frázisok határait és megadja a frázisok szintaktikai funkcióit. Ennek során felismeri az igéhez tartozó vonzatokat és partikulákat. Elvégzi a frázisok belső elemzését is: annotációval látja el a frázis fejét

[1] A szövegek elemzéséért köszönettel tartozunk Gábor Katának.

és a fejhez tartozó dependenseket. Az Alpino számunkra kiemelten fontos tulajdonsága, hogy felismeri a tagmondathatárokat, és megadja a tagmondatok egymáshoz való viszonyát (főmondat, mellékmondat, koordináció).

Az FRMG hasonló mélységű elemzést végez, mint az Alpino. Egy fontos különbség azonban, hogy az elemzés nem tartalmazott tagmondathatárra vonatkozó információt, ezért a tagmondathatár-felismerést saját szabályokkal végeztük el, amelyeket később részletezünk.



1. ábra: A munkafolyamat.

A következő lépésben az Alpino és az FRMG parszer kimenetét külön-külön részleges függőségi elemzéssé alakítottuk, hogy az elemzett korpuszok az igekinyerő algoritmus bemenetéül szolgálhassanak.

Az igei szerkezeteket kinyerő algoritmus abból az előfeltevésből indul ki, hogy (1) az ige jellemző bővítménykeretét mindig az a tagmondat tartalmazza, amelyben az ige előfordul, (2) egy tagmondat csak egy igéhez tartozó bővítményeket tartalmaz. Ebből következően a konverzió során meg kellett oldani a tagmondathatár-felismerést a francia esetében, valamint visszaállítani a teljes vagy eredeti bővítménykeretet azokban az esetekben, amikor erre szükség volt (pl. passzív igék, határozói és melléknévi igeneves szerkezetek). Ezeket utólagos átalakító szabályok hozzáadásával valósítottuk meg. A szabályok a részletes szintaktikai annotáción alapulnak, amely azt is jelöli, ha az ige valamilyen képzett formában szerepel (passzív, illetve különféle igeneves szerkezetek).

A holland esetében az alábbi átalakításokat végeztük el:

(1) Passzív szerkezetek aktívvá alakítása
(2) Segédigék törlése az összetett igeidők esetében
(3) Melléknévi igeneves szerkezetek konverziója tagmondattá

A francia elemzés esetében a fentieken túl a tagmondathatárok bejelölésére is szükség volt, így a fenti szabályokhoz továbbiakat adtunk hozzá:

(4) Melléknévi igeneves szerkezetek önálló tagmondatot alkotnak
(5) A vonatkozó névmások előtt legyen tagmondathatár
(6) A főnévi igenév előtt is van tagmondathatár, ha a főnévi igenév előtt valamilyen prepozíció áll (*de*, *pour*, *sans*, *en vue de*, *à* stb.)
(7) Legyen tagmondathatár koordinált tagmondatokat összekötő kötőszavak helyén (*et* - *és*, *puis* - *aztán*, *ou* - *vagy*, stb.)
(8) Legyen tagmondathatár az alárendelt mondatokat bevezető kötőszavak helyén (*que* - *hogy*, *quand*, *pendant que* - *amikor*, stb)
(9) Ha két ige között nincs tagmondathatár, akkor szúrjon be tagmondathatárt vessző, pontos vessző vagy kettőspont esetén.

Végül el kellett döntenünk, hogy a részletes szintaktikai annotáció mely jegyeit kívánjuk figyelembe venni az igei bővítménykeretek kinyeréséhez. Itt két ellentmondó követelménynek kell eleget tenni: egyfelől, minél több jegyet tartunk meg az eredeti elemzésből, annál részletesebben karakterizálhatjuk az igei bővítménykereteket. Másfelől, túl sok jegy alkalmazása jelentősen ronthatja az eredményeket, hiszen az irreleváns címkék növelik az adatok diverzitását. A típusok számának növekedésével párhuzamosan csökken a típusok előfordulási gyakorisága, ez pedig rontja a generált szótár minőségét.

Első megközelítésben megtartottuk az igét, az igével közvetlenül függőségi viszonyban levő összetevő fejét, valamint a fej dependensei közül az esetleges mellékneveket, illetve egyéb módosítókat a vonzatos főnevek esetében, míg a névelőket elhagytuk. A koordinált szerkezetekből (ha nem koordinált tagmondatokról volt szó) mindig csak az első összetevőt őriztük meg. A következő részben látni fogjuk, hogy bizonyos esetekben ez is túl részletes elemzésnek bizonyult, így további empirikus vizsgálatot igényel, hogy pontosan milyen mélységű elemzést érdemes végezni.

2.2 Az igei szerkezetek automatikus kinyerése

A releváns francia és holland *ige+bővítmény* szerkezeteket automatikusan nyertük ki a párhuzamos korpusz megfelelő egynyelvű részeiből. Az igei szerkezetek automatikus kinyerése során az ige mellett meglévő jellegzetes bővítménykereteket határozzuk meg a tagmondatokban a gyakori részkeretek rendszerezett összeszámlálása révén. A [9]-ben részletesen leírt módszer előnye abban rejlik, hogy felismeri, hogy melyik bővítménynél lényegi elem a konkrét fej és melyiknél csak az ige-bővítmény viszony. Így egyszerre képes meghatározni az összetett igéket és a vonzatkereteket is. A *haszнot húz* vmiből szerkezet esetén például felfedezi, hogy a lexikálisan kötött tárgy mellett egy *-ból*/*-ből* esetragos vonzat is szerepel az igei keretben.

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

Az igeiszerkezet-kinyerő módszer alapvetően tagmondatokra bontott, szintaktikailag részlegesen elemzett korpuszon dolgozik. A tagmondatok egy igét és annak bővítményeit kell, hogy tartalmazzák, a szintaktikai elemzés pedig meg kell hogy állapítsa a tagmondat igéjét, a bővítmények fejét, valamint a bővítmények igéhez való szintaktikai viszonyát. A szintaktikai viszonyt a megfelelő esetrag vagy egy elöljáró-szó jelöli. Mivel az igei szerkezet fogalmát a vonzatkeretnél tágabban értjük, mély szintaktikai annotációval rendelkező korpuszokon is futtatható az algoritmus úgy, hogy többletinformációt nyerjünk ki belőle (az algoritmus az igei vonzatokon túl a jellemző bővítményeket is megadja – akkor is, ha azok szabad határozók – sőt az igei szerkezet részét képezik a jellemző lexikai fejek is). Az 1. és 2. táblázatban példákat láthatunk az automatikusan kinyert igei szerkezetekre.

1. táblázat. A holland *'gebruiken'* ige négy leggyakoribb szerkezete.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
gebruik obj1	470	*használ vmit*
gebruik niet=mod:ADV obj1	159	*nem használ vmit*
gebruik obj1 obj1_ADJ	104	*használ vmilyen vmit*
gebruik obj1 als=predc:CP	95	*úgy használ valamit, hogy ...*

Az 1. táblázat mutatja azt is, hogy a részletes elemzés eredményeképpen a *'nem használ vmit'* illetve a *'használ valamilyen vmit'* is gyakori kereteknek minősülnek, ám felvételük egy igei kereteket tartalmazó szótárba a keretek kompozicionalitása miatt nem indokolt. A megfelelő bővítmények elhagyásával mindkét keret a *'használ vmit'* kerethez sorolódna, így növelve ezen keret gyakoriságát a korpuszban, és ezáltal a megfelelő fordítási ekvivalensek kinyerésének a valószínűségét.

A 2. táblázatban szintén szerepelnek irreleváns keretek is a mély szintaktikai elemzés eredményeként:

2. táblázat. A holland *'geven'* ige négy leggyakoribb szerkezete.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
geef obj1	170	*ad vmit*
geef obj1 obj1_ADJ	80	*ad vmilyen vmit*
geef aan:obj2 obj1	78	*ad vkinek vmit* (indirekt)
geef obj1 obj2	72	*ad vkinek vmit* (direkt)

A táblázatban látszik, hogy ha a tárgyat módosító jelzőt nem vennénk figyelembe, akkor a *'geven'* leggyakoribb szerkezetei pontosan az „elvártak" lennének.

A 3. táblázatban található példa már lexikai bővítményt is tartalmaz a jellemző esetkeret mellett. Ez a mély elemzés egy másik nem kívánt hatását szemlélteti: a parszer ugyanahhoz a felszíni szerkezethez bizonyos esetekben különböző annotációkat rendel, és ez – függetlenül attól, hogy melyik a jó elemzés – megint csak a rendelkezésre álló adatok csökkenéséhez vezet.

3. táblázat: A holland *'een beroep doen op'* elemzései.

Szerkezet	**Gyakoriság**	**Magyar megfelelő**
doe beroep=obj1 obj1_op	72	*felhívást tenni vmire*
doe beroep=obj1 op:mod	39	*felhívást tenni vmire*

Az első esetben a holland *'op'* (-rA) az ige tárgyának, a *'beroep'*-nak, míg a második esetben magának az igének a bővítménye. További probléma, hogy ennek a szerkezetnek a névelő (*een*) kötelezően része, de ez mindkét keretből hiányzik.

A következő lépésben automatikusan választottuk ki azokat az igei szerkezeteket, amelyeket akár forrásnyelvi, akár célnyelvi oldalon a szótárban szerepeltetni akartunk. Egy lehetséges megközelítés, hogy heurisztikát dolgozunk ki a „*lexikográfiai szempontból érdekes*" bővítménykeretek automatikus szűrésére. Mivel fordítási feladatról van szó, a kompozicionalitás ebben az esetben nem önmagában, hanem egy másik nyelv függvényében értelmezhető. A javasolt módszer egyik kiemelten fontos tulajdonsága a nyelvfüggetlenség. Így elképzelhető, hogy *A* nyelv egy igei szerkezete kompozicionálisan fordul le *B* nyelvre, de nem kompozicionális *C* nyelven. Ebben az esetben tehát azt kell mondanunk, hogy *A* nyelv adott kifejezése lexikográfiailag érdekes az első esetben, és érdektelen a másodikban. A nyelvfüggetlenség miatt járhatóbb megközelítési módnak tűnik az igei szerkezeteket gyakorisági alapon szűrni. Ebben az esetben feltételezzük, hogy egy szótárban a gyakran előforduló jelenségeket célszerű rögzíteni, függetlenül attól, hogy ezek fordítása transzparens-e vagy sem egy másik nyelven.

Így tehát az automatikusan kinyert igei szerkezetek közül azokat vettük fel a listánkba, amelyek legalább ötször előfordultak a párhuzamos korpusz megfelelő oldalán. Ennek a kritériumnak a holland oldalon 289 ige felelt meg, összesen 5804 kerettel, míg a francia igelista 391 igét tartalmazott 5987 különböző kerettel.

2.3 A keretek azonosítása, összevonása és a protoszótár létrehozása

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

[6]-ban csak azokat a szerkezeteket vizsgáltuk, amelyek az igén kívül is tartalmaztak valamilyen kötött lexikai elemet. Az igei szerkezetek kiválasztásakor nem törekedtünk a teljes bővítménykeret megőrzésére, így bizonyos esetekben a kitöltetlen – vagyis tipikus főnévi lemma nélkül álló – esetragokat elhagytuk. Ennek oka egyfelől az volt, hogy az eltérő igei szerkezetek összevonásával növelhettük a szükséges adatok mennyiségét. Másfelől, mivel az illesztés bemeneti korpusza nem tartalmazott sem részleges szintaktikai elemzést, sem tagmondatfelismerést, az esetek egy jelentős részében lehetetlen volt pontosan azonosítani a megfelelő prepozíciót.

Ezzel szemben a jelen kísérlet célja minden megfelelően gyakori igei bővítménykerethez fordítási megfelelőt találni, függetlenül attól, hogy tartalmaz-e kötött lexikai elemet. Az ige bővítményeit értelemszerűen csak az igéhez tartozó tagmondatban kerestük. Az illeszkedő igei keretek közül a leghosszabbakat választottuk, és ezt vontuk össze a párhuzamos korpusz elemzett változatában.

Míg az említett első kísérletben a 126 francia igei szerkezet összesen 7805-ször, és a 146 holland igei szerkezet 8029-szer fordult elő a párhuzamos korpuszban, addig a jelen kísérletben 170,229 illeszkedő francia bővítménykeret és 207,610 illeszkedő holland bővítménykeretet találtunk a párhuzamos korpuszban.

A továbbiakban a kiválogatott többszavas igei kifejezéseket egy tokenként kezeltük és így közvetlenül alkalmaztuk az működő illesztő algoritmust.

Az illesztést a GIZA++ szoftverrel végeztük [8], amely az illesztés során fordításjelölteket hoz létre, úgy, hogy a forrásnyelvi és célnyelvi lemmapárokhoz fordítási valószínűséget rendel. A fordítási valószínűség a célnyelvi és forrásnyelvi szópár feltételes valószínűségének közelítése – $P(\text{szó}_{\text{cél}}|\text{szó}_{\text{forrás}})$ – az EM (expectation maximization) algoritmus alapján [3].

A protoszótárak kiindulási alapját az így kinyert fordítási jelöltek és fordítási valószínűségeik képezték. Mivel a fordítási valószínűség 0-tól 1-ig bármilyen értéket felvehet, ebben a szakaszban még sok helytelen fordítási jelöltünk van. Ezért szükség van olyan szűrők bevezetésére, amelyek lehetővé teszik a legjobb fordításjelöltek automatikus kiválasztását a lehető legtöbb helyes fordításjelölt megtartásával. Eddigi tapasztalataink azt mutatták [5], hogy a fordítási valószínűségek és a forrásnyelvi, illetve célnyelvi korpuszgyakorisági adatok együttesen már jól használhatóak az eredmények szűrésére. Így a protoszótárban az alábbi adatok szerepelnek:

4. táblázat. Francia és holland fordítási jelöltpárok és paramétereik.

Kifejezés$_{forrás}$	**Kifejezés$_{cél}$**	**P(szó$_{cél}$|szó$_{forrás}$)**	**Gyak$_f$**	**Gyak$_c$**
prendre médicament=obj1	neem_in genees_middel=obj1	0.377261	53	32
	gebruik genees_middel=obj1	0.102349	53	21
	start gebruik=met:cmp met:cmp_van	0.0971227	53	28
	sta onder_invloed=particle drug=van:cmp	0.050697	53	11

A 4. táblázatban látható, hogy a francia *'prendre médicament*' (*gyógyszert bevenni*) szerkezetnek a legvalószínűbb holland megfelelője az '*geneesmiddel innemen*'. Ezt követi a *'geneesmiddel gebruiken'* (*gyógyszert használni*). A *'start met gebruik van'* nem teljes keret (*elkezdeni a használatát valaminek*) szintén releváns fordításnak tekinthető. A legkevésbé valószínű, ám lexikográfiai szempontból még érdekes fordítás a '*staan onder invloed van drug*' (drog hatása alatt állni).

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

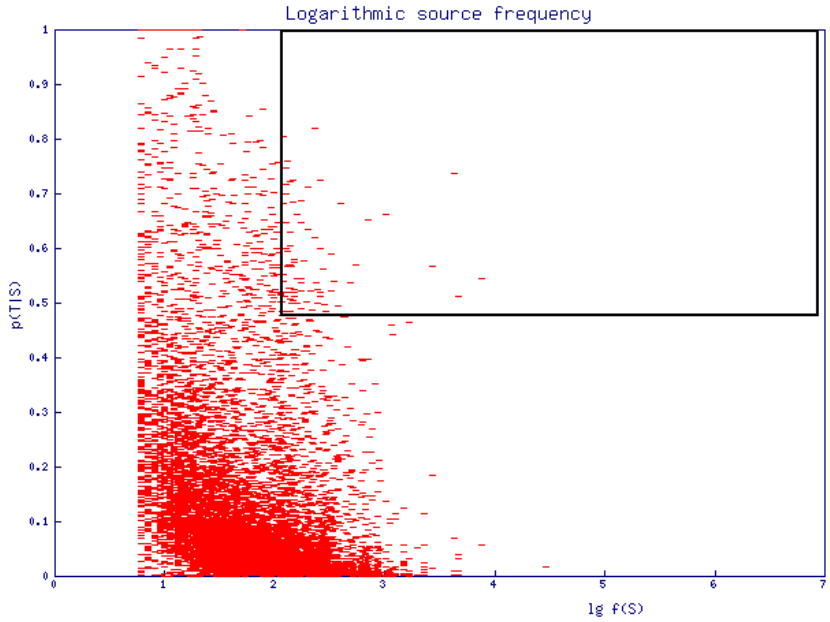
(1) A forrásnyelvi és a célnyelvi szónak is legalább 5-ször elő kell fordulnia a párhuzamos korpuszban. Ez a feltétel szükséges ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez.

(2) Hasonló gyakoriságú szavak esetén magasabb fordítási valószínűségi küszöb alkalmazása esetén magasabb lesz a jó vagy hasznos fordítási jelöltek aránya.

(3) A paraméterek beállíthatóak úgy, hogy gyakoribb forrásnyelvi szavak esetén alacsonyabb fordítási valószínűségi küszöb körülbelül ugyanolyan arányban eredményezzen jó vagy hasznos fordítási jelölteket, mint a ritkább szavak esetében egy magasabb fordítási valószínűségi küszöb.

3 Kiértékelés

Első lépésben olyan paraméterbeállítást választottunk, amely mellett feltételezhetően magas a jó vagy hasznos fordításjelöltek aránya. Így megmutathatjuk, hogy van olyan paraméterbeállítás, amely magas pontosságot eredményez, amelyből kiindulva a fedés – legalábbis részben – növelhető a paraméterbeállítások finomításával. A 2. ábrán látható a francia-holland igekeret-jelöltpárok eloszlása a forrásnyelvi kifejezés logaritmikus gyakorisága és a megfelelő fordítási valószínűség szerint. A fekete téglalap területére eső fordításjelölteket értékeltük ki. A legalább 100-szor előforduló forrásnyelvi és a célnyelvi lemmák közül azokat a fordítási jelöltpárokat választottuk ki, amelyek legalább 0,44 fordítási valószínűséggel rendelkeznek. Ezek közül 100 megfelelő keretet értékeltünk ki.



2. ábra: A francia-holland igekeret-jelöltpárok eloszlása a forrásnyelvi kifejezés logaritmikus gyakorisága és a megfelelő fordítási valószínűség szerint. A kiértékelési tartomány.

A kiértékelést két szempont alapján végeztük: egyfelől figyelembe vettük, hogy az algoritmus megtalálta-e a megfelelő igét. Másfelől azt is vizsgáltuk, hogy az illesztés a teljes keretek között történt-e. Összesen 46 esetben volt megfelelő a fordítás, úgy, hogy mind a forrásnyelvi, mind a célnyelvi oldalon teljes igei bővítménykeretek szerepeltek (46%). Ebből 54 esetben a megfelelő ige állt mindkét oldalon, de hiányos volt valamelyik, esetleg mindkét ige kerete (21 esetben a forrásnyelvi ige, 9 esetben a célnyelvi ige, 24 esetben mindkét ige kerete hiányzott).

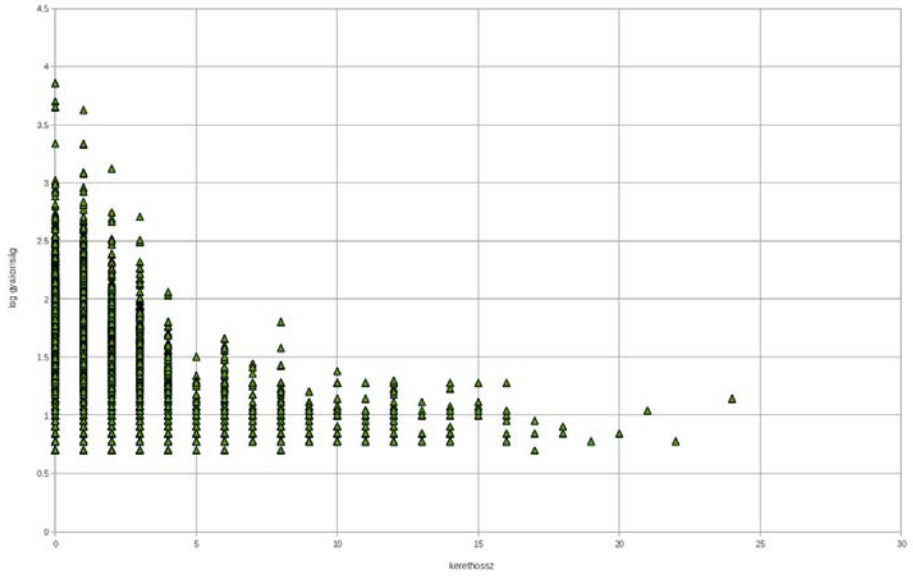
A kiértékelt keretek többnyire egy bővítményt tartalmaztak, általában egy tárgyat, de előfordultak több bővítményt tartalmazó keretek is, pl.:

avoir besoin=obj1 de:cpl *hebben obj1 nodig=predc:ADJ*
(*vkinek szüksége van vmire*)

A legjobb fordításjelöltek kiértékelése során kérdésként merült fel, hogy hogyan növelhető a jó fordításjelöltek között a teljes keretek száma? Erre egy lehetséges megoldás, hogy valamilyen alkalmas heurisztikával szűrjük a rossz kereteket az automatikusan előállított bővítménykeretlistából. Kérdés, hogy esetünkben mi számít „rossz” bővítménykeretnek. Mivel célunk általános célú szótárak építése, rossz keretnek minősülhetnek a „túl hosszú” keretek, amelyek jellemzően a korpusz valamely szaknyelvi részében (orvosi, informatikai) fordulnak elő nagy számmal. Az ilyen

keretek illesztésével a rövidebb, általánosabb kereteket kizárjuk. A leghosszabb francia keret 24 egység hosszú[2] és 14-szer fordul elő orvosi szövegekben.

A 3. ábra a francia esetében azt mutatja, hogy az egyes kerethossz alapján csoportosított kerettípusokból hány van, és az egyes keretek hányszor fordulnak elő a francia részkorpuszában.



3. ábra: A kerethossz alapján csoportosított kerettípusok száma és az egyes keretek gyakorisága a párhuzamos korpusz francia részkorpuszában.

Az ábrán jól látszik, hogy a 8 hosszúságú keretek között még vannak olyanok, amelyek viszonylag gyakoriak, így ezeket még érdemes lehet megtartani a szótár generálásánál, de az ennél hosszabbakat már nem. Mindazonáltal a keretek manuális vizsgálata azt mutatja, hogy még a 8 hosszúságú keretek is nagyon specifikusak, és egy általános célú szótár esetében legfeljebb 5 hosszúságú kereteket érdemes figyelembe venni. További empirikus vizsgálatokat igényel, hogy ez a heurisztika növeli-e a teljes keretek arányát a jó fordítási jelöltek között.

Az alkalmazott módszer érdekessége, hogy az igei szerkezetek kinyerése és a fordítási jelöltek kinyerése is felügyelet nélküli tanulással történik – vagyis az emberi intuíció kiküszöbölésével. Így a kiértékelés során azt is vizsgáltuk, hogy a kapott szerkezeteket mennyire jól karakterizálnak egy igét (*mettre*):

Az illesztés eredményeképpen előállt protoszótárból csak a 0,02-nál valószínűbb és legalább 5-ször előforduló párokat hagytuk meg. A ‘*mettre’* 5706 előfordulása 65 különböző bővítményi kerettel fordul elő. Ezek az 5611 esetben előforduló 132 holland kerettel összesen 151 fordítási párba rendeződnek. Ezeket részletesen kiértékel-

[2] A keretek hosszát a bővítmények számával mérjük: az igekinyerő algoritmusnak megfelelően a bővítmények szintaktikai funkcióját jelző morfémák és a keretben szereplő lexikai elemek ugyanolyan súllyal számítanak.

tük. A kiértékelés során igen-nem-döntést hoztunk a megfeleltetés helyességéről aszerint, hogy az adott francia keretet lehetséges-e a hozzá párosított holland kerettel fordítani a korpuszban található valamely kontextusban. Megengedtük a hiányos kereteket is, ha a konkordanciában úgy láttuk, hogy megfelelően bővíthetőek. A 151 keret 62%-át ítéltük helyesnek.

Mind a francia, mind a holland oldalon megjelöltük a hiányos kereteket, amelyek nem önálló szótári tételek, de ilyenné bővíthetőek. A '*mettre'* 65 kerete közül 10 olyan volt, amelynek csak rossz fordításai voltak, 55-höz (a keretek 85%-ához) találtunk egy vagy több helyes fordítást.

Érdekes, hogy a helytelen fordítási párok jellemzően (78% teljes francia keret és 86% teljes holland keret) a teljes keretekhez adódtak. Ezzel szemben a helyes fordítási pároknak csak 59%, illetve 63%-a teljes keret. Tehát egyértelmű trade-off van a keretek jólillesztettsége és a pontosság között.

4 Konklúziók és további teendők

Eredményeinkből látszik, hogy a javasolt módszer hasznos ötletekkel láthatja el a lexikográfusokat arra vonatkozóan, hogy mely igei tételeket kell szerepeltetni a szótárban, illetve ezen tételeknek milyen fordításai lehetnek. Mindazonáltal, a keretek sok esetben hiányosak, így sokszor kell a megfelelő konkordanciára támaszkodni a helyes igei szerkezetek visszaállításához. Így a jövőben az elsődleges célunk az, hogy a fordításjelöltek között minél teljesebb keretek szerepeljenek.

Egy lehetséges megoldás, hogy valamilyen alkalmas heurisztikával szűrjük a rossz kereteket az automatikusan előállított bővítménykeretlistából. Mivel célunk általános célú szótárak készítése, első lépésként azt kívánjuk vizsgálni, hogy a hosszú keretek rövidebb keretek alá rendezésével növelhető-e a teljes keretek aránya a fordítási jelöltpárok között.

Az eredmények általános pontosságának a növeléséhez pedig szükséges az adatok diverzitásának csökkentése, hogy minél több adat álljon az illesztő algoritmus rendelkezésére. Ehhez tovább kell szűkíteni az igeiszerkezet-algoritmus bemenetéül szolgáló nyelvtani kategóriák körét, valamint a teljes szintaktikai annotációt elegendő csak az igei szerkezeteken belül megtartani.

Bibliográfia

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
2. Bouma, G., Noord, van G., Malouf, R.: Alpino: Wide coverage computational analysis of Dutch. In: Daelemans, W., Sima'an, K., Veenstra, J., Zavrel, J. (eds): Computational Linguistics in the Netherlands 2000. Rodolpi, Amsterdam (2001) 45–59
3. Dempster, A. P., Laird, N. M., Rubin. D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B Vol. 39, No.1 (1977) 1–22

4. É. Kiss, K.: Mondattan. In: É. Kiss, K., Kiefer, F., Siptár, P. (eds.): Új magyar nyelvtan. Osiris Kiadó, Budapest (2003) 15–184
5. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography. In: Proceedings of the LREC2010 Conference. La Valletta, Malta (2010) 2798–2805
6. Héja E., Sass B.: Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban. In: MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2010) 80–90
7. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007. Birmingham, United Kingdom (2007)
8. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics Vol. 29, No. 1 (2003) 19–51
9. Sass, B.: A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In: Angelova G. et al. (eds.): Proceedings of RANLP 2009. Borovec, Bulgária (2009) 399–403
10. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2010) 102-110
11. Villemonte de la Clergerie: Convertir des dérivations TAG en dépendances. In: Atala, (ed.):17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010 (2010)

Félig kompozicionális szerkezetek automatikus azonosítása magyar és angol nyelven

Vincze Veronika[1], Nagy T. István[2], Zsibrita János[2]

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

Kivonat Jelen munkában bemutatjuk szabályalapú és gépi tanult módszereken alapuló megközelítéseinket, melyek mind angol, mind magyar nyelven képesek a félig kompozicionális szerkezetek folyó szövegben történő automatikus azonosítására. Eredményeink azt igazolják, hogy a sekély morfológiai elemzésre épülő módszereink mellett a szintaktikai információ is nagyban képes segíteni a félig kompozicionális szerkezetek automatikus azonosítását. Cikkünkben kitérünk a feladat angol és magyar nyelvű sajátságaira is.

Kulcsszavak: többszavas kifejezések, lexikális szemantika, többnyelvűség, FXtagger

1. Bevezetés

A természetes nyelvi feldolgozásban, különösen a gépi fordítás és fordítástámogatás területén az egyik legnehezebb problémát a többszavas kifejezések megfelelő kezelése jelenti. A többszavas kifejezések sikeres kezelésének első lépése, hogy felismerjük őket a folyó szövegben. Ebben a munkában a többszavas kifejezések egy altípusának, a félig kompozicionális szerkezeteknek automatikus felismerésére koncentrálunk.

A félig kompozicionális szerkezetek (FX-ek) olyan, főnévből és igéből álló többszavas kifejezések, ahol a szemantikai fej a főnév, míg az ige pusztán csak a szerkezet igeiségéért felel. Mivel jelentésük nem teljesen kompozicionális, a szerkezet elemeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Emellett a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*) [1], így azonosításuk nem valósulhat meg pusztán szintaktikai mintákat figyelembe véve. Végül, mivel a szerkezet szintaktikai és szemantikai feje nem azonos, a szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni - az angol vonzatos igékhez (phrasal verbs) hasonlóan.

A fenti okokból kifolyólag a félig kompozicionális szerkezetek kezelése különleges figyelmet érdemel a természetes nyelvi alkalmazásokban. Ennek első lépéseként azonosítani kell őket, mely célhoz különféle algoritmusok fejlesztése segíthet hozzá. Ennek megfelelően először szabályalapú megközelítéseket definiálunk, majd ezek eredményeire alapozva gépi tanuló módszerek segítségével is azonosítjuk a félig kompozicionális szerkezeteket.

2. Kapcsolódó munkák

A félig kompozicionális szerkezetek automatikus azonosítására, illetve a főnév + ige szerkezetek osztályokba sorolására már több szerző is kísérletet tett.

Van de Cruys és Moirón [2] szemantikai alapokon nyugvó rendszere igeprepozíció-főnév kombinációkat azonosít holland szövegekben. Módszerük az ige és a főnév szelekciós megkötéseire épül, illetve az igével együtt előforduló főnevek szemantikai osztályát is figyelembe veszik.

Cook és munkatársai [3] angol ige + főnév szerkezetek szó szerinti és idiomatikus használatát különítik el egymástól. Feltevésük szerint idiomatikus használatban főként a szerkezet szótári alakja fordul elő, míg szó szerinti használatban a szerkezet nagyobb szintaktikai változatosságot mutat. A szerkezet szintaktikai rögzítettségét kihasználó felügyelet nélküli osztályozó módszerük 72%-os eredményt ér el.

Bannard [4] szintén angol nyelvű ige + főnév szerkezeteket osztályoz szintaktikai rögzítettségük alapján. Az általa használt jellemzők közé tartozik a főnév névelőzhetősége, módosíthatósága, a szerkezet szenvedő szerkezetben való előfordulása stb.

Samardžić és Merlo [5] angol-német párhuzamos korpuszokban előforduló félig kompozicionális szerkezeteket vizsgálnak. Eredményeik szerint a szerkezetek párhuzamosításánál különösen nagy szerepet játszanak a gyakorisági adatok mellett a szerkezetek nyelvi jellemzői is, például a kompozicionalitás foka.

Gurrutxaga és Alegria [6] baszk nyelvű szövegekből nyernek ki idiomatikus és félig kompozicionális főnév + ige szerkezeteket statisztikai módszerek segítségével. Mivel a baszk szabad szórendű nyelv, azzal az előzetes feltételezéssel éltek, hogy az ige tágabb környezetét nézve javulni fognak az eredmények, azonban kísérleteik ezt nem támasztották alá.

Tu és Roth [7] ige + főnév párokat osztályoznak aszerint, hogy félig kompozicionális szerkezetek-e vagy sem. Mind környezeti, mind statisztikai jellemzőkkel dolgoznak, és megállapításuk szerint a többértelmű példákon a lokális környezeti jellemzők használata vezet a legjobb eredményhez.

Sass Bálint [8] beszámol egy igei szerkezetek párhuzamos korpuszból való kinyerésére szolgáló eljárásról, mely egy korábbi, igéket és azok bővítményeit kinyerő algoritmusra épül. A módszer lényege, hogy a tagmondatok igéit egymás mellé rendelve egy komplex ige jön létre, melyhez a bővítményeket halmazként rendeljük hozzá, felcímkézve őket aszerint, hogy melyik nyelvű részkorpuszból származnak. Az így kapott reprezentációból az eredeti algoritmus segítségével lehet kigyűjteni az egyes nyelvekre jellemző igei szerkezeteket.

A félig kompozicionális szerkezetek automatikus azonosítását célzó módszerek nagy része kiindulási alapnak tekinti a szintaxist, azaz általában ige-tárgy párokat osztályoznak [3,4,9,7]. Ezzel szemben mi nem a szintaktikai mintázatok alapján megszűrt FX-jelölteket szeretnénk osztályozni, hanem folyó szövegben szeretnénk azonosítani őket, nem feltétlenül szintaktikai információk segítségével. Kísérleteink közben azonban kiemelt figyelmet szentelünk a szintaktikai információk hozzáadott értékének.

3. A félig kompozicionális szerkezetek automatikus felismerése

A félig kompozicionális szerkezetek automatikus azonosítására szabályalapú és gépi tanulási módszereket is definiáltunk. Angol és magyar nyelvre alapjában véve ugyanazokat az eljárásokat alkalmaztuk, természetesen figyelembe véve az adott nyelv sajátosságait.

Módszereink kiértékeléséhez három korpuszt használtunk. A SzegedParalellFX párhuzamos korpusz [10] angol és magyar nyelven ugyanazokat a szövegeket tartalmazza, melyekben összesen 1100 angol nyelvű és 1112 magyar nyelvű FX található. A Szeged Korpuszban szintén be vannak jelölve a félig kompozicionális szerkezetek [11]. Kísérleteinkhez a sajtónyelvi részkorpuszokat használtuk. Az angol nyelvű Wiki50 korpuszban [1] többszavas kifejezések és névelemek vannak annotálva, így a félig kompozicionális szerkezetek is be vannak jelölve. Noha a korpuszokban a félig kompozicionális szerkezetek melléknévi igenévi és főnévi alakjai is be vannak jelölve, jelen munkánkban csak az igei alakok felismerésére koncentrálunk. A felhasznált korpuszok adatait az 1. táblázat mutatja.

1. táblázat. A felhasznált korpuszok adatai

Korpusz	**Mondat**	**Token**	**Igei FX**
Wiki50 (angol)	4.350	114.570	368
SzegedParallelFX (angol)	14.262	298.948	745
SzegedParallelFX (magyar)	14.262	240.399	753
Szeged Treebank (újságcikkek - magyar)	10.210	182.172	458

3.1. Szabályalapú módszerek

Számos szabályt fogalmaztunk meg a félig kompozicionális szerkezetek automatikus azonosítására. Az angol nyelvű szövegeket a Stanford elemzési lánc segítségével tokenizáltuk, majd elemeztük szófajilag [12] és szintaktikailag [13]. A SzegedParalellFX magyar nyelvű szövegeit a `magyarlanc` [14] csomaggal tokenizáltuk és elemeztük szófajilag. A Szeged Korpuszból származó szövegek esetén az etalon szófaji és dependenciaelemzésekre hagyatkoztunk, illetve az összevethetőség kedvéért a `magyarlanc` által nyújtott szófaji elemzésekkel is végeztünk kísérleteket.

A **POS-szabályok** („POS”) módszer esetében különféle szófaji mintákat adtunk meg, például VB.? (NN|NNS) angolra vagy N V a magyarra. Amennyiben ezek illeszkedtek a szöveg egy szegmensére, azt megjelöltük mint félig kompozicionális szerkezetet. Mivel további módszereink morfológiai információkra épülnek, pontosabban az ige vagy a főnév természetére tesznek megszorításokat, a POS-szabályokra való illeszkedés előfeltétele a többi módszer alkalmazhatóságának.

A **végződés** („vég”) módszer alapja, hogy az FX-ek főnévi komponense legtöbbször igéből képzett főnév. Ebben az esetben azokat az FX-jelölteket fogadtuk el, amelyekre illeszkedett egy szófaji minta, és a főnév az előre definiált n-gramok (képzők) egyikében végződött.

A **leggyakoribb ige** („ige”) módszer azon megfigyelésen alapszik, hogy általában a leggyakoribb igék szerepelnek funkcióigeként (az angolban a *do*, *make*, *take* stb., míg a magyarban *ad*, *vesz*, *hoz* stb.). így azokat az FX-jelölteket fogadtuk el, amelyek illeszkedtek a szófaji mintákra, és az igei komponens lemmája megegyezett az előre megadott leggyakoribb igék egyikével.

A **szótő** („tő”) módszer a főnév szótövét vizsgálja. Mint fentebb említettük, a főnévi komponens igen gyakran igéből származik, így az angolban azt néztük meg a Porter stemmert használva [15], hogy a főnév szótöve egybeesik-e egy igei szótővel (*to make a decision - to decide*) vagy maga a főnév egybeesik-e egy igével (*to have a walk - to walk*). A magyarban pedig a hunmorph elemző [16] segítségével állapítottuk meg a főnév szótövét, és vizsgáltuk meg, hogy annak van-e igei elemzése.

A félig kompozicionális szerkezetek azonosításában a szintaktikai információk is hasznosak lehetnek. Az angolban a szerkezet két tagja között általában `dobj` vagy `prep` viszony szerepel (tárgyi vagy prepozíciós vonzat esetében), míg a magyarban `obj` vagy `obl` (tárgy vagy egyéb argumentum). A **szintaxis** módszert alkalmazva azokat az FX-jelölteket fogadtuk el, amelyek tagjai a fenti relációk egyikében álltak egymással.

A fenti módszereket kombináltuk is egymással: vagyis vettük a különféle módszerek unióját $\cup$ (egy potenciális FX jelölt abban az esetben került elfogadásra, amennyiben legalább az egyik módszer elfogadta azt), és a metszetüket $\cap$ (csak akkor jelöltünk szóösszetételt FX-nek, amennyiben minden szabály elfogadta azt). Eredményeinket a 2. táblázat szemlélteti.

3.2. A szabályalapú módszerek eredményei

A 3. táblázat mutatja a szabályalapú módszereink eredményét a négy felhasznált korpuszon. Jól látszik, hogy három korpusz esetében a leggyakoribb ige módszer bizonyul a legsikeresebbnek, jóval magasabb F-mértéket ér el, mint a többi módszer vagy azok kombinációi. Az egyetlen kivételt a SzegedParalellFX angol állománya jelenti, ahol is az ige és tő módszerek metszete a legeredményesebb. Ez valószínűleg annak köszönhető, hogy a korpuszban nagy arányban fordulnak elő tipikus főnév + tipikus ige kombinációk. A végződés jellemző a SzegedParalellFX-en bizonyul hasznos információnak, a másik két korpuszon önmagában még ront

2. táblázat. Szabályalapú megközelítések eredményei, fedés/pontosság/F-mérték.

Megközelítés	Wiki50			ParalellFX angol			ParalellFX magyar			Szeged Treebank		
POS	77,14	6,32	11,68	79,40	5,07	9,52	65,55	7,67	13,74	74,56	5,75	10,69
Vég	17,14	9,47	12,20	15,24	10,5	12,43	21,45	12,79	16,02	19,30	6,53	9,76
Ige	55,24	34,32	42,34	54,56	28,81	37,73	43,83	30,19	35,76	58,77	24,28	34,36
Tő	54,29	7,72	14,64	61,55	7,66	13,62	21,05	16,14	18,27	16,67	7,85	10,67
Vég ∩ Ige	9,52	43,48	15,64	10,24	48,31	16,90	15,15	40,36	22,03	18,42	32,81	23,60
Vég ∪ Ige	62,86	19,64	29,93	59,64	19,02	28,84	50,13	18,21	26,71	59,65	12,39	20,51
Vég ∩ Tő	14,29	10,79	12,30	11,07	11,14	11,10	19,30	16,31	17,68	15,79	8,37	10,94
Vég ∪ Tő	57,14	7,60	13,42	65,71	7,74	13,84	23,19	12,90	16,58	20,18	6,32	9,62
Ige ∩ Tő	40,95	42,57	41,75	43,45	38,87	41,03	15,01	46,09	22,65	16,67	35,19	22,62
Ige ∪ Tő	68,57	8,93	15,81	72,74	8,25	14,82	49,87	20,52	29,07	58,77	14,44	23,18
Vég ∩ Ige ∩ Tő	8,57	52,94	14,75	7,62	47,41	13,13	13,67	46,36	21,12	15,79	39,13	22,50
Vég ∪ Ige ∪ Tő	70,48	8,70	15,48	74,29	8,05	14,53	50,54	17,77	26,30	59,65	11,97	19,94

is az eredményeken, viszont kiegészítve a leggyakoribb ige jellemzővel már mindenütt javít a rendszer teljesítményén. A szótő jellemző pedig a Szeged Korpusz kivételével mindenhol javulást eredményezett: feltehetőleg arányaiban kevesebb a tipikus (igéből képzett) főnévi komponenst tartalmazó félig kompozicionális szerkezet ebben a korpuszban, mint a többiben.

Míg a leggyakoribb ige az igei komponensre, a szótő és végződés pedig a főnévi komponensre tesz megszorításokat. Így a módszerek uniója a fedésre van jó hatással, hiszen a nem tipikus főnév + tipikus ige és a tipikus főnév + nem tipikus ige párokat egyaránt meg lehet találni. A módszerek metszete pedig a pontosságot javítja, hiszen így csak a tipikus főnév + tipikus ige párokat találjuk meg.

3. táblázat. Szabályalapú megközelítések eredményei a Szeged Treebanken, fedés/pontosság/F-mérték.

Megközelítés	pred. POS			etalon POS			pred. POS + szint.			etalon POS + szint.		
POS	74,56	5,75	10,69	84,21	6,70	12,41	76,32	6,92	12,69	85,09	7,77	14,23
Vég	19,30	6,53	9,76	21,93	7,35	11,01	19,30	7,64	10,95	21,93	8,56	12,32
Ige	58,77	24,28	34,36	69,30	28,11	40,00	60,53	26,44	36,80	70,18	29,20	41,24
Tő	16,67	7,85	10,67	20,18	9,35	12,78	16,67	9,00	11,69	20,18	10,80	14,07
Vég ∩ Ige	18,42	32,81	23,60	20,18	35,38	25,70	18,42	35,00	24,14	20,18	35,94	25,84
Vég ∪ Ige	59,65	12,39	20,51	71,05	14,57	24,18	61,40	14,31	23,22	71,93	16,33	26,62
Vég ∩ Tő	15,79	8,37	10,94	18,42	9,55	12,57	15,79	9,68	12,00	18,42	11,11	13,86
Vég ∪ Tő	20,18	6,32	9,62	23,68	7,38	11,25	20,18	7,35	10,77	23,68	8,54	12,56
Ige ∩ Tő	16,67	35,19	22,62	19,30	38,60	25,73	16,67	38,00	23,17	19,30	40,00	26,04
Ige ∪ Tő	58,77	14,44	23,18	70,18	17,02	27,40	60,53	16,35	25,75	71,05	18,75	29,67
Vég ∩ Ige ∩ Tő	15,79	39,13	22,50	17,54	41,67	24,69	15,79	41,86	22,93	17,54	42,55	24,84
Vég ∪ Ige ∪ Tő	59,65	11,97	19,94	71,05	14,14	23,58	61,40	13,81	22,54	71,93	15,83	25,95

A Szeged Korpusz etalon szófaji annotációja lehetővé tette azt is, hogy összevethessük a `magyarlanc` által elemzett és az etalon szófaji kódokat tartalmazó szövegeken a szabályalapú módszerek teljesítményét. Az eredményeket a 3. táblázat első két oszlopa mutatja. Egyértelműen kiderül, hogy jobb eredményeket lehet elérni, ha az etalon kézi címkéket használjuk, hiszen így a szófaji egyértelműsítés hibái kiküszöbölődnek. Különösen látványos javulás érhető el a leg-

gyakoribb ige jellemző esetében, ami valószínűleg arra vezethető vissza, hogy a `magyarlanc` gyakran minősíti hibásan melléknévnek a múlt idejű igéket (amelyek homonímek az ige befejezett melléknévi igenévi alakjával), például *adott*. Az etalon címkék használata átlagosan 2,75% javulást eredményezett az F-mértékben.

4. táblázat. Szabályalapú megközelítések eredményei szintaktikai információval (fedés/pontosság/F-mérték).

Megközelítés	**Wiki50**			**ParalellFX**			**angol Szeged Treebank**		
POS	73,33	8,85	15,79	72,98	6,89	12,59	76,32	6,92	12,69
Vég	15,24	11,03	12,80	14,52	12,82	13,62	19,30	7,64	10,95
Ige	53,33	42,11	47,06	51,19	34,82	41,45	60,53	26,44	36,80
Tő	51,43	10,87	17,94	56,19	10,16	17,21	16,67	9,00	11,69
Vég ∩ Ige	7,62	38,10	12,70	9,76	55,03	16,58	18,42	35,00	24,14
Vég ∪ Ige	60,95	24,90	35,36	55,95	23,06	32,66	61,40	14,31	23,22
Vég ∩ Tő	13,33	12,73	13,02	10,60	14,02	12,07	15,79	9,68	12,00
Vég ∪ Tő	53,33	10,53	17,58	60,12	10,18	17,40	20,18	7,35	10,77
Ige ∩ Tő	40,00	50,00	44,44	40,48	44,04	42,18	16,67	38,00	23,17
Ige ∪ Tő	64,76	12,45	20,89	66,90	10,99	18,88	60,53	16,35	25,75
Vég ∩ Ige ∩ Tő	7,62	50,00	13,22	7,26	53,98	12,80	15,79	41,86	22,93
Vég ∪ Ige ∪ Tő	66,67	12,15	20,56	68,33	10,64	18,42	61,40	13,81	22,54

Mivel számos korábbi munka szintaktikai információból kiindulva kísérelte meg a félig kompozicionális szerkezetek automatikus felismerését, mi is fokozott figyelmet fordítottunk a szintaxis szerepére. Legjobb tudomásunk szerint magyar nyelvű dependenciaelemző még nem áll rendelkezésre, így magyar nyelvi méréseinkhez a Szeged Korpusz etalon dependenciaannotációját használtuk fel.

Amennyiben pusztán szintaktikai információt használunk fel a félig kompozicionális szerkezetek azonosítására, azaz a korpuszban előforduló ige-tárgy párokat minősítünk annak, csupán 17,69-es F-mértéket érünk el a Wiki50 korpuszon (fedés: 59,51 és pontosság: 10,39). Mivel módszereink arra épülnek, hogy a baseline módszer által meghatározott lehetséges FX-ek köréből további megszorítások segítségével válasszuk ki a tényleges FX-eket, így olyan baseline-t érdemes választani, amely nagy fedéshez vezet. E célnak pedig a POS-szabályok sokkal inkább megfelelnek (76,63-as fedés a Wiki50 korpuszon), így a továbbiakban a szintaktikai információk hozzáadott értéket vizsgáljuk meg az egyes korpuszokon.

A 3. és 4. táblázat összevetéséből látszik, hogy a szintaktikai információ javít a rendszer teljesítményén, különösen a leggyakoribb ige (és kombinációi) esetében. Az átlagos javulás F-mértékben 2,3% a Wiki50, 2,26% a SzegedParalellFX és 1,52% a Szeged Korpusz esetében. A 4. táblázat utolsó oszlopa azt is mutatja, hogy a Szeged Korpuszon akkor érjük el a legjobb eredményeket, ha etalon szófaji kódokat és szintaktikai információt használunk az FX-ek azonosításában, átlagosan 4%-kal javítva az F-mértéket a predikált szófaji kódokra épülő rendszerhez képest.

3.3. Gépi tanulási módszerek

Szótárillesztéses megközelítéseket használtunk baseline megoldásnak a gépi tanulási módszerek esetében. Mivel mindkét nyelven rendelkezésünkre állt két annotált korpusz, ezért az ezeken előforduló FX-ekből lemmatizált listákat hoztunk létre. Az azonos nyelvű korpuszokra a másikról gyűjtött listát jelöltük rá. Így például a Wiki50 esetében az angol SzegedParallelFX-ről gyűjtött lista került illesztésre. A különböző korpuszokon így elért eredmények a 5. táblázatban láthatók.

5. táblázat. A szótáralapú megközelítés eredményei.

Korpusz	**Fedés**	**Pontosság**	**F-mérték**	**Szótárméret**
Wiki50	8,57	81,81	15,51	587
SzegedParalellFX angol	9,01	73,07	16,04	287
SzegedParalellFX magyar	29,5	40,14	34,01	1215
Szeged Treebank	30,7	39,77	34,65	578

Az eddig ismertetett megközelítéseken túl implementáltuk az FXtagger nevű, gépi tanuló alapú megközelítésünket is. Vizsgálatainkban a Conditional Random Fields (CRF) [17] szekvenciális tanuló MALLET [18] implementációját használtuk, az alábbi alapjellemzőkkel ([19] alapján a feladat sajátságaira szabva):

- **Felszíni jellemzők:** kis/nagybetűs kezdet, szóhossz, a szó belsejében előforduló különleges karakterek (számok, nagybetűk stb.), karakter bi- és trigramok, toldalékok;
- **Szótárak:** személynevek, cégnevek, helynevek, a leggyakoribb funkcióigék, főnevek szótövei;
- **Gyakorisági jellemzők:** a token gyakorisága, a kis- és nagybetűs alakok előfordulásának aránya, a nagybetűs és mondatkezdő alakok előfordulásának aránya;
- **Nyelvi jellemzők:** szófaj, függőségi viszonyok;
- **Környezeti jellemzők:** mondatbeli pozíció, a szó környezetében előforduló leggyakoribb szavak, idézőjelek a szó körül stb.

Ezt az általános jellemzőteret egészítettük a szabályalapú megközelítések jellemzőkre transzformált verzióival. így a leggyakoribb ige és a szótő módszereket szótáralapú jellemzőként, a POS-szabályokat és a mondat szavai közti szintaktikai kapcsolatokat nyelvi jellemzőként, míg a végződés megközelítést felszíni jellemzőként alkalmaztuk a CRF tanítása során. Mivel a magyar nyelv részletesebb morfológiai elemzést tesz lehetővé, ezért magyar nyelvű gépi tanulás során a jellemzőket még kiegészítettük ezekkel a részletesebb jellemzőkkel. Továbbá minden esetben szótáralapú jellemzőként használtuk a szótárillesztés baseline megközelítésnél használt listákat.

Kísérleteinkhez a korpuszokat 70%:30% arányban osztottuk fel tanító és kiértékelő adatbázisra. Mivel a korpuszok több témában is tartalmaznak szövegeket (újságcikkek, szépirodalom, tankönyvi mondatok stb.), minden egyes dokumentumot a fenti arányoknak megfelelően osztottunk fel a tanító és a kiértékelő adatbázis között. Eredményeink a 6. táblázatban láthatók.

6. táblázat. A gépi tanult megközelítés eredményei a különböző korpuszokon.

Korpusz	**Fedés**	**Pontosság**	**F-mérték**
Wiki50	42,86	56,96	48,91
SzegedParalellFX angol	37,91	55,55	45,07
SzegedParalellFX magyar	61,0	67,78	64,21
Szeged Treebank etalon	44,73	62,96	52,03
Szeged Treebank predikált	43,86	56,82	49,51

3.4. A gépi tanulási módszerek eredményei

A szótáralapú megközelítések eredményeiben igen nagy kontraszt mutatkozott a két vizsgált nyelvben. Ez a módszer magyar nyelvű korpuszokon kétszer jobb F-mértéket ért el, mint az angol nyelvűeken. Ugyanakkor az angol nyelvű korpuszokon a megközelítés pontossága jóval magasabb volt, mint a magyarokén. A fedésben mutatkozó különbségeket az magyarázhatja, hogy a magyar nyelvű korpuszok jóval homogénebbek voltak az angolokénál. Az enciklopédia domén (Wiki50), mely több különböző témát ölel fel, egészen más jellegű, mint a homogénebb SzegedParallelFX, nagyrészt újságcikből és regényekből álló domén, mely hatással lehet az FX-ek eloszlására is. Mivel a két magyar nyelvű korpusz mindegyikében találhatók újságcikkek, ezért a belőlük kinyert FX-listák kevésbé voltak eltérőek. A SzegedParalellFX korpuszon mért eredmények közti különbségeket magyarázhatja az alkalmazott listák mérete. Mivel a Szeged Treebank jóval nagyobb, mint a Wiki50, ezért az ezekből a korpuszokból összeállított listák mérete is nagyon eltérő. Ugyanakkor ezen baseline megközelítés pontossági értékei szerint a félig kompozicionális szerkezetek kevésbé többértelműek angolban, mint a magyar nyelvben, azaz a listákban előforduló FX-jelölt nagyobb valószínűséggel lesz a valóságban is FX.

Az 5. táblázat pontossági értékei is igazolják, hogy a félig kompozicionális szerkezetek automatikus azonosítása során hasznos információ lehet a kontextus is. Így például a *titokban tartja a kapcsolatot Imrével* szövegrészletben a *titokban tarja* és a *tartja a kapcsolatot* is lehetséges FX. Ebben az esetben a szövegkontextus segíthet eldönteni, hogy melyik szekvencia az adott szövegben az FX. A folyó szövegekben előforduló félig kompozicionális szerkezetek automatikus azonosítása így nagyban segítheti az olyan alkalmazásokat, mint a gépi fordítás vagy az információkinyerés. Ugyanakkor előfordulhat olyan eset is, amikor a felhasználót alapvetően a szövegből kigyűjthető FX-ek listája érdekli alapvetően. Ebben az esetben elegendő minden potenciális FX azonosítása a szövegben, nem

szükséges annak eldöntése, hogy az adott szekvencia FX-ként viselkedett-e az adott kontexusban.

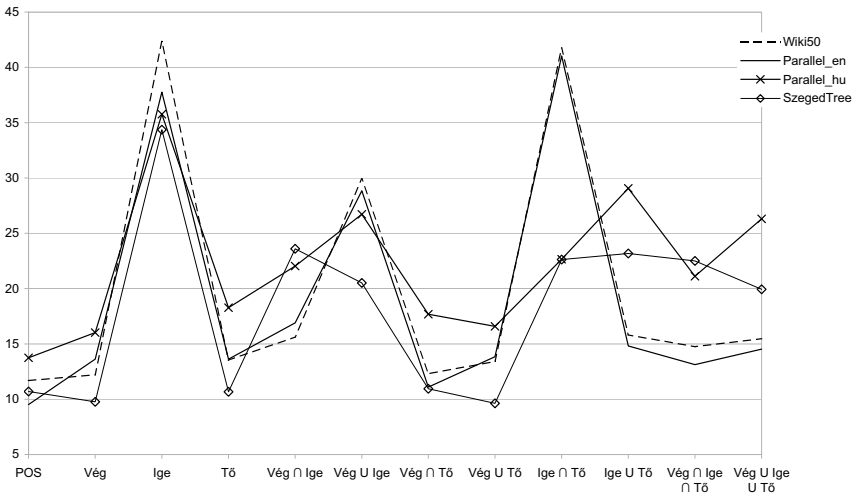
Az FXtaggerrel elért eredmények az 6. táblázatban láthatóak. A gépi tanuló megközelítéssel elért eredmények minden korpuszon meghaladták mind a szótáralapú baseline módszer, mind a szabályalapú rendszerek eredményeit. Vagyis a félig kompozicionális szerkezetek automatikus azonosítására hatékony reprezentációt voltunk képesek adni a CRF lineáris tanuló számára kibővített jellemzőtér segítségével. Mint ahogy megfigyelhettük, a korpuszokról gyűjtött szótárak kedvező hatással voltak a pontosságra, míg a POS-szabályok a fedést javították. A gépi tanuló módszerek ezen jellemzők kedvező kombinálásával érhették el a legjobb eredményeket a különböző korpuszokon.

Szembetűnő, hogy az angol nyelvű korpuszokon elért eredmények szerényebbek a magyar nyelven elérteknél. Ezt magyarázhatja, hogy megközelítéseink alapvetően a morfológiai jellemzőkre támaszkodnak, így hatékonyabbnak bizonyultak a morfológiailag jóval gazdagabb magyar nyelv esetében. Az etalon POS-címkék pozitív hatását jól mutatja a Szeged Treebanken mért két eredményünk. A SzegedParalellFX korpusz magyar nyelvű változatán elért legmagasabb F-mértéket többek közt az ebben az esetben alkalmazott nagyobb FX-lista magyarázhatja.

4. Eredmények

Az általunk definiált szabályalapú megközelítések eredményei azt igazolják, hogy már sekély morfológiai elemzések segítségével is versenyképes eredményeket lehet elérni félig kompozicionális szerkezetek automatikus azonosítása során. Hatékony jellemzőnek bizonyult a lemmatizálás, szótövesítés, szófaji egyértelműsítésen kívül egy funkcióige-lista is. Ugyanakkor a szintaktikai információk integrálása tovább javítja a rendszer teljesítményét. A félig kompozicionális szerkezetek felismerése ennélfogva leghatékonyabban a szintaktikai elemzést követően, egy utófeldolgozó lépésben valósulhat meg, annak végeredményét pedig jól tudják hasznosítani a magasabb rendű alkalmazások, például az információkinyerés és a gépi fordítás.

A különböző szabályalapú módszerek jellemzőkre való transzformálásával megvizsgáltuk a gépi tanuló algoritmusok hatékonyságát is. Általánosan elmondható, hogy a gépi tanuló módszerekkel magasabb F-mértéket tudtunk elérni, mint a szabályalapú megközelítésekkel. Ugyanakkor az eredményekből kitűnik, hogy a szabályalapú módszerek jobb fedést tudnak elérni, míg a gépi tanuló megközelítés jórészt jó pontosságának köszönheti sikerét. Ahogy a 6. táblázatban is látszik, a gépi tanuló megközelítés mind a négy vizsgált korpuszon 50% fölötti pontosságot volt képes elérni, míg a szabályalapú megközelítések vagy egyáltalán nem képesek ilyen magas pontosságra, vagy csak igen alacsony fedés mellett.



1. ábra. Szabályalapú eredmények a korpuszokon.

5. Az angol és magyar eredmények összevetése

Az angol és magyar korpuszokon elért eredményeket az 1. ábra szemlélteti. Bizonyos módszerek esetében alapvető különbségeket figyelhetünk meg a nyelvek között. érdekes módon a leggyakoribb ige és a szótő metszete sokkal jobb eredményt ért el az angol korpuszokon, mint a magyarokon, ugyanakkor e két módszer uniója a magyar korpuszokon teljesít sokkal jobban. Ennek az lehet az oka, hogy feltehetőleg az angol korpuszokban több olyan FX fordul elő, amelyek tipikus ige és tipikus főnév kombinációja, míg a magyarokban a tipikus ige + nem tipikus főnév párok vannak túlsúlyban.

További számottevő eltérést figyelhetünk meg mindhárom módszer metszete kapcsán: sokkal jobb eredményhez vezet a magyarban, mint az angolban. Ez talán azzal magyarázható, hogy a metszet megköveteli, hogy egy igei tövű főnév adott képzőben végződjön. A magyarban ez definíció szerint megvalósul (igéből képzők segítségével tudunk főnevet képezni: *dönt - döntés*), ugyanakkor az angolban a konverzió művelete is létrehozhat igéből főnevet (például *walk - walk*). Utóbbi megfelel a szótő definíciójának, de a végződésének már nem, így az ilyen típusú főneveket tartalmazó FX-eket nem lehetséges azonosítani a módszerek metszetével.

A nyelvek közti eltérések egy újabb vetületét jelenti a leggyakoribb igék száma. Míg az angolban a 12 leggyakoribb igével lehetett 40% körüli eredményeket elérni, addig a magyarban nagyobb (17 elemű) igelistával is szerényebb eredményekhez jutottunk. E jelenség magyarázatát keresve összevetettük a SzegedParalellFX két részében található FX-igék számát. Míg angolban összesen 100 ige fordult elő, melyek eloszlása megfelel a Zipf-törvénynek, addig a magyarban 179 ige fordult elő, kiegyenlítettebb eloszlásban. Tehát az angolban kevesebb

ige is nagyobb hányadát fedi le az FX-eknek, mint a magyarban. Mindez azt is mutatja, hogy az FX-igelisták bővítésével várhatóan jobb eredményeket lehet elérni mindkét nyelven.

6. Összegzés

Ebben a cikkben bemutattuk szabályalapú és gépi tanult módszereken alapuló megközelítéseinket, melyek mind angol, mind magyar nyelven képesek a félig kompozicionális szerkezetek automatikus azonosítására sekély morfológiai jellemzők segítségével. Eredményeink összevethetők más, szintaxison alapuló megközelítésekkel. Módszereinket két különböző nyelven és három korpuszon teszteltük, melyeken hasonló eredményeket értünk el. Eredményeink azt mutatják, hogy mind angol, mind magyar vonatkozásban egy adott nyelvre és doménre szabott funkcióige-lista és a főnév szótöve bizonyul a leghasznosabb jellemzőnek, illetve az angol anyagban a szintaktikai jellemzők beépítése is számottevően javít a rendszer teljesítményén. Gépi tanult megközelítésnek lineáris CRF tanuló algoritmust alkalmaztunk, melynek alap jellemzőterét kiegészítettük a szabályalapú módszerek jellemzőkre transzformált verzióival. FXtagger nevű, gépi tanuló megközelítésünk érte el a legmagasabb F-mértékeket az összes vizsgált korpuszon.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze, V., Nagy T., I., Berend, G.: Multiword expressions and named entities in the Wiki50 corpus. In: Proceedings of RANLP 2011, Hissar, Bulgaria (2011)
2. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, Association for Computational Linguistics (2007) 25–32
3. 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, Association for Computational Linguistics (2007) 41–48
4. 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, Association for Computational Linguistics (2007) 1–8

5. 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, Association for Computational Linguistics (2010) 52–60
6. Gurrutxaga, A., Alegria, I.n.: 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, Association for Computational Linguistics (2011) 2–7
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, Association for Computational Linguistics (2011) 31–39
8. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 102–110
9. Tan, Y.F., Kan, M.Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, Trento, Italy, Association for Computational Linguistics (2006) 49–56
10. Vincze, V., Felvégi, Z., R. Tóth, K.: Félig kompozicionális szerkezetek a Szeged-Paralell angol–magyar párhuzamos korpuszban. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, University of Szeged (2010) 91–101
11. Vincze, V.: Félig kompozicionális szerkezetek a Szeged Korpuszban. In Tanács, A., Szauter, D., Vincze, V., eds.: VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2009) 390–393
12. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of EMNLP 2000, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 63–70
13. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Annual Meeting of the ACL. Volume 41. (2003) 423–430
14. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, University of Szeged (2010) 275–283
15. Porter, M.F.: An algorithm for suffix stripping. In Sparck Jones, K., Willett, P., eds.: Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 313–316
16. Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: hunmorph: Open Source Word Analysis. In: Proceedings of the ACL Workshop on Software, Ann Arbor, Michigan, Association for Computational Linguistics (2005) 77–85
17. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
18. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
19. Szarvas, G., 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

II. Korpusz, ontológia

Jelentés-egyértelműsített szabadalmi korpusz

Nagy Ágoston, Almási Attila, Vincze Veronika

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
vizipal@gmail.com, {vinczev,nagyagoston}@inf.u-szeged.hu

Kivonat: A tanulmány célja, hogy megállapítsuk, hogy az angol nyelvű szabadalmakban milyen arányban fordulnak elő többjelentésű szavak, valamint azt, hogy ezek a valóságban hány különböző jelentéssel fordulnak elő e szövegekben. Kutatásaink során az A23K osztályba tartozó 60 szabadalmat tartalmazó korpuszunkban található szövegekre összpontosítunk. Előfeltételezéseink szerint a szakkifejezések és terminusok nagy része a főnév osztályba sorolható, ezek pedig adott doménen belül általában egyféleképpen használatosak. Az eredmények is azt igazolják, hogy a szabadalmakban kevesebb jelentés jelenik meg a gyakorlatban, mint amennyi a szótárakban található.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg. A keresőrendszer hatékony működéséhez a szabadalmak morfológiai és szintaktikai elemzésén túl szükséges azok szemantikai feldolgozása is, melynek előfeltétele a szavak jelentésének előzetes meghatározása, azaz a jelentés-egyértelműsítés.

A tanulmány célja, hogy megállapítsuk, hogy az angol nyelvű szabadalmakban milyen arányban fordulnak elő többjelentésű szavak, valamint azt, hogy ezek a valóságban hány különböző jelentéssel fordulnak elő e szövegekben.

Cabré [1] alapján az az előfeltételezésünk, hogy a főnevek és igék a szabadalmakban általában csak egy jelentésben fordulnak elő, mivel ezek főleg terminusok, amelyeknek alapfeltétele, hogy lehetőleg csak egy fogalmat denotáljanak. Ettől függetlenül előfordulhat, hogy egy terminus több fogalmat jelöl, de egy doménen belül csak egyet, így ideális esetben a terminusok nem lehetnek poliszémek, csak homonímek.

2 A jelentés-egyértelműsítési feladat

A jelentés-egyértelműsítés egy szöveg adott szavának egy olyan meghatározással vagy jelentéssel történő párosítását jelenti, amely az adott szóhoz társítható más lehetséges jelentésektől élesen elkülönül. Így a feladat szükségszerűen két lépésből tevődik össze: (1) a vizsgált szöveg minden releváns szavának meg kell határozni a lehet-

séges jelentéseit, illetve (2) az adott szó minden egyes előfordulásához társítani kell a megfelelő jelentést. Az első lépésben leginkább előre megadott jelentésmeghatározásokat alkalmaznak, amelyek például a következőkből állhatnak:

- hétköznapi szótárakban megadott jelentések
- különféle szemantikai jegyek, kategóriák vagy kapcsolódó szavak (pl. szinonimák)
- kétnyelvű szótárakban megadott információk (idegen nyelvű megfelelők)

A második lépésben a szóalakok és jelentések összekapcsolása két fő információforrás alapján történhet meg:

- tág értelemben vett kontextus: különféle információt tartalmaz a szó szövegkörnyezetében, a diskurzusban stb.
- külső tudásforrások: lexikális, enciklopédikus tudás

A jelentés-egyértelműsítő eljárások hatókörük alapján és a jelentésmegkülönböztetés foka szerint két-két főbb csoportra oszthatók. Hatókör tekintetében a teljes szókincsre alkalmazható (all-words WSD) és előre megadott szóalakokon működő (lexical sample WSD) módszereket különböztethetünk meg, míg a jelentésmegkülönböztetés részletessége szerint aprólékos vagy finom (fine grained), illetve durva (coarse grained) szinteket különböztethetünk meg.

A *lexical sample* alapú módszer sokkal kevesebb előzetes munkát (pl. jelentésmeghatározások megalkotása) és időráfordítást igényel, mivel nem szükséges az adott korpusz összes többjelentésű elemének előzetes definiálása. Ezzel szemben az *all-words* módszer egy jóval nagyobb mértékű vállalkozás, amely akkor lehet hasznos, ha egy általános korpuszt kívánunk létrehozni, mert ebben az esetben jobban meg lehet figyelni, hogy milyen jelentés milyen szövegkörnyezetben fordul elő.

Durva jelentésmegkülönböztetés esetén nagyobb jelentésmezők, jelentésklaszterek jelennek meg. Ezek feldolgozhatósága egyszerűbb, és az egyértelműsítés a gépi tanuló számára – és egyben az emberi annotátor számára is – könnyebb. Finom jelentésmegkülönböztetés esetén viszont sokkal aprólékosabb különbségeket lehet kódolni, ami mindenképpen hasznos lehet bizonyos alkalmazásokban, mert specifikusabb dolgokra lehet rákeresni, de a korpusz elkészítése sokkal idő- és munkaerőigényesebb feladat. A túlzott jelentésmegkülönböztetés bizonyos esetekben még az emberi annotátorok számára is indokolatlannak tűnik, gyakoriak az eltérő annotációk, hiszen minél több a jelentés, annál nagyobb a tévesztés valószínűsége. Így, mind informatikai, mind pedig nyelvészeti szempontból 3-5 egymástól pontosan elkülöníthető jelentés felvétele tűnik a legmegfelelőbbnek, mert ezt mind az emberi annotátorok, mind pedig a különféle számítógépes algoritmusok számára is ideális működési hatékonyságot tesz lehetővé (lásd [6]).

3 Korpusz és módszer

Kutatásaink során az A23K osztályba tartozó 60 gyógyszerészeti és gyógyászati segédeszközöket leíró szabadalmakat tartalmazó korpuszunkban található szövegekre [7] összpontosítunk. Annak eldöntésére, hogy mely szónak hány jelentése van, a legújabb, 3.0-s Princeton WordNetet (PWN) használtuk [8]. Ebből adódóan az egyértelműsítést csak azokra a szavakra tudjuk elvégezni, amelyek ebben az ontológiában is szerepelnek, azaz főnevekre, igékre és melléknevekre. Noha a WordNet határozószavakat is tartalmaz, ezekkel nem foglalkoztunk, mert a határozószavak előfordulási aránya igen csekély a szövegekben, továbbá a szemantikus keresés szempontjából kis jelentőséggel bírnak. Mivel a PWN finom jelentésmegkülönböztetést alkalmaz, így a lehetséges jelentések száma szóalakonként magasnak mondható.

A többértelmű kifejezések kigyűjtését 60 szabadalmi főigénypontón végeztük el. Ezeket a főigénypontokat az Apache UIMA keretrendszerében az OpenNLP modullal mondatokra bontottuk és tokenizáltuk. Ezt követően a Stanford POS-tagger segítségével minden tokenhez hozzárendeltük annak szótövét és Penn Treebank szerinti szófaji kódját (pl. NNS többes számú főnév) [5]. Eztán kigyűjtöttük a korpuszban előforduló összes főnevet, igét és melléknevet, majd megnéztük, hogy a WordNetben ezen szavak többértelműek-e vagy sem. Ehhez a Javába is beilleszthető JAWS (Java API for WordNet Searching) alkalmazást [3] használtuk. Ezután a többértelmű szavakat a szövegkörnyezetükkel együtt elmentettük a SemEval és SensEval workshopokon [2] is használatos XML formátumba.

A korpusz annotálását két független nyelvész végezte a Sensetagger program segítségével. Azokat a szavakat egyértelműsítettük, amelyek legalább háromszor előfordultak a korpuszban, a későbbiekben azonban – hasonló elvek alapján – bővíthető az annotáció. 15 szó előfordulásait mindkét annotátor bejelölte, ezáltal lehetővé vált a korpusz konzisztenciaszintjének mérése. A szavakat szófajuk szerint annotáltuk, tehát például a *form* szó igei és főnévi jelentéseit egymástól teljesen elkülönítve kezeltük, a szófaji egyértelműsítő modul elemzésének megfelelően.

4 Eredmények

Ebben a fejezetben az elkészült korpusz statisztikáit és az elért eredményeket ismertetjük.

4.1 A jelentések eloszlása

A korpuszban található többértelmű főnevek, melléknevek és igék eloszlása az 1. táblázatban látható. Hangsúlyozzuk, hogy itt a többértelműséget pusztán a wordnetbeli jelentések alapján határoztuk meg, nem pedig a valós korpuszbeli eloszlások alapján.

1. táblázat: A WordNet alapján a szabadalmakban előforduló többértelmű szavak aránya szófajonként.

	Összes	Többértelmű	
Főnév	744	284	38,17%
Melléknév	310	115	37,1%
Ige	162	135	83,33%
Összes	1216	534	43,91%

A táblázat jól mutatja, hogy elméleti szinten leginkább a szabadalmak igéire jellemző a többértelműség.

Ezen listából azon szavakat annotáltuk kézzel, amelyek legalább háromszor fordultak elő a vizsgált korpuszban. Ezek konkrét száma szófaji lebontásban és az összesre kivetítve a 2. táblázat első oszlopában olvasható. A második oszlop mutatja az annotált szavak arányát az összes előforduló többértelmű szóhoz viszonyítva. A harmadik oszlop tartalmazza azon elemek számát, amelyek az annotáltak közül legalább két jelentéssel bírnak a szabadalmakban, végül az utolsó mutatja, hogy a korpuszban többértelmű szavak aránya mekkora az annotált szavak számához képest.

2. táblázat: Az annotált szavak aránya az összes többértelmű szó függvényében.

	Annotáltak száma	Annotáltak aránya az összes előforduló többértelmű szóhoz képest	Annotált és legalább kétértelmű szavak száma	Legalább kétértelmű szavak aránya az annotáltak közül
Főnév	164	57,74%	15	9,14%
Melléknév	52	45,22%	2	3,84%
Ige	69	51,11%	12	17,39%
Összes	285	53,37%	29	10,17%

A táblázatból jól látható, hogy az annotálás során a lehetséges többértelmű szavak kicsivel több mint a felét annotáltuk kézzel. A harmadik és a negyedik oszlopból kiderül, hogy az igék azok, amelyek a legnagyobb arányban bírnak több jelentéssel a szabadalmakban: ezen igék aránya 17,4%, míg a főneveknél ez az arány 9%, a mellékneveknél pedig 4%.

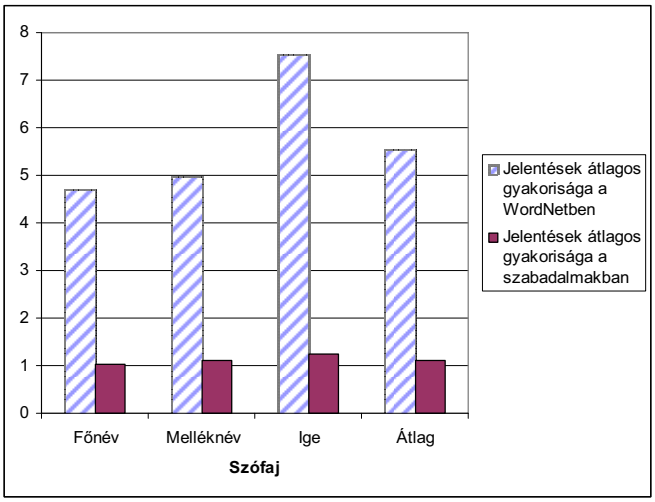
A vizsgált többértelmű szavak esetén megnéztük, hogy azok átlagosan hány jelentéssel fordultak elő mind a WordNetben, mind a szabadalmakban. A 3. táblázatban foglaljuk össze az átlagos jelentésszámot a különböző szófaji kategóriákra vonatkoztatva.

3. táblázat: Jelentések átlagos száma a WordNetben és a szabadalmakban.

	Jelentések átlagos gyakorisága a WordNetben	Jelentések átlagos gyakorisága a szabadalmakban
Főnév	4,7115	1,0385
Melléknév	4,9817	1,0976
Ige	7,5362	1,2319
Átlag	5,5509	1,1193

A 3. táblázatból jól látható, hogy a ténylegesen vizsgált és kézzel is annotált szavak esetében is az igék rendelkeznek a legtöbb jelentéssel a WordNetben, átlagban 7,5-del, míg a főnevek és a melléknevek jelentésének átlagos száma 5. A szabadalmak esetén azonban azt vehetjük észre, hogy a jelentések átlagos száma szófaji kategóriától függetlenül 1 körül van, és ez a szám az igéknél a legnagyobb, egészen pontosan 1,2319. Ez megerősíti azt a feltételezésünket, hogy a szabadalmakban nagyrészt terminusként fordulnak elő a kifejezések.

Az 1. ábra mutatja szófaji kategóriákra lebontva, hogy az adott szófaj esetén mennyi az átlagos jelentésszám a WordNetben (bal oszlop), illetve a szabadalmakban (jobb oldali oszlop).



1. ábra. Jelentések átlagos száma a WordNetben és a szabadalmakban.

Az igék között 4 darab háromértelmű (*form, reduce, make, have*) és 8 darab kétértelmű szó található. A *form* ige esetében az alábbi három jelentés figyelhető meg a WordNetben előforduló 7 jelentés közül a szabadalmakban:

4. táblázat: A *form* ige jelentései.

Jelentés száma	WordNetbeli jelentés	Példa a szabadalmakban
1	to compose or represent	
2	create (as an entity)	[…] adding to a second fluid bed dryer the fourth feed stream to **form** the granular detergent composition; […]
3	give shape or form to	[…] deforming the films to **form** a multiplicity of recesses […]
4	develop into a distinctive entity	
5	establish or impress firmly in the mind	
6	make something, usually for a specific function	A water resistant suntan gel capable of **forming** […] a water-resistant film on skin […]
7	assume a form or shape	

A wordnetbeli jelentések közül így kevesebb, mint fele használatos a szabadalmakban. Az ötös számmal ellátott jelentés például nagyon kis valószínűséggel fordulhatna elő akármilyen szabadalomban.

A szabadalmakban két jelentéssel rendelkező igék a következők: *provide, determine, combine, contain, comprise, treat, mix* és *produce*. A többi mind egy jelentéssel rendelkezik.

A melléknevek esetében kizárólag az *oral* és *lower* szó rendelkezett kettő jelentéssel a szabadalmakban, a többi mind egyjelentésű volt. Az első szó szabadalmakban előforduló két jelentését és a wordnetbéli jelentéseket az alábbi táblázat tartalmazza:

5. táblázat: Az *oral* szó jelentései.

Jelentés száma	WordNetbeli jelentés	Példa a szabadalmakban
1	of or relating to or affecting or for use in the mouth	A composition for treating diabetes to be taken in **oral** doses
2	of or involving the mouth or mouth region or the surface on which the mouth is located	tablet capable of being chewed or disintegrated in the **oral** cavity […]
3	a stage in psychosexual development when the child's interest is concentrated in the mouth; fixation at this stage is said to result in dependence, selfishness, and aggression	
4	using speech rather than writing	

A főnevek közül egyedül a *system* szónak volt kettőnél több jelentése a szabadalmakban, összesen 3 a wordnetbeli 9 helyett. Ez a három jelentés a következő volt: (1) *instrumentality that combines interrelated interacting artifacts designed to work as a coherent entity*, (2) *a group of independent but interrelated elements comprising a unified whole* és (3) *a procedure or process for obtaining an objective*. Ezen kívül 14 darab főnévnek volt legalább két jelentése a szabadalmakban.

A szabadalmakban előforduló jelentések aránya arra mutat rá, hogy noha a jelentés-egyértelműsítési feladatot finom megkülönböztetésként fogtuk fel, hiszen a WordNet alapján határoztuk meg a jelentéseket, a valóságban elégségesnek bizonyul a durva jelentésmegkülönböztetés, azaz általában 2-3 jelentéssel rendelkeznek a többértelmű szavak a szabadalmakban. Tapasztalataink azt is igazolják, hogy a gyógyszerészeti szabadalmak jelentés-egyértelműsítése nem igényli speciális gyógyszerészeti jelentéstár létrehozását, mivel egy általános célú jelentéstár (WordNet) is alkalmasnak bizonyult a feladatra.

4.2 Egyetértési ráta

A korpusz annotálását két független nyelvész végezte a Sensetagger program segítségével. Minden szófajból az öt leggyakoribb többértelmű szó előfordulásait mindkét annotátor egyértelműsítette, így mérhetővé vált az egyetértési ráta. A 6. táblázat mutatja a szófajonkénti és az összesített adatokat a mindkét annotátor által jelölt korpuszrészen.

6. táblázat: A két annotátor közötti egyetértési ráta.

	Előfordulás	Egyetértés
Főnév	211	96,68%
Ige	179	93,85%
Melléknév	62	100%
Összesen	452	96,08%

A 6. táblázat jól mutatja, hogy az annotátorok közti egyetértés igen magasfokúnak mondható. A szintén WordNet-jelentésekre épülő magyar nyelvű WSD-korpusz [6] egyetértési rátája 84,78%-os volt, amihez képest 11,4%-kal jobb teljesítményt értünk el a minta alapján. Ez arra enged következtetni, hogy szakszövegekben könnyebb feladat a jelentés-egyértelműsítés, hiszen egy adott doménen belül kisebb valószínűséggel használatosak a szavak többféle jelentésben (noha a *család* szó többértelmű, botanikai kontextusban szinte kizárólagosan a rendszertani kategóriát jelöli). Bár a magyar WSD-korpusz is homogén szövegeket tartalmaz (HVG-cikkek), azok nyelvezete és tematikája mégsem annyira kötött, mint a szabadalmaké (vö. [4]).

Különösen a melléknevek egyértelműsítése bizonyult könnyű feladatnak, noha itt számottevően kevesebb példát kellett címkézni, mint a főnevek és igék esetében. Meg kell tovább említeni, hogy a melléknevek nagy többsége egyjelentésűként fordult elő a szabadalmakban, ami tovább könnyítette az annotálást. Az egyértelműsítésre kiválasztott mintában a *form* ige bizonyult a legnehezebbnek: itt az annotátorok pusztán 52,6%-ban értettek egyet. Ennek valószínűleg az lehet az oka, hogy két jelentést (’lét-

rehoz' és 'valamilyen célra létrehoz') egymáshoz közel állónak, így nehezen megkülönböztethetőnek ítéltek az annotátorok. Az eltérően annotált esetek nagy része e két jelentést érintette.

5 Összegzés és további célok

Tanulmányunkban bemutattuk a gyógyszerészeti szabadalmakat tartalmazó jelentés-egyértelműsített korpuszunkat. A wordnetbeli és a korpuszban előforduló jelentések aránya azt tükrözi, hogy szakszövegekben, jelesül a szabadalmakban kevesebb jelentés jelenik meg a gyakorlatban is, mint ahogy azt az adatbázis alapján várhatnánk. Ez némileg megkönnyíti mind az annotátorok, mind a gépi egyértelműsítés feladatát.

Az elkészült korpuszt a jövőben szeretnénk jelentés-egyértelműsítő algoritmusok tesztelésére használni, melyek beépülnek majd a szemantikus keresőbe.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Cabré, M. T.: Terminology. Theory, methods and applications. John Benjamins, Philadelphia PA (1998)
2. Erk, K., Strapparava, C. (eds.): Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Uppsala, Sweden, July (2010)
3. Java API for WordNet Searching (JAWS), http://lyle.smu.edu/~tspell/jaws/index.html
4. Osenga, K.: Linguistics and patent claim construction. Rutgers Law Journal Vol. 38, No. 61 (2006) 61–108
5. Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml
6. Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of 6th International Conference on Language Resources and Evaluation. LREC 2008, Marrakech, Morocco (2008) 3344–3349
7. Vincze, V., Nagy Á., Klausz, Á., Almási, A., Kiss, M., 2010: Nyelvészeti problémák a szabadalmak feldolgozásában. In: Tanács, A., Vincze, V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 168–179
8. WordNet – A lexical database for English, http://wordnet.princeton.edu/

Korpuszépítés ómagyar kódexekből

Simon Eszter, Sass Bálint, Mittelholcz Iván

MTA Nyelvtudományi Intézet
{eszter,sass.balint,mittelholcz}@nytud.hu

Kivonat Az annotált nyelvi erőforrások elérhetősége egyre fontosabb szerepet kap a nyelvészet több területén: a nyelvtechnológiai fejlesztéseken kívül az elméleti kutatásoknak is kiváló alapanyagot szolgáltatnak a korpuszok. A Magyar Generatív Történeti Szintaxis című projekt keretében felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar szövegemléket. A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk – a szkenneléstől az online lekérdező felületig.

1. Bevezetés

Az annotált nyelvi erőforrások elérhetősége egyre fontosabb szerepet kap a nyelvészet több területén: a nyelvtechnológiai fejlesztéseken kívül az elméleti kutatásoknak is kiváló alapanyagot szolgáltatnak a korpuszok. A történeti korpuszok az adatok és a nyelvi jelenségek gazdag tárházát adják – de csak akkor, ha a releváns információ elektronikusan interpretálható és előhívható módon van tárolva bennük. A Magyar Generatív Történeti Szintaxis című projekt célja, hogy diakrón szintaktikai vizsgálatokat végezzen magyar nyelvű szövegeken, melyhez elsődleges fontosságú egy elektronikus nyelvtörténeti adatbázis létrehozása. A projekt időtartama alatt (2009-2013) felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) szövegemléket, a középmagyar korból (1526–1772) pedig különféle szempontok szerinti arányos válogatást úgy, hogy minden nyelvjárás, műfaj, regiszter súlyának megfelelően képviselve legyen benne.

Napjainkban a korpuszépítési munkálatok során elsősorban már digitalizált szövegekből indulnak ki; de nem ez a helyzet a történeti dokumentumokkal. Az elektronikus formátumok (sőt az elektromosság) előtti korból származó szövegekből való korpuszépítés sokkal idő- és munkaigényesebb folyamat, és bizonyos esetekben más módszereket is igényel, mint a mai szövegek esetében. A tény, hogy az ómagyar kor több mint 6 évszázadot fog át, amelynek során nem volt egységes hangjelölési rendszer, vagyis az egyes szövegekben levő speciális karakterek halmaza különböző, tovább nehezíti a helyzetet. A helyesírás ezekben a századokban távolról sem volt egységes, ráadásul egy kódexet általában több kéz jegyez, ami még tovább növeli a heterogenitást a szövegekben. Ezek és más, később részletezett okok miatt a sztenderd előfeldolgozó lépések (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) nem végezhetők teljesen automatikusan, és nagyon sok kézi ellenőrzést igényelnek.

A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk – a szkenneléstől az online lekérdező felületig. A 2. fejezetben a korpusz anyagának összegyűjtését írjuk le, majd a 3. fejezetben bemutatjuk a korpusz felépítését, valamint az ezzel párhuzamos szövegfeldolgozási lépéseket. A 4. fejezet az online lekérdező felület leírását adja, végül a korpuszépítéssel kapcsolatos további feladatainkat tárgyaljuk.

2. A korpusz anyagának összegyűjtése

A reprezentativitás a korpuszok egyik lényegi tulajdonsága, kivéve abban az esetben, ha egy holt nyelvet vagy egy nagyon speciális nyelvi réteget vizsgálunk. Ez a helyzet az ómagyar korpusz esetében is, amely terveink szerint az összes ómagyar korból fennmaradt szövegemléket tartalmazni fogja. Szövegemlék alatt az összefüggő ómagyar mondatokat tartalmazó nyelvemlékeket értjük, az ún. szórványemlékekkel, amelyekben csak sporadikusan fordulnak elő magyar szavak vagy nevek, jelen projektben nincs lehetőségünk foglalkozni. Nem szerepelnek továbbá a korpuszban azok a szövegek sem, amelyeket még soha nem adtak ki nyomtatásban, vagyis a nyelvtörténeti átírási munkát is nekünk kellene elvégezni.

A fenti megszorításokat figyelembe véve a feldolgozandó ómagyar anyag 47 kódexet, 27 rövidebb szövegemléket és 244 misszilist (elküldött levelet) foglal magában, vagyis mindösszesen körülbelül 2 millió szövegszót. Ebből több mint 770 ezer már elérhető, kereshető állapotban van. A középmagyar kori szövegek kiválogatása még folyamatban van.

A korpuszépítés első lépése a valamilyen elektronikus szöveges formátumban már meglévő nyelvtörténeti anyagok összegyűjtése volt. A különböző forrásokból származó, változatos fontkészleteket használó, jellemzően Microsoft szövegszerkesztő eszközökkel előállított dokumentumokat egységes, UTF-8 kódolású, sztenderd Unicode-karaktereket tartalmazó sima szövegfájlokká alakítottuk. Egy másik forrásunk a Számítógépes Nyelvtörténeti Adattár volt, amelyben több ómagyar kódex ábécérendes adattára elérhető. A kódexfeldolgozási munkálatok még a hetvenes években kezdődtek a Debreceni Egyetemen Jakab László vezetésével. Az adattárban a kódex címszavai ábécérendbe rendezve szerepelnek. A hozzájuk tartozó betűhű szövegszavakat a lelőhely (lapszám, sorszám) megjelölésével közlik, mellettük számokkal rögzítették az adatra vonatkozó helyesírás-történeti, szótörténeti, hangtani, szófajtani, jelentéstani és alaktani tudnivalókat. Ez a fajta adatkódolási módszer még a hetvenes évekből maradt, mivel annak idején még lyukkártyán rögzítették az információkat. Ebből a táblázatos formából állítottuk vissza a kódexek eredeti betűhű szövegét, továbbá az egyes szövegszavakhoz tartozó morfológiai elemzést az általunk használt morfológiai elemző kimeneti formátumára átalakítva.

Az ómagyar szövegek nagy részének azonban nincsen elektronikusan elérhető szöveges változata, így ezeket a számítógép által olvasható és feldolgozható formára kell hoznunk. Ez a rövidebb szövegek esetében általában begépeléssel, a hosszabbak esetében szkenneléssel, optikai karakterfelismerő (OCR) program alkalmazásával és kézi ellenőrzéssel történik.

3. Az annotáció kidolgozása

Ahhoz, hogy a korpuszban a nyelvi jelenségek kereshetők legyenek, vagyis az adatbázis használható segédeszköze legyen az elméleti és nyelvtörténeti kutatásoknak, a releváns információknak elektronikusan interpretálható és előhívható módon kell tárolva lenniük. Ennek megvalósításához a sztenderd szövegfeldolgozó lépéseket (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) kell megtennünk, a történeti szövegek esetében azonban ezek nem problémamentesek. Bizonyos lépések automatizálhatók, de munkaigényesebb módszereket és több kézi ellenőrzést igényelnek, mint a mai nyelvet reprezentáló korpuszok esetében.

A korpusz felépítése, vagyis az egyes szövegszavakhoz tartozó annotációs szintek párhuzamosan alakulnak a szövegfeldolgozottsági szintekkel, melyeket az 1. táblázatban láthatunk. Ezek alapján hat annotációs szintet és öt feldolgozó lépést különíthetünk el, melyeket ebben a fejezetben ismertetünk részletesebben.

1. táblázat. Szövegfeldolgozottsági szintek.

(1) kiadott kódex szkennelve
→ OCR
(2) nyers OCR-kimenet
→ *kézi* javítás, kódolás
(3) betűhű elektronikus forma
→ *félautomatikus* normalizálás
(4) normalizált forma
→ *automatikus* morfológiai elemzés
(5) szótövesített és morfológiailag elemzett forma
→ *kézi* egyértelműsítés
(6) egyértelműsített korpusz

3.1. Szkennelés

Néhány kódex beszkennelt verziója megtalálható a Magyar Elektronikus Könyvtárban, sőt ezek egy része ún. „szendvics” PDF, vagyis a kép mögött megtalálható az OCR-ezett szöveg is. Ennek ellenére ezeket nem tudtuk használni: a képek felbontása nem elég jó az OCR-ezéshez, a mögöttes szöveg pedig nem esett át kézi ellenőrzésen, vagyis meglehetősen sok benne a hiba. Így minden kódexet, amit nem tudtunk szöveges formában megszerezni, minimum 300 dpi felbontásban be kellett szkennelnünk.

3.2. OCR

Az ómagyar kódexekben található nagyszámú különleges karakter kezelése miatt az OCR programmal szemben alapvető elvárásunk volt a taníthatóság. A

szóba jöhető nyílt forráskódú szoftverek (pl. Tesseract) tanítása túl időigényes lett volna, ezért végül az Abby FineReader mellett döntöttünk. Ez ugyan nem nyílt forráskódú, de meglehetősen könnyen tanítható, és elég jó minőségű kimenetet ad.

Az OCR program teljesítményét másokhoz hasonlóan (pl. [1]) nem karakterszinten, hanem szópontossággal (*word accuracy, WAcc*) mértük (az írásjelek felismerésétől eltekintettünk). Az előzetes elvárásoknak megfelelően az eredmények azt mutatják, hogy a pontosság nagyban függ a kódexekben alkalmazott helyesírástól. Kniezsa [2] az ómagyar kori kódexek kezeinek helyesírását három nagy típusba sorolja; a kiértékelésnél ezt a kategorizálást követtük. A mellékjel nélküli helyesírás a latinban nem szereplő magyar hangokat több betű kombinációjával írja le; a mellékjeles helyesírás egy rokonhang betűjének mellékjeles változatával jelöli ezeket; a harmadik típus pedig ezek keveréke. A kiértékeléshez három kódexet választottunk a három különböző típusból, továbbá összehasonlítási alapként egy rövidebb mai magyar szövegen is kiértékeltük a szoftver teljesítményét.

A legjobban a mellékjel nélküli helyesírással boldogult a program: ez nagyjából megegyezik a mai magyar szövegek felismerésében nyújtott pontossággal. A mellékjeles és keverék helyesírású kódexekben használt speciális karakterek nagy száma a tanítás ellenére is kb. 30%-kal rontotta a pontosságot.

2. táblázat. Az OCR szópontossága helyesírási típusok szerint.

kódex	helyesírás	tokenszám	felismert	WAcc (%)
Kulcsár	mellékjel nélküli	36.321	35.258	97,07
Müncheni	mellékjeles	74.657	50.790	68,03
Czech	keverék	11.478	7.910	68,91
–	mai magyar	5.121	5.068	98,97

3.3. A betűhű szöveg

A betűhű szöveg elkészítésekor nem a kódexek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követjük, vagyis nem törekszünk tökéletes paleográfiai pontosságra. A szabványosság előnyei miatt a teljes korpuszt sztenderd UTF-8 kódolású Unicode karakterekkel tároljuk és jelenítjük meg. Mindenképpen szükséges egy, az egész korpuszra kiterjedő szigorúan egységes formátum, ez teszi lehetővé, hogy a lekérdezéseket az egész anyagra vonatkoztathassuk. Ugyanakkor viszonylag nagy erőfeszítést kíván ennek az egységességnek a megvalósítása, mivel az egyes nyelvemlékek írásmódja, a bennük előforduló speciális ómagyar karakterek halmaza meglehetősen különbözik egymástól. A különféle ékezetes és többszörösen ékezetes karaktereket a Unicode megfelelően kezeli, de előfordulnak olyan régi magyar karakterek is, melyek a Unicode-ban nincsenek reprezentálva. Ezeket a karaktereket egy kiválasztott

Unicode karakterrel helyettesítjük, mégpedig úgy, hogy az adott helyettesítő karaktert kizárólag az adott hiányzó eredeti karakter helyett használjuk a korpuszban.

3.4. Normalizálás

A magyar írásosságot a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. Az ómagyar korban a helyesírás még egyáltalán nem volt egységesítve, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még annyira sem várható el. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. *Vylag uilaga* [világ világa])), vagy kettős hangértéke van egy-egy betűnek (pl. *zerzete zerent* [szerzete szerint]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is (pl. az *u,v,w* több évszázadon át jelölhette az *u,ú,ü,ű,v* hangok bármelyikét).

Ezért szükség van egy ún. *normalizálás*i lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. [3]). A szövegfeldolgozásnak ez a lépése kritikus fontosságú, enélkül ugyanis a (félig) automatikus annotáció hatékonysága a következő lépésekben drámaian visszaesik [4].

Mivel a normalizálás nyelvtörténeti szakértelmet kívánó, rendkívül időigényes manuális munka, megpróbáltuk kiváltani gépi eljárással. Az általunk épített gépi normalizáló az ómagyar tokenekhez átírási lehetőségeket rendel, melyek közül a normalizálást végző nyelvész ki tudja választani a megfelelő kimenetet (részletesen lásd [5]).

A normalizálás során két alapelvet tartunk szem előtt. Egyrészt a ma nem létező összes szót, toldalékot, morfológiai konstrukciót megtartjuk, vagyis morfémát nem toldunk be, és nem hagyunk el. Másrészt viszont elhagyunk minden fonológiai és helyesírási esetlegességet, vagyis egységes, amennyire lehet, a mainak megfelelő helyesírásra törekszünk. Ez utóbbi azt is jelenti, hogy egy adott szót mindig ugyanúgy írunk le – ezt nevezzük az egységesség elvének.

A normalizálási lépés során történik meg a szöveg tokenekre és mondatokra való bontása is – mindkettő kézzel. Tokenizáláson jelen esetben azt értjük, amikor az ómagyar szövegben a szavakat a mai helyesírásnak megfelelően összevonjuk, illetve szétválasztjuk, természetesen a megfelelő módon jelölve a változtatásokat. Mivel ebben a korban a mai írásjelek nagy része még ismeretlen volt, továbbá amit használtak, azt se következetesen tették, a mai értelemben vett automatikus

mondatra bontás teljesen lehetetlen vállalkozásnak tűnik. Ezért ezt a szövegfeldolgozási lépést is manuálisan végezzük el.

3.5. Morfológiai elemzés és egyértelműsítés

A normalizált szövegváltozat képezi a morfológiai elemző bemenetét. Mivel a normalizálás során az ómagyar szöveget mai magyarra írjuk át, az ez utóbbira kifejlesztett automatikus morfológiai elemzőt viszonylag könnyen tudjuk alkalmazni a nyelvemlékek feldolgozására. Jelen projektben a *Humor* elemzőt használtuk [6]. Az egyik normalizálási alapelvünk, hogy minden morfológiai konstrukciót megtartunk, ezért természetesen ki kellett bővítenünk a lexikont és a szabályhalmazt bizonyos ma már nem létező, de az ómagyarban még használt nyelvi jelenségek leírásával. A morfológiai elemző kimenetének egyértelműsítését viszont – a gépi normalizáló kimenetének kezeléséhez hasonlóan – kézzel végezzük.

4. Korpuszlekérdező eszköz

A korpusszal párhuzamosan készül a hozzá tartozó korpuszlekérdező rendszer, amelynek segítségével a teljes ómagyar korpuszt kutathatjuk. A jó korpuszlekérdező eszközök lehetővé teszik azt, hogy kifinomult, nyelvészetileg releváns lekérdezéseket fogalmazzunk meg általuk. Az ilyen lekérdezések sok esetben különféle nyelvi szinteken megjelenő információra hivatkoznak. Hogy ez megvalósulhasson, adatbázisunk párhuzamosan tartalmazza az 1. táblázatban látható hat szövegfeldolgozottsági szintnek megfelelő nyelvi adatokat. Ezenfelül lehetővé tesszük a több szintre való egyidejű hivatkozást akár egy kérdésen belül is. Ha például az a kérdésünk, hogy milyen szavak szerepelnek egy igealak és egy igekötő között, akkor az elemzések szintjén (6) kell megfogalmazni a kérdést. Ha gyakorisági listát készítünk a korpusz egy részéből, akkor ezt megtehetjük például a szótövekből kiindulva, de rá lehet kérdezni közvetlenül az *nç* végű szavakra is, ekkor a (3) szinthez fordulunk.

A korpusztalálatok megjelenítése független a lekérdezéstől, abban az értelemben, hogy igény szerint bármilyen – akár a lekérdezésben nem is szereplő – szövegfeldolgozottsági szintet is megjeleníthetünk.

A korpusz anyaga vertikális fájlok formájában készül el. Ezek `.csv` formátumú táblázatok, melyek soronként egy szövegszót tartalmaznak, az egyes szövegfeldolgozottsági szintekhez tartozó információt pedig a megfelelő oszlopban, kiegészítve egy ,,Értelmezés" és egy ,,Megjegyzés" oszloppal. Ezt a formát XML-lé alakítjuk, így végezzük el a validációs lépéseket, melyek az adatbázis konzisztenciáját ellenőrzik. Egy következő átalakító lépés során alakul ki az alkalmas bemenet az *Emdros* [7] korpuszkezelő rendszer számára, melyre a lekérdezőfelület épül.

A lekérdező felület az 1. ábrán látható. A felület középső részén hivatkozhatunk az egyes szövegfeldolgozottsági szintekre. Az itt megadott adatokból az *OK* gomb megnyomására áll elő maga a lekérdezés a bal oldali szövegmezőben az Emdros lekérdezőnyelvén, ez szerkeszthető, és a *Mehet* gombbal futtatható.

1. ábra. A korpuszlekérdező felülete. A feltüntetett példában azokra a tokenekre keresünk, melyeknél a normalizált alak kezdete a *jonh* sztring.

2. ábra. Az 1. ábrán látható lekérdezés eredményének részlete: korpuszpozíciók, ahol a normalizált alak kezdete a *jonh* sztring.

Az 1. ábrán bemutatott lekérdezés eredménye a 2. ábrán látható. A találatok felett a lókuszjelölő található, mely a kódex azonosítójából, az oldalszámból és az adott szó egyedi azonosítójából áll. Az egyes találatokat táblázatos formában jelenítjük meg: a betűhű alak zölddel, a normalizált alak feketével, az értelmezés – az ómagyar *jonh* mai magyar megfelelője a *szív* szó – pedig kékkel.

Végül lássunk egy valódi ómagyar szintaxisra vonatkozó elméleti nyelvészeti kutatási kérdést, melynek megválaszolásához segítséget nyújthat a korpusz. A mai magyarban tagadás esetén az igekötő követi az igét (vö: *nem jön be*), az ómagyar viszont az igekötő + tagadószó + ige (vö: *be nem jön*) sorrendet használja legtöbbször. A szófajok sorozatára vonatkozó megfelelő lekérdezések a 3. ábrán láthatók. Ezt a jelenséget mutatja a Jókai-kódexből származó alábbi példamondat is: „Ver touaba **kÿ nem futott**” (Vér továbbá ki nem futott.).

Mai magyar szórend:

```
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
[W FOCUS w_6e ~ 'Vpfx']
```

Ómagyar szórend:

```
[W FOCUS w_6e ~ 'Vpfx']
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
```

3. ábra. A tagadott ige és igekötő sorrendi viszonyára vonatkozó lekérdezések. A `w_6e` jellemzővel a (6) szinten elérhető morfológiai elemzésre kérdezhetünk rá, a tagadószó kódja `Mod`, az ige kódja `V`, az igekötőjé pedig `Vpfx`.

A *Régi Magyar Konkordancia* nevet viselő lekérdezőfelület szabadon elérhető a `http://corpus.nytud.hu/rmk` címen.

5. További feladatok

Elsődleges feladatunk a teljes ómagyar anyag betűhű szöveges formában való előállítása és kereshetővé tétele. A normalizálást, valamint a morfológiai elemzést és egyértelműsítést csak a korpusz egy részén fogjuk végrehajtani.

Az ómagyar szövegek eleve adott heterogenitása mellett további problémákat okoz az is, hogy a különböző korokban kiadott nyomtatott kódexátiratok tipográfiai kényszerűségek miatt azonos karaktereket eltérően jelenítenek meg. Terveink között szerepel ezen esetlegességek kiküszöbölése, vagyis a különbözőképpen jelölt karakterek azonos sztenderd Unicode-karakterrel való lecserélése.

A középmagyar anyagok esetében már fontos szerepet játszik a reprezentativitás kérdése, ugyanis ebből a korból lényegesen több nyelvemlékünk származik, vagyis a teljes anyag feldolgozására ebben a projektben nem vállalkozhatunk.

A középmagyar szövegemlékek kiválogatásánál két fő szempontot tartunk szem előtt: csak a már szöveges formátumban elérhető dokumentumokkal foglalkozunk, és ezeket Dömötör [8] műfaji beosztását követve kategorizáljuk úgy, hogy minden regiszter megfelelően képviselve legyen a korpuszban.

Köszönetnyilvánítás

Az ómagyar korpusz építése a Magyar Generatív Történeti Szintaxis projekt keretében valósul meg. A projektet az OTKA NK 78074. számú pályázata támogatja. Köszönetet mondunk Novák Attilának, aki a morfológiai elemzést és a Jakab László-féle táblázatok átalakítását végzi.

Hivatkozások

1. Volk, M., Marek, T., Sennrich, R.: Reducing OCR Errors by Combining Two OCR Systems. In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), Lisbon, Portugal, Faculty of Science, University of Lisbon (2010)
2. Kniezsa, I.: Helyesírásunk története a könyvnyomtatás koráig. Akadémiai Kiadó, Budapest (1952)
3. McEnery, T., Hardie, A.: Lancester Newsbooks Corpus. (2003)
4. Rayson, P., Archer, D., Baron, A., Culpeper, J., Smith, N.: Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In: Proceedings of Corpus Linguistics, University of Birmingham (2007)
5. Oravecz, C., Sass, B., Simon, E.: Semi-automatic normalization of Old Hungarian codices. In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), Lisbon, Portugal, Faculty of Science, University of Lisbon (2010)
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, College Park, Maryland, USA (1999) 261–268
7. Petersen, U.: Emdros – a text database engine for analyzed or annotated text. In: COLING 2004. (2004) 1190–1193
8. Dömötör, A.: Régi magyar nyelvemlékek. Akadémiai Kiadó, Budapest (2006)

Nem lexikalizált fogalmak a Magyar WordNetben

Vincze Veronika, Almási Attila

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
vinczev@inf.u-szeged.hu, vizipal@gmail.com

A Magyar WordNet (HuWN) építése során az annotátorok viszonylag nagy számú olyan fogalommal találkoztak, melyeknek nem volt megfelelőjük a magyar nyelvben. E dolgozatban bemutatjuk a HuWN-be bevezetett nem lexikalizált synsetek két (non-lex és t non-lex) típusát, megvizsgáljuk a non-lex jelenség hátterét, statisztikákat is közlünk, a két wordnetből vett példákkal rávilágítunk bizonyos problémákra, majd megoldásokra is javaslatot teszünk többszavas kifejezések kezelésének kérdését is körüljárjuk és egy esetleges jövőbeli HuWN revízió non-lex irányú felülvizsgálatát is javasoljuk.

1 Bevezetés

A wordnetek olyan lexikai adatbázisok, amelyek jelentésük alapján klaszterekbe rendeződnek és különféle szemantikus és lexikai relációk segítségével kapcsolódnak össze egy konceptuális hierarchiába (lexikai ontológiába). Eredetileg azért alkották meg ezeket, hogy bemutassák, hogyan szerveződnek a nyelvi ismeretek az emberi elmében [6].

A wordnetek méretüket tekintve ugyan eltéréseket mutatnak, de ezeket – különösen a Princeton WordNetet (PWN) – tekintik egy adott nyelv legnagyobb nyelvi információt tartalmazó adatbázisainak.

A wordnetek létrehozásánál a többnyelvűség is fontos szempont: az építők rendszerint a PWN-hez igazítják új adatbázisaikat, így azokat olyan – mind egy-, mind pedig többnyelvű – alkalmazásokban lehet felhasználni a számítógépes nyelvészetben mint pl. a jelentés-egyértelműsítés, a gépileg támogatott fordítás, dokumentumklaszterezés stb.

Azonban két nyelv sosem fedi egymást teljesen sem a konceptuális, sem pedig lexikai szinten. Dolgozatunkban fogalmak megfeleltetése szempontjából vetjük össze a magyar és angol wordnetet, ismertetjük a felmerült problémákat és megoldási javaslatokat is teszünk. Először röviden bemutatjuk a magyar és angol wordnetet, majd példákkal világítjuk meg a nem lexikalizált (*non-lex*) és technikailag nem lexikalizált (*t non-lex*) synseteket. Ezt követően arra teszünk javaslatot, hogy hogyan kerülhetjük el a *non-lex* címke alkalmazását, végül pedig rámutatunk arra, hogy noha ideális esetben egy, a nyelv konceptuális hierarchiáját ábrázoló wordnetnek nem kellene non-lex elemeket tartalmaznia, mégis hasznosnak bizonyulhatnak olyan kutatási területek számára, mint a pszicholingvisztika, néprajz és kontrasztív nyelvészet.

2 Wordnetek a nagyvilágban

Az első wordnetet a Princeton Egyetemen hozták létre angol nyelvre. A ’90-es évek óta folyamatosan fejlesztik és mostanra a legnagyobb angol nyelven hozzáférhető lexikai adatbázissá vált, mely könnyen illeszthető különféle számítógépes alkalmazásokhoz. A Princeton WordNet 3.0 hozzávetőleg 155 000 szót és mintegy 117 000 synsetet tartalmaz.

Azóta egyéb wordneteket is létrehoztak, így pl. a EuroWordNetet, holland, olasz, spanyol, német, francia, cseh és észt nyelvekre [2]; a BalkaNetet, az EuroWordNet kiterjesztéseként bolgár, görög, török, szerb és román nyelvekre [9,10]. Ezeken kívül wordneteket fejlesztettek még arab, horvát, kínai, dán, szlovén, lengyel, orosz, perzsa, hindi, tulu, dravida, tamil, telegu, szanszkrit, bodo, asszámi és filippínó nyelvekre [3,8].

A Magyar WordNetet (HuWN) a Magyar Tudományos Akadémia Nyelvtudományi Intézete, a Szegedi Tudományegyetem Informatikai Tanszékcsoportja és a MorphoLogic Kft. Fejlesztette ki egy hároméves projekt keretében [1,5]. A HuWN jelenleg több mint 40 000 synsetet tartalmaz, melyből 2 000 synset a gazdasági, 650 synset pedig a jogi szakontológia részét képezi.

A HuWN alapjául a Princeton WordNet 2.0 szolgált, pontosabban a BalkaNet Concept Setbe (BCS) tartozó synsetek lettek kiválogatva és magyarra fordítva. A wordnet készítői ezt követően szerkesztették, javították és kiterjesztették őket szinonimákkal a VisDic szerkesztőprogram segítségével. Később a fogalmak körét koncentrikusan terjesztették ki, azaz a már meglévő synsetek „utódait” synsetjelöltekként kezelték. A végső döntést, arról, hogy felvegyék őket vagy sem, több tényező is befolyásolta, mint pl. a fogalom gyakorisága vagy jelenléte más wordnetekben [5].

3 Nem lexikalizált synsetek

A munka kezdetén a magyar wordnet fejlesztői az úgynevezett expand[1] módszer mellett döntöttek. Ez azt vonta maga után, hogy a HuWN a PWN hierarchiáját örökölte. A HuWN főnévi és melléknévi része a következő módszer alapján lett felépítve: a PWN csomópontjait automatikusan magyar synsetjelöltekhez kapcsolták és a relációkat átvették. Az alapstratégia az volt, hogy egy kétnyelvű angol-magyar szótár magyar szócikkeit hozzákapcsolták a PWN 1.6 főnévi/melléknévi synsetjeihez.

A HuWN létrehozása gyakorlatilag azt jelentette, hogy a PWN synseteket magyarra fordították. Azonban, mivel nincs teljes átfedés a nyelvek fogalmai között, kulturális, életkörülmények és egyéb tényezők eltéréséből adódóan a nyelvek gyakran csak rájuk jellemző fogalmakkal rendelkeznek, s ezeknek más nyelvekben csak hozzávetőleges megfelelőik vannak, és nem fordíthatók, fejezhetők ki egyetlen szóval [4].

Így a PWN építési elvek teljes átvételének és alkalmazásának negatív következményei lettek volna a HuWN-re; egyrészt kevésbé tükröződött volna a magyar lexikalizáció, másrészt a PWN konceptuális szerkezetének egy az egyben magyarra

[1] Kiterjesztéses modell

történő átültetése további nehézségeket okozott volna, különösen a többnyelvű alkalmazásokra tekintettel [7].

Azért, hogy ne legyenek „lyukak" a fában, azaz a magyar és angol wordnet a lehető legnagyobb mértékben átfedjen, meg kellett találni az ilyen synsetek megfelelő kezelésének módját. Bevezettük a *non-lex* címkét olyan synsetek jelölésére, melyek (szó szintjén) nem léteznek az adott nyelv lexikonjában. Ezek a synsetek körülírás formájában tartalmazzák az angol synsetnek megfelelő fogalmat, de definíciót és példát nem.

POS: n NL: yes
ID: ENG20-04138222-n BCS: 3
Synonyms: (hajó jobb oldala):0
Domain: aeronautic

NL jelöli a *non-lex*-t; a synsetnek nincs definíciója, példája, értelmező szótárbeli linkje és literálja.

Alább statisztikákat közlünk a HuWN nem lexikalizált synsetjeit illetően. Látható, hogy a HuWN egészét tekintve minden huszadik, a BCS részt tekintve pedig minden tizenkettedik synset nem lexikalizált.

1.táblázat: (Technikai) nem lexikalizált synsetek a HuWN-ben

	HuWN	BCSHu
Synsetek	42 292	8 446
Nem lexikalizált	1 999	463
Technikai nem lexikalizált	454	271
Nem lexikalizált synsetek % -a	5,799	8,69

Most pedig megadjuk azokat a kritériumokat, amelyek alapján egy synset a non-lex synset kategóriába sorolható. Először, lehetséges, hogy a fogalom az adott nyelvben nem fordul elő (különösen kulturális különbségeknek köszönhetően). Másodszor, a fogalom kifejezhető produktív vagy kompozicionális szerkezetekkel (pl. melléknév + főnév szerkezetekkel), azaz nincs mód arra, hogy egyetlen szóval fejezzük ki őket. Harmadszor, a fogalom több más, egyetlen szóval kifejezhető fogalmat foglal magában, így a másik nyelvben csupán egy listával fejezhető ki. Negyedszer, úgy tűnik, hogy a PWN több következetlenséget vagy hibás definíciót, hipernima relációt tartalmaz, melyeket a HuWN építői nem kívántak követni és ehelyett a problémás synseteket *non-lex* címkével látták el.

3.1 A nem lexikalizált synsetek típusai

A nem lexikalizált synsetek hat fő osztályba sorolhatók, melyekre példákat alább láthatunk.

3.1.1 Kulturálisan meghatározott fogalmak

Ezek a fogalmak a kultúrák, életstílus, földrajzi elhelyezkedés stb. különbségeiből fakadnak. Mivel a magyar és amerikai kultúra, (népi) hagyományok és társadalmi háttér igen eltérő, vannak olyan fogalmak, melyeknek vannak ugyan szó szerinti megfelelőik a másik nyelvben, ahogy az alábbi példákból is látszik, azonban nem tükrözik az eredeti szavak által előhívott érzéseket, hangulatokat, azaz, azt, ami az anyanyelvi beszélő eszébe jut, amikor hallja őket [11].

Példák a magyar nyelvből:

- **Luca széke** – *Luca's chair* (az angol fordítás semmit sem árul el a kapcsolódó népi hiedelemről);
- **Máglyarakás** – *stake* (a magyarban ez egy sütemény, melynek jelentése nem adható vissza az angol szóval).

Példák az angol nyelvből:

- ***Anglia*** – Anglia latinul (a magyarban nincs megkülönböztetés, mivel a magyarban az England megfelelője Anglia);
- ***Sassenach*** – angol személyt jelölő skót terminus; nincs lexikalizált magyar megfelelője.

3.1.2 Gyűjtőfogalmak

A nem lexikalizált synsetek egy másik csoportja olyan elemeket tartalmaz, amelyeknek nincs megfelelőjük az adott nyelvben. Igen gyakran bizonyos, ebbe az osztályba tartozó gyűjtőfogalmakat csak körülírással vagy lista megadásával lehet kifejezni a másik nyelvben. Például:

Learned profession:1, a jog- , orvos- és teológia tudományának gyűjtőneve, melyet a magyar nem tud kifejezni egyetlen szóval, csak a három területet tudjuk felsorolni.

Ami a **drug:1**-et illeti, a HuWN-ben nincs egyszavas megfelelője, mivel a magyarban jól elkülönül a gyógyszer a kábítószertől, bár az utóbbit használják orvosi értelemben olyan anyagok jelölésére, melyeknek nagyon erős és tartós fájdalomcsillapító hatásuk van.

3.1.3 Fosztóképzővel ellátott synsetek

A nem lexikalizált synsetek egy másik, alappéldája a fosztóképzővel képzett melléknevek/főnevek olyan prefixumokkal, mint a *non-*, *in-*, *un-* stb. Néhány esettől eltekintve, az ilyen fosztóképzővel képzett lexikai egységek magyar megfelelőit negatív határozókkal képezzük, és ezek együtt nem alkotnak lexikalizált synseteket; például: *unattractive* – nem vonzó; *ill-timed* – rosszul időzített; *incongruity* – meg nem egyezés stb.

3.1.4 Melléknév + főnév szerkezetek

A magyarban bizonyos PWN-ben található fogalmakat melléknév + főnév szerkezetekkel fejezünk ki és ezeket nem tekintjük lexikai egységeknek, mert vagy produktívak, vagy pedig jelentésük teljesen kompozicionális.

Például az **Englishman:1**/**Englishwoman:1** (*English male* 'angol férfi' *English woman* 'angol nő') nem lexikalizált egységek a HuWN-ben, mert a magyarban nincs nyelvtani nem. Másrészt az *Englishman* magyar megfelelője, az 'angol' bekerülhetett volna a HuWN-be. Ugyanakkor az **Englishwoman:1** magyar megfelelője, az 'angol nő' nem vehető fel a HuWN-be.

A HuWN sajnos nem túl következetes e tekintetben. Lásd pl. **Scotsman:1**-t, melyet megfelelően 'skót'-nak vettek fel. A magyarban a 'skót', 'angol', 'magyar' szavaknak nincs neme, e szavak mégis elsősorban az adott nemzet hímnemű tagjára utalnak és nőnemű párjukat a 'nő' hozzáadásával kapjuk meg. A 'skót nő' összetételt azonban már produktív szerkezetnek (melléknév + főnév) és nem többszavas kifejezésnek tekintjük (, mely a magyarban a fenti szerkezetek feltétele a HuWN-be való bekerülésre), ezért nem vettük fel a magyar wordnetbe.

3.1.5 Nyelvtani különbségek

Némely esetben a nem lexikalizált synset nyelvtani különbségekből adódik. A **people:1**-nek (embercsoport) konceptuális szinten van, de lexikai szinten nincs megfelelője a magyarban: például a *200 people* magyarra a 'kétszázan' szóval adható vissza, ahol az esetrag az angol főnévnek felel meg.

Példa a nem lexikalizált melléknevekre a HuWN-ben a **comfortable:1**, **uncomfortable:2** synsetek. A HuWN-be nem lehetséges felvenni a cselekvés ágensét és experiensét egy synsetbe, ami viszont a PWN-ben gyakran előfordul.

3.1.6 Átvételek

Idővel bizonyos nem lexikalizált fogalmak lexikalizálódnak. E folyamat egyik tipikus területe a technológia, melynek fogalmai egyre gyorsuló ütemben terjednek világszerte. Néhány évvel ezelőtt, amikor a HuWN épült, pl. az *RV* (*recreational vehicle*) *non-lex* címkét kapott, ám most már teljes jogú lexikalizált synsetként felvehető lenne a HuWN-be.

3.2 Technikai nem lexikalizált synsetek

A wordnetépítés során gyakran előfordult, hogy két hipernima relációban lévő angol synsetnek egy magyar megfelelője volt; a két fogalom csak a konceptuális szinten különül el, lexikai szinten azonban nem találunk két külön szót. Ez azzal a következménnyel járna a HuWN-re, hogy a magyar szó önmaga hipernimája lenne. Ez volt a fő oka annak, hogy bevezettük a technikai nem lexikalizált (*t non-lex*) címkét.

A *t non-lex* címkét a következő esetekben használjuk: szófaji eltérés, azonos literálok hipernima relációban, azonos literálok *similar_to* relációban.

3.2.1 Eltérő szófaj

Különbségeket a két nyelv lexikonjában is találunk. Némely esetben a synset megfelelője a célnyelvben más szófajú, de a wordnetekben megengedett négy szófaj egyike. Például az *afraid* szó az angolban melléknév, viszont a magyarban a 'fél' igével adható vissza. Ezekben az esetekben vettük hasznát az ún. *eq_xpos_synonym* relációnak, mely eltérő szófajok közt jelöl szinonímiát és a magyar synset pedig *t non-lex* címkét kapott.

3.2.2 Azonos literálok hipernima relációban

A *t non-lex* címkézés második esete két azonos literál hipernima relációban lévő synsetekben. A címkézést azzal indokolható, hogy automatikusan könnyebb lehetséges hibákat azonosítani. Ha ugyanaz a literál *x* és *y* synsetben is megjelenik és azok hipernima relációban vannak, akkor valószínű, hogy az annotátor hibázott.

Az is a wordnetépítés egyik alapelve, hogy a fogalmat helyettesíteni lehet a hipernimájával, ezért ésszerűnek tűnt, hogy a hiponimát nem vettük fel a HuWN-be.

Lásd a következő példát:

1 **curtain:1**	**függöny:2**
2 **drop curtain:1**	(függöny) *t non-lex*

Ebben az esetben a HuWN *t non-lex* synsetjének van egy szinonimája a 'színházi függöny', mely egy kollokáció és teljes joggal felvehető lett volna a wordnetbe. A hiponima helyzetben lévő azonos literál törlésének szabályának felfüggesztésével egy kéttagú synsetet kapunk ('függöny', 'színházi függöny'). Az a különös ebben a synsetben, hogy a két tag nem valódi szinonima, mivel nem minden esetben felcserélhetők:

*Előadás után a **függöny** leereszkedett.*
*Az egész várost felkutattam megfelelő anyagért **színházi függöny** készítéséhez.*

Az első mondatba csak a 'függöny' illeszkedik megfelelően, a 'színházi függöny' furcsán hangzik; a melléknév ('színházi') felesleges. A második esetben azonban ez annyiban módosul, hogy a melléknévi rész használata nélkül a 'függöny' (**curtain:1** a PWN-ben) általánosabb jelentése is előfordulhat.

3.2.3 Azonos literálok központi és szatellit synsetekben

Az ontológia melléknévi részében is alkalmaztuk a *t non-lex* címkét. Mivel építése az antonim párokon és a hozzájuk asszociáció révén kapcsolható, szinonim szatellit synseteken alapul, lehetséges, hogy amíg angolban eltérő szó szerepel a központi és szatellit synsetben, addig a magyarban mindkét helyen ugyanaz a synset jelenik meg. A wordnetépítés szabályai nem engedik meg, hogy azonos literálok szerepeljenek a központi és szatellit synsetben (vö. a hiper- és hiponima azonossága). Ebből következően ismét azt az eljárást követtük, hogy a központi synset lexikalizált marad és a specifikusabb szatellit synset kapja a *t non-lex* címkét.

Például a {**wide:1**; **broad:1**}'s szatellit synsetje a {**heavy:5**; **thick:5**}, de a magyarban a 'széles' mindkettőt lefedi, ezért a központi synset a {**széles:2**}, a szatellit synset pedig a {**széles:0**}.

A *t non-lex* címkével ellátott synseteknek – szemben a *non-lex* synsetekkel – van definíciója, példája és, a legtöbb esetben, ÉKSz-linkje is. Azért választottuk ezt a megoldást, mert ezek a synsetek létező fogalmak a magyarban, szavakkal kifejezhetők, és csak a wordnet szerkezetének köszönhető, hogy a *t non-lex* címkét kell alkalmaznunk.

4 Nem lexikalizált synsetekhez kapcsolódó wordnet hibák

Itt a PWN és HuWN néhány problémás synsetjét mutatjuk be megoldásaikkal együtt.

4.1 Problémák a fában

Bizonyos esetekben a synset és hipernimája nincs összhangban. Például a **location:1** PWN synset definíciója a következő: *a point or extent in space* ('térbeli pont vagy kiterjedés'); egyik hiponimája a **bilocation:1,** melynek definíciója**:** *the ability (said of certain Roman Catholic saints) to exist simultaneously in two locations* ('az a képesség (, melyet bizonyos római katolikus szentekről állítanak), hogy valaki egy időben, két helyen van jelen' (unique beginner synset: **entity:1**). Szerintünk a reláció nem megfelelő, mert a definíciók nem összeegyeztethetők és csak úgy tűnik, hogy szabályszerű hiper-hiponima párt alkotnak. Ehelyett a *bilocation*t az **ability:2**, **power:3**/**képesség:2**-höz kellene kapcsolni éppen PWN-ben szereplő definíció alapján vagy pedig a **phenomenon:1**/**jelenség:1**-hez. Ha a PWN szerkezetét meg akarjuk őrizni a HuWN-ben, a synsetet *non-lex*-nek kellene címkézni és egy új synsetet kellene létrehozni a megfelelő hipernima alatt (**képesség:2** vagy **jelenség:1**).

A PWN kritikátlan másolásának következményei helytelen synset relációk is lettek: pl. **alsó állkapocs:1/lower jaw:1 → állkapocs:2/jaw:1** hipernima relációban vannak, noha a megfelelő a *holo_part* ('része') reláció lenne.

4.2 Lexikalizált synsetek *non-lex* címkével

Bizonyos esetekben – meglátásunk szerint – a HuWN annotátorai vétettek hibát. Például a **labor:1** jelenleg egy *non-lex* synset, miközben teljes joggal lehetne lexikalizált a 'fizikai munka' kollokációval fordítva. Hasonlóképpen a **seating:1, area:1**-t is fel lehetett volna venni mint 'ülőhely'.

A synsetek egy másik csoportja a HuWN-ben – melyet helytelenül *non-lex* címkével láttak el – az, melyben a literálok birtokos esetben vannak (**rear:2**/'hátulja'; **front:2**/'eleje').

4.3 Lexikalizáltként felvett non-lex synsetek

A non-lex synsetek egy érdekes példája a **bow and arrow:1**/**íj és nyílvessző:1**. Meglátásunk szerint a synsetet helytelenül jelölték lexikalizáltnak, mivel – bár két része egy egységet alkot – a kilövőszerkezet és a lövedék nem alkotnak egy fogalmat a magyarban.

A PWN kritikátlan másolásának másik példája egy teljességgel nem létező (bár lehetséges) synsethez, a **fúvóeszköz:1**/**blower:1**-hez vezet a magyarban.

A PWN-ben, úgy tűnik, vannak olyan synsetek, melyek nyilvánvalóan nem alkotnak egységes fogalmat. A **small/large definite/indefinite quantity**, **creating from raw materials**, **sound property**, **change of integrity**, **creating by removal** stb. synseteket *non-lex*-nek tekintjük.

4.4 Öröklési problémák

Bizonyos synseteknek két vagy több hipernimája van a fában. Arra kívánunk rámutatni, hogy csak abban az esetben szabad megengedni a több hipernimát, ha a hiponim synsetek a hipernima összes jellemzőjét örökölhetik. Példa lehet erre a **relaxant:1**, melynek két hipernimája van (*drug* vagy *treatment*). A fában a synset a **treatment:1**-től terjed egészen az **act:2** legfelső szintű fogalomig. A fenti esetben a synset nemcsak a *drug*, hanem a *treatment* tulajdonságait is örökli, ami ahhoz az ellentmondáshoz vezet, hogy (hiponimája,) a *Valium* egyszerre entitás és emberi tevékenység.

5 A non-lex problémák lehetséges megoldásai

A magyar wordnetben található non-lex synsetek nagy száma felveti a wordnetépítési elvek felülvizsgálatának kérdését. A non-lex synsetek tulajdonképpen nem képezik részét az adott nyelvnek, és a nagyszámú non-lex elemet tartalmazó wordnetek aligha tükrözik megfelelően az adott nyelv fogalmi hierarchiáját. Azért, hogy megoldjuk ezeket a problémákat, azt javasoljuk, hogy csökkentsük a non-lex synsetek számát a következőkben ismertetendő módszerekkel.

5.1 Hiponima nélküli non-lex synsetek

Azt javasoljuk, hogy a hiponima nélküli non-lex synseteket töröljük a fából. Mivel a hipernimák minden kontextusban helyettesíthetik hiponimáikat, ez az eljárás nem ássa alá bizonyos fogalmak kifejezhetőségét. Ez a következő példák esetében lehet hasznos:

1 **freedom:1**	**szabadság:1**
2 **liberty:1**	(szabadság)

Magyarban nincs jelentéskülönbség a két PWN-fogalom közt, így a fában lejjebb elhelyezkedő non-lex synsetet törölni kell. Ez a megoldás egyéb kultúra- és földrajz-specifikus synsetek esetében is alkalmazható.

5.2 Gyűjtőfogalmak

Azokat az gyűjtőfogalmakat, amelyeket vissza lehet adni egy lista megadásával, egyszerűen törölni kell a fából és összes hiponimáit a hipernimájához kell csatolni. Például:

cycling:1 (kerékpározás, motorozás)

Ebben az esetben a 'kerékpározás' és 'motorozás' fogalmakat két külön synsetbe kell felvenni és a **sport:1** alá kell bekötni.

5.3 A fa újraépítése

Bizonyos esetekben a fa újraépítése tűnik a legmegfelelőbb megoldásnak. Legelőször is, hadd mutassuk be a problémát az alábbi PWN-ből és HuWN-ből vett farészlettel (a magyar átírások megfelelnek a PWN definícióinak):

1 **building:1**	**épület:1**
2 **place of worship:1**	(istentisztelet helye) *non-lex*
3 **church**	(keresztény templom) *non-lex*
temple:1	(nem keresztény templom) *non-lex*

A PWN-ben a **church:2** és a **temple:1** azonos szintű hiponim synsetjei **a place of worship:1**-nek, és jelenleg nincs lexikalizált megfelelőjük a magyar wordnetben. Azért, hogy „megszabaduljunk" három non-lex synsettől, azt javasoljuk, hogy a 'templom' synsetet (, mely magyarban valamely vallás istentiszteleti helyének, épületének felel meg), hipernima pozícióba kell helyezni párhuzamosan a **place of worship:1**-gyel. A másik két PWN synsetnek a magyarban nincs megfelelője, így helyük üresen marad.

1 **place of worship:1**	1 **templom:1**
2 **church:2**	(-)
temple:1	(-)

5.4 Többszavas kifejezések integrálása

A következő példa elgondolkodtatott az alapvető wordnetépítési elvekről:

1 **gutter:2**, **sewer:3**, **toilet:3** ('WC, ablak, csatorna; kidobhatod az ablakon')

A *misfortune resulting in lost effort or money* (’kárba veszett erőfeszítés vagy pénz’) jelentésű synsetet az annotátorok nem találták lexikalizálható elemnek. Ez arra a tényre vet fényt, hogy a HuWN sokkal inkább lexikai wordnet, mintsem konceptuális. Gyakran a magyar wordnet építői inkább a szóalakra figyeltek, mint a fogalomra, ezért nincs a PWN synsetnek lexikalizált megfelelője a magyarban. Azonban a fő gond az, hogy az angol literálok egy többszavas kifejezés részei (ebben az esetben egy idiómáé), melyeket mint (konceptuális) egységet (, azaz synsetet) lehetett volna felvenni. Mivel a legtöbb többszavas kifejezésnek megvan a megfelelője a másik nyelvben, a megfelelő synsetet könnyebben meg lehet találni.

A probléma megoldására azt javasoljuk, hogy a teljes idiómát vegyük fel egy lexikai egységként a wordnetek igei részében (az idiómák jellemzően komplex predikátumok), melyeket aztán könnyen lehet párosítani anélkül, hogy a névszói összetevők megfelelőit kellene keresnünk a másik nyelvben. Ezek alapján a következő synsetek állnak elő:

be in the gutter, go down the sewer, be in the toilet ’lehúzhatja a WC-n’, ’kidobhatja az ablakon’

Az idióma felvétele mint nyelvi egység sokkal hasznosabb a többnyelvűség szempontjából, mert így könnyebb azok megfelelőit megtalálni a másik nyelvben mint egyes részeit, másrészt pedig az egész idióma felvételre kerül, s nemcsak főnévi, igei vagy melléknévi részei[2]. Egyúttal az idiómák részeihez kapcsolódó non-lex synseteket is fel lehet számolni.

7 Az eredmények értékelése

A non-lex elemek kulturális vagy konceptuális különbségeket tükröznek és így nyelvek közti hasonlóság megállapítására szolgálhatnak. A magyar wordnet jelen formájában tartalmaz non-lex elemeket, de amennyiben valamikor sor kerül a felülvizsgálatára, érdemes lenne bizonyos elemeket törölni vagy lexikalizált elemként felvenni (ha hibásan non-lex synsetként lettek jelölve), így a HuWN igazán tükrözni tudná a magyar nyelv konceptuális hierarchiáját.

Azonban a *non-lex* jelölés több szakterületen is hasznos lehet, pl. a pszicholingvisztikában, ahol különböző nyelvek beszélői mentális fogalmainak hierarchiáját vetik össze – a non-lex synsetek expliciten jelzik ezeket a különbségeket. A kultúraspecifikus synseteknek a néprajz vehetné hasznát. A nyelvi különbségekből adódó non-lex synsetek (pl. fosztóképzős melléknevek) pedig hozzájárulhatnának az elméleti és kontrasztív nyelvészet kutatásaihoz.

A fentiekre alapozva tehát azt javasoljuk, hogy a magyar wordnetet két változatban kellene létrehozni: az egyiket, amennyire csak lehetséges, a PWN-hez kellene kötni, így megőrizve annak hierarchiáját (non-lex synsetekkel); a másiknak nem kellene non-lex elemeket tartalmaznia, hogy a magyar nyelv hierarchiáját tükrözze. A két verziót így a kutatási céloknak megfelelően lehetne felhasználni.

[2] E szófajok és a határozószavak alkotják a wordneteket.

8 Összegzés

Ebben a dolgozatban bemutattuk a két, HuWN-be bevezetett *non-lex* címkét (*non-lex* és *t non-lex*) és megvizsgáltuk, hogy mi áll a non-lex jelenség mögött: elsősorban kulturális és/vagy nyelvi különbségekre vezethetők vissza. Megpróbáltunk megoldásokkal is szolgálni a szükségtelen synsetek törlésével vagy a fa újrarendezésével.

Bár az adott nyelv hierarchiáját ábrázoló wordnetnek nem volna szabad non-lex elemeket tartalmaznia, mégis hasznosnak bizonyulhatnak különféle kutatási területek (pszicholingvisztika, néprajz stb.) szempontjából. Így azt javasoljuk, hogy amennyiben sor kerül a magyar wordnet revíziójára, a non-lex elemeket törölni kellene és így a magyar konceptuális hierarchiát tükröző wordnetet kapnánk, melyet elsősorban magyar nyelvű kutatásokra lehetne felhasználni, az eredetileg kiadott verzió pedig többnyelvű kutatások referencia-adatbázisaként szolgálhatna.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy.: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In: Proceedings of the Third International WordNet Conference. South Jeju Island, Korea (2006) 291–292
2. Alonge, A., Bloksma, L., Calzolari, N., Castellon, I., Marti, T., Peters, W., Vossen P.: The Linguistic Design of the EuroWordNet Database. Computers and the Humanities. Special Issue on EuroWordNet Vol.32, No. 2–3 (1998) 91–115
3. Bhattacharyya, P., Fellbaum, C., Vossen, P. (eds.): Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fourth Global WordNet Conference. Narosa Publishing House, Mumbai, India (2010)
4. Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., Broda, B.: Words, Concepts and Relations in the Construction of Polish WordNet. In: Proceedings of the Fourth Global WordNet Conference (2008) 167–68
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. University of Szeged, Szeged (2008) 311–320
6. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an On-line Lexical Database. International Journal of Lexicography Vol.3, No.4 (1990) 235–244

7. Raffaelli, I., Tadić, M., Bekavac, B., Agić, Ž.: Building Croatian WordNet. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008) 349–359
8. Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008)
9. Tufiş, D. (ed.): Romanian Journal of Information Science and Technology. Special Issue on BalkaNet Vol.7, No.1–2 (2004)
10. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology. Special Issue on BalkaNet Vol.7, No.1–2 (2004) 9–43
11. Zidoum, H.: Towards the Construction of a Comprehensive Arabic WordNet. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008) 531–544

A *Magyar szóelemtár* megalkotása és a *Magyar gyökszótár* előkészítő munkálatai

Kiss Gábor[1], Kiss Márton[1], Sáfrány-Kovalik Balázs[2], Tóth Dorottya[3]

[1] TINTA Könyvkiadó, 1116 Budapest, Kondorosi út 17.
{kissgabo, kissmarci}@tintakiado.hu
[2] PPKE ITK (hallgató), 1083 Budapest, Práter utca 50/a.
safba@digitus.itk.ppke.hu
[3] ELTE BTK (hallgató), 1088 Budapest, Múzeum krt. 4/A.
tdorottya90@gmail.com

Antal László a morféma fogalmát tisztázó 1959-es cikke [1] után 1964-ben Egy magyar morfématár ügyében című cikkében [2] ezt írja: „A modern nyelvészeti felfogás a nyelvet jelrendszernek, kódnak tekinti. […] A nyelv teljes leírásához […] minden, az adott időpontban élő kódtag, jel listába vétele is hozzátartozik. […] Amennyiben a nyelv alapvető eleme a morféma, úgy jogosult és szükséges olyan szótár készítése, amely morfémákat tartalmaz, pontosabban anyagát a morféma kategóriájában mutatja be. […] Persze az ilyen szótár valójában »morfématár« lesz, bár nevezhetnénk morfémaszótárnak is." Antal László ötletéből és Kiss Gábor egy korábbi tanulmányából [10] kiindulva készítettük el a magyar morfémaszótárt, azaz a *Magyar szóelemtárat*. Kiindulásunk a *Magyar értelmező kéziszótár* (ÉKSz2) [16] 73.542 címszava volt. Munkánk eredményeképpen összeszámolhatóvá vált, hogy 16.272 tőmorféma, 518 tőváltozat, 705 fiktív tő, 54 igekötő, 949 toldalék, illetve előtag és 82 idegen szótő építi fel a szótár címszavait. Átlagosan 2,46 morféma alkot egy-egy címszót. A *Magyar szóelemtár* jó kiindulási alap a *Magyar gyökszótár* elkészítéséhez.

1 Bevezetés

Bárczi Géza is felhívja a figyelmet arra, hogy „a nyelvi jelenségek általában nehezen határolhatók el pontosan szétváló kategóriákba" [3]. Ez különösen érvényes a magyar összetett és továbbképzett szavak morfológiai felbontása során, hiszen a szinkrón és a diakrón nyelvi síkok között nincs éles határvonal. Természetesen sok esetben a szóképzés, illetve a szóösszetétel elhomályosulhat, eltűnhet a nyelvhasználó előtt. Nyelvtörténetileg a *folt, jobbágy, jószág, kendő, kopár, kopasz, mond, orom, ország* szavak képzett szavak; a szóösszetétel ténye pedig a következő szavakban felismerhetetlen a mai nyelvhasználó előtt: *ifjú, férj, ezüst, arc, kengyel, ünnep, lány, sármány, kesztyű, nyolc.* Külön csoportot képeznek azok a szóösszetételek, melyek a nyelvhasználó számára bizonytalanul homályosak: *holnap, tegnap, testvér.* A kérdésről legutóbb alapos tanulmányt T. Somogyi Magda [19] tett közzé.

2 A *Magyar szóelemtár* munkálatai

Nem előzmény nélküli a magyar lexikográfiában és számítógépes nyelvészetben, hogy valamely szótár címszavainak sorát géppel dolgozzák fel. Papp Ferenc az egyik első magyar számítógépes nyelvészeti munkaként az 1960-as évek második felében végezte el a *Magyar Nyelv Értelmező Szótára* címszavainak kódolását Debrecenben. E munkálat eredményeképpen született meg a Magyar Nyelv Szóvégmutató szótára, amely az a tergó elrendezés mellett információt tartalmaz a címszó tőtípusáról, összetettségéről és ragozási típusáról is [14,15,10].

A *Magyar morfématár* munkálata során az ÉKSz[2] címszavaiban bejelöltük a szóelemhatárokat a következő módon: *ágyú+golyó+, áll+kapocs+, angóra+nyúl+*. A gépi reprezentáció során az elemhatároló jeleket szögletes zárójelben elhelyezett kódokkal valósítottuk meg: *ágyú*[1]*golyó*[1]*, áll*[1]*kapocs*[1]*, angóra*[1]*nyúl*[1]. A kódolás során a következő hat szóelem-kategóriát különböztettük meg, és jelöltük:

1. **szótő** [1]**:** *asztal*[1]*láb*[1]*; andrás*[1]*kereszt*[1]*; anya*[1]*csavar*[1]
2. **szótőváltozat** [2]**:** *alv*[2]*ó*[5]*; árk*[2]*ol*[5]*; asztmá*[2]*s*[5]*; bányá*[2]*sz*[5]
3. **fiktív tő** [3]**:** *acsar*[3]*og*[5]*; ápor*[3]*odik*[5]*; ford*[3]*ul*[5]*; ugr*[3]*ik*[5]
4. **igekötő** [4]**:** *át*[4]*gázol*[1]*; be*[4]*cipel*[1]*; meg*[4]*nyom*[1]
5. **toldalék (képző) vagy előtag** [5]**:** *ad*[1]*omány*[5]*; ág*[1]*as*[5]*; akaszt*[1]*ó*[5]*; anti*[5]*anyag*[1]
6. **idegen szó** [6]**:** *baseball*[6]; *know*[6]-*how*[6]; *kick*[6]-*box*[6]

Az ÉKSz[2] címszavainak felbontása, azaz a kódolás során számos kérdés merült fel, amelyek legtöbbször a szinkrón és a diakrón nyelvi síkok érintkezéséből, illetve egymásba csúszásából adódtak. Hiszen döntés kérdése, hogy például a *szarvas, sertés, farkas* szavakat egyetlen elemnek vagy több elemből állóknak tekintjük: *szarv*[1]*as*[5] <–> *szarvas*[1]*; serté*[2]*s*[5] <–> *sertés*[1]*; fark*[2]*as*[5] <–> *farkas*[1]. Általában a felbontás és a szétválasztás mellett döntöttünk, példaképpen néhány szó, amely felbontásra került: *étvágy, kerít, laktanya, növény.* Továbbá irányelvünk volt, hogy akkor jelölünk szóelemhatárt, ha a szóelemek kapcsolódása a mai magyar beszélő számára „átlátható'', érzékelhető.

Magyar szóelemtárból 6 részlet 10-10 kódolt címszóval:

in[3]dít[5]
in[3]dít[5]ás[5]
in[3]dít[5]ék[5]
in[3]dít[5]ó[5]
in[3]dít[5]ó[5]áll[1]ás[5]
in[3]dít[5]ó[5]gomb[1]
in[3]dít[5]ó[5]kar[1]
in[3]dít[5]ó[5]kulcs[1]
in[3]dít[5]ó[5]motor[1]
in[3]dít[5]ó[5]ok[1]

köpű[1]
köpü[2]l[5]
köpü[2]l[5]ő[5]
kör[1]
kőr[1]
kő[1]rács[1]
kő[1]rajz[1]
kő[1]rak[1]ás[5]
kör[1]bástya[1]
kör[1]be[5]

le[4]ad[1]
le[4]ad[1]ás[5]
le[4]ad[1]ó[5]
le[4]akaszt[5]
le[4]alacsony[1]ít[5]
le[4]alacsony[1]ít[5]ó[5]
le[4]alacsony[1]odik[5]
le[4]aláz[1]
le[4]aláz[1]kodik[5]
le[4]aláz[1]ó[5]

madám[1]
madár[1]
madár[1]berkenye[1]
madár[1]kép[1]ű[5]
madar[3]ász[5]
madar[3]ász[5]ik[5]
madár[1]birs[1]
madár[1]cseresznye[1]
madár[1]csicserg[2]és[5]
madár[1]dal[1]

mamut[1]
mamut[1]birtok[1]
mamut[1]cég[1]
mamut[1]fenyő[1]
mamut[1]jöv[2]edelem[5]
mamut[1]sziv[2]attyú[5]
ma[1]nap[1]ság[5]
mancs[1]
mandarin[1]
mandátum[1]vizsgá[2]l[5]ó[5]

prém[1]
prém[1]es[5]
prémcsi[1]
prém[1]ez[5]
prém[1]gallér[1]
prém[1]gallér[1]os[5]
premier[1]
premier[1]ajándék[1]
premier[1]film[1]
premissza[1]

utó-[1]
utó[1]él[1]et[5]
utó[1]idény[1]
utó[1]ját [1]ék[5]
utó[1]rend[1]el[5]és[5]
utó[1]szül[1]ött[5]
utó[1]vég[1]re[5]
utó[1]bb[5]
utó[1]bb[5]i[5]
utó[1]d[5]

A kódolás ellenőrzéséhez a *Magyar szóelemtárat* elhelyeztük a világhálón, majd szerkesztő-, illetve konkordanciakészítő és lekérdező felületet hoztunk létre, amelynek segítségével a kiindulási szótár címszójegyzékében szétszórtan elhelyezkedő elemek kódolását egységesítettük.

Pl.: *anya*[1]*serté*[1]*s*[5]*; bacon*[1]*serté*[2]*s*[5]*; híz*[1]*ó*[5]*serté*[2]*s*[5].

3 Eredmények

A munkálat során létrehoztuk a *Magyar szóelemtárat*, amelyet a következő elemek építenek fel:

16.272 egyedi tőmorféma 96.645 előfordulással,
518 egyedi tőváltozat 4616 előfordulással,
705 egyedi fiktív tő 5988 előfordulással,
54 egyedi igekötő 11.275 előfordulással,
949 toldalék, ill. előtag 62.282 előfordulással,
82 idegen szótő 108 előfordulással.

A *Magyar szóelemtár* internetes elérhetősége: (felhasználónév: MSZNY, jelszó: szoelem) http://tintakiado.hu/szotar/szoelemtar/

MAGYAR SZÓELEMTÁR [kereső] [konkordancia]								
	szó-elem	szótő	módosult tő	fiktív tő	toldalék, előtag	igekötő	idegen szó	szum-ma
1.	ó	49			3988			4037
2.	ik				3845			3845
3.	ás	31			3473			3504
4.	ő	13			3107			3120
5.	és	2			2456			2458
6.	el	20			988	1394		2402
7.	ít				2074			2074
8.	i	57		2	1832			1891
9.	es			118	1668			1786
10.	z	7			1774			1781
11.	meg	27				1748		1775
12.	ki	11				1669		1680
13.	os				1646			1646
14.	et				1428			1429
15.	s	7			1339			1346
16.	l	1		1	1209			1211
17.	ol				1210			1250
18.	at				1150			1150
19.	ség				1139			1139
20.	be	1			70	1057		1128

A Magyar szóelemtár *internetes felületének konkordanciarészlete. Az első 20 szóelem összes előfordulása szerint sorba rendezve*

A *Magyar szóelemtár* felépítése után számszerűen rendelkezésünkre áll, hogy az egyes szóelemek milyen mértékben, hányszor vesznek részt az ÉKSz[2] címszavainak felépítésében. A következő 20 tőszó mindegyike több mint 250 alkalommal szóalkotó elem, gyakorisági sorrendben: *fa, köz, ház, szer, fog, kép, rend, von, áll, egy, szín, él, víz, szám, fő, gép, hely, jár, szó, tan.*

Lexikográfiai és szótárírói segédeszközként is használható, hiszen a *Magyar szóelemtárból* például kikereshetővé vált annak a 156 tőszónak a listája, amely tőszóként nem, hanem csak összetételi tagként szerepel a *Magyar értelmező kéziszótárban*. Pl: *-arábikum, -istók, -pipőke, csicseri-, esztrád-, kardán-*. Ugyancsak listázhatóvá váltak a címszójegyzék összetett szavaiban található tulajdon- és keresztnevek. Pl.: *leiter*[1]*jakab*[1]*, szent*[1]*János*[1]*áldás*[1]*, borzas*[1]*kata*[1].

Lehetővé vált a magyar nyelv számos szóalapú (értelmező és egyéb típusú) szótára után egy morfémaalapú szótárnak az elkészítése.

4 A *Magyar gyökszótár* munkálatainak előkészítése

A magyar szótárkiadás a 20. században mindvégig olyan értelmező szótárakat adott ki, melyek címszavai ábécérendben követték egymást. Azonban a szavak szótári besorolásának és közreadásnak nemcsak ez a mechanikus besorolás az egyetlen módja, hanem elképzelhető és megvalósítható egy olyan szótári közreadás, ahol a szavakat felépítő szóelemek (szavak, toldalékok) alkotják a rendező elvet, legyenek a szóelemek szókezdő, szó belseji vagy szóvégi helyzetben. Ezt a szemléletet valósította meg Kresznerics Ferenc 1838-ban kiadott *Magyar szótár gyökérrenddel és deákozattal* című munkájában [12]. Minta Kresznerics Ferenc szótárából:

DUG
DUG dugja, bele dugja, bé dugja, el dugja, ki dugja; DUGA donga, dugába dől; DUGACS dugacsol, dugacsolja, bé dugacsolja; DUGASZ s.g-dugasz, dugaszol, dugaszolja, be dugaszolja, el dugaszolja; DUGÁS bé dugás, el dugás; DUGDOS dugdossa, bé dugdossa; DUGGAT duggatja; DUGGOGAT, duggogatja; DUGÓ dugni való; DUGTIG dugulás, bé dugulás, dugult, dugultság; DUGVA

PÖR
PÖR vas-pör; PÖRCEN, pörcenet, óra-pörcenés; PÖRD pördít, pördíti, meg pördíti, pördíthető, pördíthetetlen, pördül, bé pördül, el pörült, PÖRG, PÖRÖG pögec, pörgeldik, pörgés, pörgés, pörget, pörgeti, pörgetés, pörgettet, pörgettyű, pörgetve, pörgő, pörgő óra, pörgő rokka

A Czuczor–Fogarasi-szótár ábécérendben közreadott (és értelmezett) szavainak a sorát rendre megtöri és keresztbeszövi a szóelem, azaz a hajdani szerzők által használt terminussal, a gyökök szerinti csoportosítás [11].

A *Magyar morfématár* elkészülte után lehetővé vált egy olyan magyar gyökszótár összeállítása, melynek anyagának vezérlő elve az ÉKSz[2] címszavainak egy olyan közreadása, ahol egy-egy szócikkben együtt látjuk mindazokat a szavakat, amelyekben megtalálható egy adott szóelem (a 19. századi terminussal élve gyök).

5 Mutatvány a készülő *Magyar gyökszótárból*

-oda *képző* (53 db)
állat|óv|oda, fésű|s|fon|oda, finom|fon|oda, fiók|ir|oda, fogad|ó|ir|oda, fon|oda, for|dít|ó|ir|oda, gőz|mos|oda, gyűrű|s|fon|oda, hang|verseny|ir|oda, hir|det|ő|ir|oda, ing|atlan|ir|oda, ir|oda, ir|oda|bútor, ir|oda|ép|ül|et, ir|oda|gép, ir|oda|ház, ir|oda|i, ir|oda|igaz|gat|ó, ir|oda|kis|asszony, ir|oda|kukac, ir|oda|szer, ir|oda|technika, ir|oda|tiszt, jegy|ir|oda, kabinet|ir|oda, kém|ir|oda, luxus|száll|oda, men|et|jegy|ir|oda,

mos|oda, ok|mány|ir|oda, óv|oda, óv|oda|pedagógus, panasz|ir|oda, párt|ir|oda, sajtó|ir|oda, sport|usz|oda, száll|oda, száll|oda|i, száll|oda|ipar, száll|oda|lánc, száll|oda|portá|s, száll|oda|tolvaj, száll|oda|tűz, száll|oda|váll|al|at, szín|i|tan|oda, tan|oda, terv|ez|ő|ir|oda, tud|akoz|ó|ir|oda, usz|oda, varr|oda, verseny|ir|oda, verseny|usz|oda

iskola, iskolá- *főnév* (97 db)

alap|iskola, balett|iskola, be|iskolá|z, elő|iskola, fa|iskola, fest|ő|iskola, fiú|iskola, fő|iskola, fő|iskola|i, fő|iskolá|s, gyakor|l|ó|iskola, had|apró|d|iskola, hegedű|iskola, inas|iskola, ipar|iskola, iskola, iskola|beteg|ség, iskola|busz, iskola|dráma, iskola|ép|ül|et, iskola|ér|ett, iskola|év, iskola|fenn|tart|ó, iskola|gép, iskola|gyakor|l|at, iskola|hagy|ott, iskola|hajó, iskola|i, iskola|igaz|gat|ó, iskola|játék, iskola|ker|ül|ő, iskola|könyv, iskola|köpeny, iskola|kötel|es, iskola|kötel|ez|ett|ség, iskola|lát|ogat|ás, iskola|lov|ag|l|ás, iskola|mester, iskola|mul|aszt|ás, iskola|orvos, iskola|pad, iskola|parancs|nok, iskola|példa, iskola|pénz, iskola|rádió, iskola|rend|szer, iskola|rep|ül|és, iskola|ruha, iskolá|s, iskolá|s|kor, iskolá|s|kor|ú, iskola|szék, iskola|szer, iskola|szolga, iskola|társ, iskola|táska, iskola|tej, iskola|televízió, iskola|tévé, iskola|típus, iskola|titkár, iskola|udvar, iskola|ügy, iskola|város, iskolá|z, iskolá|z|ás, iskolá|z|atlan, iskolá|z|ik, iskolá|z|ott, iskolá|z|tat, ismét|l|ő|iskola, kadét|iskola, kis|iskolá|s, közép|iskola, közép|iskolá|s, leány|iskola, lő|iskola, magán|iskola, magas|iskola, minta|rajz|iskola, munka|iskola, nép|fő|iskola, nép|iskola, nyelv|iskola, párt|fő|iskola, párt|iskola, reál|iskola, szak|iskola, szak|közép|iskola, szín|i|iskola, tan|onc|iskola, tánc|iskola, ugr|ó|iskola, vív|ó|iskola, zene|iskola, zongora|iskola, zug|iskola

for- *ige* (86 db)

alá|for|dít, alá|for|dul, át|for|dít, át|for|dul, be|for|dít, be|for|dul, bele|for|dít, bele|for|dul, egy|for|dul|ó|s, el|ford|ít, el|ford|ul, elő|for|dul, év|for|dul|ó, ezr|ed|for|dul|ó, fel|for|dít, fel|for|dul, fel|for|dul|ás, fél|for|dul|at, fel|for|dul|t, félre|for|dít, félre|for|dul, for|dít, for|dít|ás, for|dít|ó, for|dít|ó|gép, for|dít|ó|i, for|dít|ó|ir|oda, for|dít|ó|korong, for|dít|ó|program, for|dít|ó|s, for|dít|ó|szó|tár, for|dít|ott, for|dít|va, for|dul, for|dul|ás, for|dul|at, for|dul|at|os, for|dul|at|szám, for|dul|at|szám|lál|ó, for|dul|ó, for|dul|ó|pont, hátra|for|dít, hátra|for|dul, hova|for|dít|ás, ker|ül|-for|dul, két|for|dul|ó|s, ki|for|dít, ki|for|dul, kocsi|for|dul|ó, kor|for|dul|ó, kor|szak|for|dul|ó, kör|for|dul|at, körül|for|dul, le|for|dít, le|for|dít|hatatlan, le|for|dul, lépcső|for|dul|ó, meg|for|dít, meg|for|dul, moz|d|ony|for|dít|ó, mű|for|dít, mű|for|dít|ás, mű|for|dít|ó, nap|for|dul|ó, nyers|for|dít|ás, oda|for|dul, pál|for|dul|ás, posta|for|dul|ta, próba|for|dít|ás, rá|for|dít, rá|for|dít|ás, sors|for|dít|ó, sors|for|dul|at, sors|for|dul|ó, szak|for|dít|ó, száz|ad|for|dul|ó, száz|ad|for|dul|ó|s, szembe|for|dul, tér|ül|-for|dul, test|for|dul|at, törzs|for|dít|ás, tükör|for|dít|ás, út|for|dul|ó, világ|fel|for|dul|ás, vissza|for|dít, vissza|for|dít|hatatlan, vissza|for|dul

olvas *ige* (61 db)

át|olvas, bele|olvas, be|olvas, el|olvas, elő|olvas, fel|olvas, fel|olvas|ás, fel|olvas|ó|ül|és, gáz|le|olvas|ó, gond|ol|at|olvas|ás, gond|ol|at|olvas|ó, gyors|olvas|ás, hír|olvas|ó, hozzá|olvas, kártya|le|olvas|ó, ki|olvas, ki|olvas|ó, kotta|olvas|ás, lap|olvas|ó, le|olvas, le|olvas|ó, meg|olvas, név|sor|olvas|ás, olvas, olvas|ás, ol-

vas|ás|mód, olvas|at, olvas|atlan, olvas|gat, olvas|hatatlan, olvas|ható, olvas|mány, olvas|mány|os, olvas|ni|való, olvas|ó, olvas|ó|jegy, olvas|ó|jel, olvas|ó|könyv, olvas|ó|kör, olvas|ó|köz|ön|ség, olvas|ó|lámpa, olvas|ó|léc, olvas|ó|napló, olvas|ó|próba, olvas|ó|szem|üveg, olvas|ó|szerkeszt|ő, olvas|ó|szolgá|l|at, olvas|ó|tábor, olvas|ó|terem, olvas|ott, olvas|ott|ság, olvas|tat, össze|olvas, össze|olvas|ás, rá|olvas, rá|olvas|ás, tér|kép|olvas|ás, újra|olvas, új|ság|olvas|ó, végig|olvas, vissza|olvas

farok, fark- *főnév* (27 db)
egér|fark|kóró, fark, fark|all|ó, fark|a|pénz, fark|atlan, fark|csigolya, fark|csont, fark|csóvá|l|ás, fark|inca, fark|os, fark|toll, fark|úsz|ó, farok, farok|csigolya, farok|csont, farok|felület, fecske|fark, fecske|fark|köt|és, fecske|fark|ú, hód|fark|ú, ló|fark, nyúl|fark|fű, nyúl|fark|nyi, ökör|fark|kóró, róka|fark|ú, rozsda|fark|ú, ürge|fark

-ékony képző (31 db)
áll|ékony, alusz|ékony, boml|ékony, csal|ékony, fár|ad|ékony, fáz|ékony, fog|ékony, fog|ékony|ság, foly|ékony, foszl|ékony, gyúl|ékony, hajl|ékony, hajl|ékony|ság, hat|ékony, herv|ad|ékony, híz|ékony, ill|ékony, izgul|ékony, lobb|an|ékony, máll|ékony, moz|g|ékony, múl|ékony, nyúl|ékony, olv|ad|ékony, robb|an|ékony, rom|l|ékony, rug|ékony, sim|ul|ékony, talál|ékony, tan|ul|ékony, vált|oz|ékony

Bibliográfia

1. Antal, L.: A morfémáról. Magyar Nyelv Vol. LV. (1959) 16–22
2. Antal, L.: Egy magyar morfématár ügyében. In.: Tanulmányok a magyar nyelv életrajza köréből. Nyelvtudományi Értekezések 40. sz. Akadémiai Kiadó, Budapest (1964) 22–27
3. Bárczi, G.: Magyar történeti szóalaktan I. A szótövek. (Egyetemi Magyar Nyelvészeti Füzetek.) Tankönyvkiadó, Budapest (1958)
4. D Bartha, K.: Magyar történeti szóalaktan II. A magyar szóképzés története. (Egyetemi Magyar Nyelvészeti Füzetek.) Tankönyvkiadó, Budapest (1958)
5. Benkő, L. (főszerk.): A magyar nyelv történeti-etimológiai szótára I–III. Akadémiai Kiadó, Budapest (1967–1976)
6. Benkő, L.: Magyar fiktív (passzív) tövű igék. Akadémiai Kiadó, Budapest (1984)
7. Czuczor, G., Fogarasi, J. (szerk.): A magyar nyelv szótára I–VI. Pest (1862–1874) [Reprint kiadása: Pytheas Kiadó, 2010.]
8. Hegedűs, R.: Magyar nyelvtan. Formák, funkciók, összefüggések. Tinta Könyvkiadó, Budapest (2005)
9. Keszler, B.: A szóképzés. In: Keszler, B. (szerk.): Magyar grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 307–346
10. Kiss, G.: A Magyar Nyelv Értelmező Szótára címszavainak összetettsége. In: Horváth, K., Ladányi, M.: Elemszerkezetek és linearitás. A jelentés és szerkezet összefüggése. Bessenyei György Könyvkiadó, Nyíregyháza (1998)
11. Kiss, G.: A Czuczor–Fogarasi-szótár helye a magyar szótárirodalomban. Életünk. Vol. XLIX. No. 3–4 (2011) 84–91
12. Kresznerics, F.: Magyar szótár gyökérrenddel és deákozattal, I–II. Buda (1831–1832) [Hasonmás kiadása: Tinta Könyvkiadó, 2010.]
13. Ladányi, M.: Produktivitás és analógia a szóképzésben: elvek és esetek. (Segédkönyvek a nyelvészet tanulmányozásához 76.) Tinta Könyvkiadó, Budapest (2007)

14. Papp, F.: A magyar nyelv szóvégmutató szótára. Akadémiai Kiadó, Budapest (1969)
15. Papp, F.: A debreceni thészaurusz. Magyar Tudományos Akadémia Nyelvtudományi Intézete, Budapest (2000)
16. Pusztai, F. (főszerk.): Magyar értelmező kéziszótár (ÉKSz2). Akadémiai Kiadó, Budapest (2007)
17. Simonyi, Zs.: Tüzetes magyar nyelvtan. Magyar hangtan és alaktan. Magyar Tudományos Akadémia, Budapest (1885)
18. T. Somogyi, M.: Toldalékrendszerezésünk vitás kérdései. (Segédkönyvek a nyelvészet tanulmányozásához 3.) TINTA Könyvkiadó, Budapest (2000)
19. T. Somogyi, M: A felújított és megújított képzők. A nyelvújítás hatása a képzőrendszerre. In: Bakró-Nagy, M., Forgács, T. (szerk.): A nyelvtörténeti kutatások újabb eredményei. VI. Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék, Szeged (2011) 229−247
20. Tompa, J. (szerk.): A mai magyar nyelv rendszere. Leíró nyelvtan, I−II. Akadémiai Kiadó, Budapest (1961)
21. H. Varga, M.: Egyszerű vagy összetett képzők? Magyar Nyelvőr Vol. 124 (2000) 514−519
22. Veenker, W.: Mitteilungen der Societas Uralo-Altaica. Heft 3. Verzeichnis der Ungarischen Suffixe und Suffixkombinationen. Hamburg, kézirat (1968)

III. Szintaxis, morfológia, névelem-felismerés

A sekély mondattani elemzés további lépései

Recski Gábor

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

1. Bevezetés

A sekély mondattani elemzés (shallow parsing), mely a mondatok fő összetevőinek azonosítását jelenti a mély mondatszerkezet feltérképezése nélkül, számos nyelvtechnológiai eljárás kulcsfontosságú lépése. A legnagyobb mondattani egységek pontos azonosítása nélkülözhetetlen lehet a gépi megértésben, a gépi fordításban, de az információkinyerésben és -visszakeresésben is. Cikkünkben elsőként bemutatjuk, hogyan képes az eredetileg főnévi csoportok azonosítására kifejlesztett `hunchunk` rendszer a megfelelő tanulóadat birtokában tetszőleges kategóriájú frázisok azonosítására. A 2.1 fejezetben röviden összefoglaljuk a tanulóadat előállításának és a rendszer tanításának menetét, a 2.2. részben a `hunchunk` felépítéséről ejtünk néhány szót, végül a 2.3 fejezetben értékeljük a rendszer teljesítményét.

A mondat sekély szerkezetének megismeréséhez elengedhetetlen, hogy azonosítani tudjuk a több, gyakran nem szomszédos szóból álló igei szerkezeteket. A 3.1 fejezetben egy olyan eszközt ismertetünk, mely azonosítja egy ige és a tőle különálló igekötő kapcsolatát – felhasználva ehhez a rendelkezésre álló morfológiai elemzést, valamint az egyes igekötős igék gyakoriságáról meglévő ismereteinket is. Ugyancsak a mondatszerkezet hatékonyabb feltérképezését segíti elő, ha képesek vagyunk észlelni az igéből és annak infinitívuszi bővítményéből álló szerkezeteket - a 3.2. fejezetben erre teszünk kísérletet.

2. Mondattani egységek azonosítása

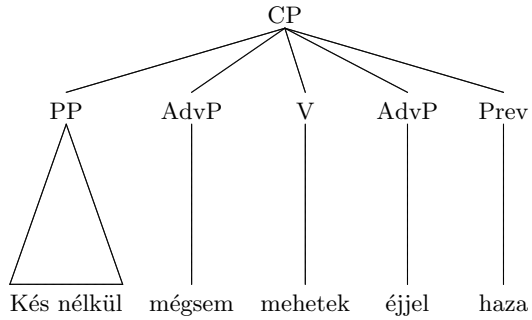
A `hunchunk` rendszer [1] magyar főnévi csoportok azonosítására készült, azonban megfelelő tanulóadat birtokában tetszőleges olyan nyelvfeldolgozási feladatra alkalmas, mely szószintű címkézésként is megfogalmazható. A Szeged Treebank [2] segítségével a főnévtől különböző mondattani kategóriákra is készíthetünk tanulóadatot, így lehetővé téve, hogy a `hunchunk` a legmagasabb szintű mondattani egységeket azonosítsa.

2.1. Tanítás

A Szeged Treebank egy vegyes műfajú, több mint 80000, szintaktikailag teljesen annotált mondatot tartalmazó korpusz. A tanítóadat előállításához a mondat-

tani elemzés legfelső két szintjét használjuk – a legfelső szinten a tagmondatok (CP) különülnek el, az ezek alatti legmagasabb szintű egységek azok, melyeket azonosítani szeretnénk. A korpuszból ugyancsak kinyerhető az egyes szavakra vonatkozó morfológiai információ MSD-kódolásban, ezt a korpusz készítésekor átalakítottuk a KR-formalizmusnak megfelelő alakra [3], mivel az általunk használt `hunmorph` morfológiai elemző [4] is ezt a formátumot követi.

Az egyes frázisokhoz tartozást a szavakhoz rendelt címkék jelzik. A címkézés során a Start/End konvenciót alkalmazzuk [5], mely az elterjedtebb IO és IOB konvencióknál [6] több címkét igényel, ugyanakkor lehetővé teszi többféle frázisbeli pozíció megkülönböztetését: míg az előbbi megoldások vagy egy címkével (I-NP) jelölik a frázishoz tartozó szavakat, esetleg a frázist kezdő szót jelölik külön szimbólummal (B-NP), addig az általunk használt jelölés a chunkhoz nem tartozó szavakon (O) kívül négy címkét használ (B-NP, I-NP, E-NP, 1-NP), melyek rendre a frázis elején, közepén és végén álló, valamint az önmagában frázist alkotó szavakat jelölik. Így a korpuszban található, 1. ábra szerinti elemzéssel bíró mondat az újonnan létrejött korpuszban a 1. táblázat szerinti címkézést kapja.



1. ábra. Mondattani elemzés

1. táblázat. Címkézés

Kés	nélkül	mégsem	mehetek	éjjel	haza	.
B-PP	E-PP	1-ADVP	O	1-ADVP	O	O

Az egyes mondattani kategóriák nagyon különböző gyakorisággal fordulnak elő maximális frázisként a korpuszban (l. 2. táblázat). Mint látható, melléknévi frázis csak elvétve fordul elő tagmondat közvetlen összetevőjeként, akkor is általában hibás annotáció következményeként (vö. *A köd mint* $[_{AdjP}$ melegvizes] *rongy feküdt az arcomon*).

2. táblázat. Kategóriák megoszlása a korpuszban

NP	268726	73.58%
ADVP	79536	21.78%
PP	16925	4.63%
ADJP	34	0.00%
Összesen	365221	100%

2.2. A `hunchunk` rendszer

A `hunchunk` egy felügyelt tanulásra épülő, szószintű címkézési feladatokat ellátó eszköz, melyet sikerrel alkalmaztunk főnévi csoportok azonosítására és tulajdonnév-felismerésre [1,7]. A rendszer a maximum entrópia módszerrel tanul [8], majd egy-egy mondat legvalószínűbb címkézését rejtett Markov-modellekkel [9], az egyes címkék közötti átmenetvalószínűségek figyelembevételével keresi meg. Az újfajta modell tanítása során változtatás nélkül alkalmaztuk azt a jegykészletet és azon beállításokat, melyek a maximális főnévi csoportok azonosítása során a legsikeresebbnek bizonyultak. Változást a folyamatban csupán az jelentett, hogy a sokszorosára bővült címkekészlet (5 helyett 21 különböző címke) jelentősen növeli mind a tanítás, mind a címkézés idejét.

2.3. Kiértékelés

A tanítást a korpusz 90 százalékán végeztük, a fennmaradó 10 százalékon mértük az eszköz teljesítményét. A rendszer teljesítményét két adat, a pontosság és a fedés jellemzi, a helyesen megtalált frázisok arányát előbbi az összes azonosított frázis arányában, utóbbi a tényleges frázisok arányában mutatja. A szakirodalomban megszokott módon a két érték harmonikus közepeként előálló ún. F-pontszámmal jellemezzük a rendszer általános teljesítményét. A `hunchunk` eredményei az egyes mondattani kategóriákon, valamint összesítve, a 3. táblázatban láthatók. Az `AdjP` kategóriát, mivel a tanulóadatban is nagyon ritkán és szabálytalanul voltak jelen, a címkéző is csak elvétve és látszólag „ok nékül” választotta, ennek hatása azonban elhanyagolható a rendszer összteljesítménye szempontjából.

3. táblázat.

	Pontosság	Fedés	F1
NP	89.36%	88.80%	89.08
ADVP	92.68%	92.99%	92.83
PP	88.70%	88.02%	88.36
ADJP	0.00%	0.00%	0.00
összesen	**90.06%**	**89.68%**	**89.87**

3. Igék

A sekély mondattani elemzés lehetővé teszi, hogy egy-egy mondaton belül azonosítsuk a főbb argumentumokat. Az állítmány azonosításához azonban olyan eszközre is szükségünk lesz, mely felfedezi az elvált igekötőket és a több szóból álló igei komplexumokat. A Szeged Treebank mindkét fajta függőségi viszonyt kódolja, így az elkészült eszközök teljesítményét módunkban áll kiértékelni.

3.1. Igekötők

A Szeged Treebankben található morfológiai elemzésből – csakúgy, mint a `hunmorph` morfológiai elemző kimenetéből – egyértelműen azonosíthatók az önmagukban álló igekötők. Célunk, hogy minél pontosabban tudjuk azonosítani, mely igéhez tartoznak. A kezdeti legegyszerűbb eljárásunk minden igekötőhöz a hozzá a mondatban legközelebb álló igét párosítja; ez a módszer az igekötő-ige párokat csupán 82% körüli F-pontszámmal azonosítja. A pontosságot kis mértékben javítja, ha az igét csak az igekötőhöz legközelebb álló írásjelek között keressük.

A legjelentősebb hibaosztályt az infinitívuszi konstrukciók okozzák (vö. *fel akar mászni*) – ha az infinitívusz mellett álló segédige kiváltja az igekötő elválását, akkor a segédige közelebb kerül az igekötőhöz, mint az infinitívusz alakban álló ige. Kálmán C. és mtsai [10] felsorolják azon segédigéket, melyek leggyakrabban az igekötő és ige közé kerülnek: *akar, bír, fog, kell, kezd, kíván, lehet, mer, óhajt, próbál, szabad, szándékozik, szeret, szokik, talál, tetszik, tud* (pp. 81-82)[1]; jelentős javulást érünk el, ha ezen igéket kizárjuk a keresésből. Célszerű volt továbbá kizárni a létigét, mivel különböző alakjaiban ugyancsak gyakran kerül egy ige és annak igekötője közé (vö. *meg lehet szokni*, *meg van csinálva*). A különböző eljárásokkal elért eredményeket a 4. táblázat összesíti.

4. táblázat. Igekötő-ige párok azonosítása

	Pontosság	Fedés	F1
baseline	82.81%	82.37%	82.59
+írásjelek között	84.41%	82.55%	83.47
+segédige szűrés	97.06%	93.41%	95.20
+létige szűrés	**97.52%**	**95.32%**	**96.41**

A hibák szemrevételezéséből kiderül, hogy azok túlnyomó többségét már a korpusz valamilyen apró hibája okozza. Így például nem járhat sikerrel az eljárás, ha bárhol is téves vagy hiányos az igék és igekötők morfológiai elemzése, vagy éppen a kiértékelés alapjául szolgáló mondattani annotációba csúszik apróbb hiba. Végül a hibaforrás sok esetben a korpuszban szereplő kétféle annotáció

[1] A segédigék beférkőzési hajlandóságáról tett megállapításokat [11] korpuszalapú vizsgálattal is megerősítette.

következetlensége egyes nem egyértelmű esetekben. Pl. az alábbi mondatban: *Vaksötét volt a fenékben, csak tapogatva jutott előre* az *előre* szó morfológiai elemzése szerint igekötő, a szintaktikai annotáció alapján azonban bővítmény. A jelenség fordítottja is előfordul: az *ide figyeljen* mondatban hiába jelez igekötő-ige viszonyt a korpusz, az algoritmusunk nem tudja azonosítani, mivel az *ide* szó a morfológiai elemzés szerint nem igekötő, hanem határozó. Ezen szavak grammatikai státuszának vizsgálata nyilvánvalóan túlmutat jelen cikk határain, az azonban kijelenthető, hogy az általunk eltévesztett párosítások jelentős része olyan szerkezeteket érint, amelyekről a kézi annotátorok sem hoztak következetes döntéseket.

3.2. Komplex igék

A több szóból álló igei szerkezetek egy másik gyakori, ámde könnyen azonosítható típusát adják a már említett, egy finit és egy *-ni* végű igéből álló szerkezetek. Magas pontosság érhető el a fentihez hasonló baseline módszer néhány triviális javításával. A módszer itt is csupán annyi, hogy a morfológia elemzés szerint infinitívuszi jeggyel bíró igéket a hozzájuk legközelebbi finit igéhez kapcsoljuk, nem lépve át közben írásjelet. A módszer pontosságát az 5. táblázat mutatja.

5. táblázat. Infinitívuszok és finit igék párosítása

Pontosság	Fedés	F1
97.02%	**96.35%**	**96.69**

Ez a baseline módszer az infinitívuszok két gyakori előfordulását is rosszul ismeri fel, ezek adják a hibák legnagyobb részét. Egyrészt nem kezeljük két infinitívusz függőségi viszonyát (vö. *Sürgősen igyekeznem kell Almirába jutni*), így a példamondatban a *jutni* szót nem az *igyekeznem* szóval kapcsoljuk össze. Ha azonban csak annyit módosítunk az algoritmuson, hogy nem követeljük meg a választott ige finitségét, akkor a módszer rosszul kezelné az olyan mondatokat, melyben egy finit igéhez több, egymást követő infinitívusz is társul, pl: *A madzagnagyiparos hűlni és zsibbadni kezdett.*

A másik nagy hibaosztályt a koordinált és vesszővel elválasztott infinitívuszok adják. Mivel a fenti eljárást nem egész mondatokon, hanem két írásjel közé eső szósorozatokon végezzük, így ha egy infinitívuszt mégis írásjel választ el a hozzá tartozó finit igétől, akkor ezt a párosítást biztosan nem találjuk meg (vö. *a szakadt ing mögött mégiscsak olyan szív dobog, amelyik tudott szeretni, fájni és aggódni is valamikor.*) Ha azonban általánosságban megengedjük az írásjeleken átívelő függőséget, akkor ez számos téves párosításhoz és így a pontosság jelentős romlásához vezet a fedés kismértékű növekedése mellett.

Mindkét problémára legalább részben megoldást jelentene, ha egy előfeldolgozási lépésben felismernénk a koordinált szerkezeteket. Ez egyúttal újabb hasznos eljárás lenne az alapvető mondatszerkezet feltérképezésére, így remélhetőleg a jövőben erre is sort keríthetünk.

4. Összefoglalás

Cikkünkben három, a magyar mondatok sekély szerkezetének feltérképezését szolgáló eljárást mutattunk be, melyeket a Szeged Treebank korpusz segítségével értékelünk ki. Megmutattuk, hogy a tagmondatok közvetlen összetevőit alkotó maximális frázisok a főnévi csoportokhoz hasonló hatékonysággal azonosíthatóak a felügyelt tanulásra alapuló `hunchunk` eszközzel. A cikk második felében két egyszerű eljárást írtunk le, melyek képesek morfológiailag elemzett szövegből kinyerni az elvált igekötőjű igéket és az ige+infinitívusz szerkezeteket. Mindkét eljárás 96 százalék feletti F-pontszámot ér el. Az igekötők és igék párosításakor a hibák legnagyobb részéért a korpuszban fellelhető ellentmondások felelnek, míg az infinitívuszok esetében a pontosság valószínűleg jelentősen javítható, amennyiben a több egymást követő infinitívuszi alakot tartalmazó mondatok szerkezetéről előzetesen több információt nyernénk ki.

Hivatkozások

1. Recski, G., Varga, D., Zséder, A., Kornai, A.: Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban [Identifying noun phrases in a parallel corpus of English and Hungarian]. VI. Magyar Számítógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics] (2009)
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue. (2005) 123–131
3. Rebrus, P., Vajda, P., Halácsy, P., Rung, A., Trón, V.: Általános célú morfológiai elemző kimeneti formalizmusa [Output formalism of a general-purpose morphological analyzer]. II. Magyar Számítógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics] (2004)
4. Trón, V., Kornai, A., Gyepesi, G., Németh, L., Halácsy, P., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the Workshop on Software, Association for Computational Linguistics (2005) 77–85
5. Uchimoto, K., Ma, Q., Murata, M., Ozaku, H., Isahara, H.: Named entity extraction based on a maximum entropy model and transformation rules. In: ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2000) 326–335
6. Sang, E.F.T.K., Veenstra, J.: Representing text chunks. In: EACL. (1999) 173–179
7. Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica **16** (2006) 293–301
8. Ratnaparkhi, A., et al.: A maximum entropy model for part-of-speech tagging. In: Proceedings of the conference on empirical methods in natural language processing. Volume 1. (1996) 133–142
9. Rabiner, R.L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. IEEE. Volume 77. (1989) 257–286
10. Kálmán C., G., Kálmán, L., Ádám Nádasdy, Prószéky, G.: A magyar segédigék rendszere. Általános Nyelvészeti Tanulmányok (1989) 49–103
11. Modrián-Horváth, B.: Gesichtspunkte zu einer funktionalen Typologie der Ungarischen Infinitiv regierenden Hilfsverben. Acta Linguistica Hungarica **56**(4) (2009) 405–439

Közösségkeresés alapú felügyelet nélküli szófaji egyértelműsítés

Berend Gábor[1], Vincze Veronika[2]

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

Kivonat Az előadásban bemutatjuk felügyelet nélküli szófaji egyértelműsítő módszerünket, mely közösségkeresésre épül. A közösségkereső eljárás bemenetéül szolgáló, a szóalakok fölött értelmezett hasonlósági gráf költséges számítására való tekintettel az elosztott rendszerek területén az ún. overlay topológiák közelítésére korábban már sikeresen alkalmazott T-MAN algoritmust alkalmaztuk. Eredményeink azt igazolják, hogy sikerült átültetnünk a két különböző tudományos közösség által használt módszerek előnyeit a szófaji egyértelműsítés területére, azaz egy olyan feladatra nyújtottunk így megoldást, amelyet egy harmadik tudományos közösség tűzött ki céljául.

Kulcsszavak: szófaji egyértelműsítés, közösségkeresés, felügyelet nélküli tanulás, modularitás

1. Bevezetés

A szófaji egyértelműsítés a természetes nyelvi feldolgozás egyik alapvető lépése: számos magasabb rendű alkalmazás hasznosítja jellemzőként a szófaji kódokat, azaz igen fontos, hogy a szövegszavakhoz hozzárendeljük azok szófaji elemzését. A felügyelt szófaji egyértelműsítési módszerek nagyméretű, kézzel annotált adatbázisokra épülnek. Az annotált adatbázis létrehozásához azonban szükséges egy, az adott nyelvre kidolgozott morfológiai kódrendszer is, melynek segítségével morfológiailag elemezni és egyértelműsíteni lehet az adott nyelvű szövegeket. Bizonyos nyelvekre azonban nem áll rendelkezésre ilyen kódrendszer és/vagy nagyméretű annotált adatbázis. Ez esetekben a megoldást a félig felügyelt vagy felügyelet nélküli szófaji egyértelműsítési módszerek jelenthetik, melyek segítségével az ilyen nyelvekre is lehetséges hatékony szófaji egyértelműsítőt építeni.

A felügyelt szófaji egyértelműsítési módszerek a szövegszavakat előre meghatározott (a tanító adatbázisban szereplő) szóosztályokba sorolják. Azonban előfordulhat, hogy egy nyelvre többféle annotációs rendszer is létezik, más-más mennyiségű elérhető annotált adattal, ami megnehezíti a különféle szófaji egyértelműsítő módszerek hatékonyságának összevetését. Például a hunpos tagger [1]

a KR morfológiai kódrendszerre épül, ám jelenleg nem tudunk olyan kézzel annotált adatbázisról, amely a KR-kódokat használná. Így a hunpos hatékonyságát csak úgy lehetséges mérni, ha a KR-kódokat megfeleltetjük egy kézzel annotált korpuszban szereplő kódoknak, ami szintén idő- és munkaigényes feladat.

A felügyelet nélküli szófaji egyértelműsítő módszerek különféle csoportokba (klaszterekbe) sorolják a szavakat, így képesek kiküszöbölni a fenti hátrányokat, mivel a klaszterek összevethetők bármely morfológiai kódrendszer által alkalmazott csoportokkal. A módszer tovább előnye, hogy a szófaji egyértelműsítés részletességét különböző technikákkal lehetséges szabályozni. Míg egyes kódrendszerek túlságosan részletes kódokat tartalmaznak (például képzéssel kapcsolatos információkat), addig a legtöbb alkalmazás számára nem szükséges a kódok ilyen mértékű részletezése: a fő szófaj megadása általában elégségesnek bizonyul a legtöbb alkalmazás számára (például információ-visszakeresés, névelemfelismerés vagy kulcsszókinyerés). Ezzel szemben más esetekben fontos lehet a minél részletesebb morfológiai információ, például a gépi fordításban vagy a szemantikai szerepek meghatározásában a főnévi esetragok igen nagy szereppel bírnak. A szükséges részletességet a klaszterek mennyiségének befolyásolásával tudjuk biztosítani. Az aktuális feladat számára indokolt klaszterszám befolyásolására a T-MAN [2] hálózati topológiaépítő pletykaalgoritmus számára bemenetként adott gráf eltérő módokon történő felépítésével nyílik lehetőség.

Az általunk használt közösségkereső eljárás [3] a szóalakok kontextuális tulajdonságaiból épített hálózat particionálásával állítja elő az egyes lexikai csoportokat. A gráfelméleti alapokon nyugvó algoritmus a particionálandó gráfok legjobb modularitással járó felbontására ad kielégítő és gyors közelítést. Az eljárás egy további tulajdonsága, hogy mivel a különböző particionálásokat jellemző modularitás mérőszámának több lépésben végrehajtott maximalizálásával történik, így lehetőség van hierarchikus közösségek kialakítására, amelyek a felhasználási területtől függően eltérő hasznossággal bírhatnak, hiszen a szóalakok durvább és részletezettebb lexikai csoportokba sorolása is lehetséges.

Eredményeink azt igazolják, hogy megközelítésünk felveszi a versenyt az angolra alkalmazott felügyelet nélküli módszerekkel, mindemellett a módszer magyarra való alkalmazhatóságát is számszerűsítettük.

2. Kapcsolódó munkák

A felügyelet nélküli és félig felügyelt szófaji egyértelműsítés területén már számos korábbi munka született az utóbbi évtizedekben, melyek több csoportba sorolhatók. Az egyik megközelítés szerint a kívánt szófaji klaszterek számát előre meg kell adni [4,5], ugyanakkor más rendszerek a klaszterek számát az adott feladathoz igazítva határozzák meg. Míg egyes módszerek rejtett Markov-modellekre épülő felügyelet nélküli tanulásként tekintenek a problémára [6,7], addig mások magasabb dimenziós terekben végeznek számításokat, illetve megint mások gráfként közelítenek a problémához. Továbbá, bizonyos módszerek működéséhez szükség van egy előre megadott részleges szótárra vagy néhány mintapéldára is, azonban ezek nem minden esetben állnak rendelkezésre.

Számos kiértékelési metrika használatos a szakirodalomban, melyek gyakran a több szófaji klasztert előállító módszereket részesítik előnyben. A legtöbb szerző azonban az információelméletből kölcsönzött V-mérték mellett teszi le a voksát [8]. A felügyelet nélküli szófaji egyértelműsítő módszerek kiértékelése megfeleltetés alapján is történhet, amikor is a rendszer teljesítményét a létrejott klaszterek (vagy ezek egy részhalmaza) és az etalon klaszterek közti megfeleltethetőség alapján határozzák meg. A kiértékelési metrikákról [9] ír bővebben.

A hálózatelemzés kulcsfontosságú szereppel bír a felügyelet nélküli megközelítésekben, ahol a magasabb dimenziós terekben történő klaszterezés helyett gráfalapon hajtódik végre a művelet, figyelmen kívül hagyva a dimenzionalitást. A hálózatelemzési módszerek közül különösen a közösségkeresés kapott nagy figyelmet több tudományterületen is a biológiától kezdve a szociológián át az informatikáig. A gráfok particionálása kapcsán a modularitás vált meghatározó fogalommá a korábbi metrikák közül [10]. A modularitás eredetileg a gráf particionálásának hatékonyságát hivatott mérni, és később számos gráfparticionáló algoritmus – mint például a spektrális optimalizáció, mohó algoritmusok és szimulált hűtés – célfüggvényévé vált.

3. Módszertan

A közösségkereső eljárásra épülő szófaji egyértelműsítés az eltérő szóalakok fölött értelmezett hasonlósági gráf particionálásán alapul, amely hasonlósági gráf építésének és jellemző csoportokra bontásának részletes bemutatására a következőkben kerül sor.

3.1. Hasonlósági gráf

Mivel a hasonló kontextusban szereplő szóalakokról feltételezhető, hogy hasonló mondatbéli funkcióval is bírnak [11], ezért eljárásunkban a szóalakok szófaji kategóriáinak felügyelet nélküli meghatározására egy olyan eljárást valósítottunk meg, mely a szóalakok fölött értelmezett hasonlósági gráf particionálásán alapul. Algoritmusunk a szóalakokat a hozzájuk meghatározott kontextusvektorok alapján sorolja be a hasonló szerepet betöltő és általunk azonos szófajúként interpretált szavak halmazaiba. Első lépésként tehát a szóalakok fölött értelmezett, súlyozott hasonlósági gráfunkat definiáljuk.

Munkánk során a szófajuk szempontjából csoportosítandó szavak alkották azt a V szótárat, amely elemeit eltérő méretű ($1 \leq W \leq 3$) ablakok mellett vett szókörnyezet-eloszlásokkal jellemeztük. (Mind a csoportosítandó szóalakok meghatározása során, mind pedig a környezetük vizsgálata során egy egyszerű reguláris kifejezés segítségével a numerikus kifejezéseket egységesen kezeltük.) A különböző méretű és nyelvű korpuszok feldolgozása során egy-egy szóalakot, a bal és jobb oldalukon, eltérő $w \leq W$ pozíciókon számított $2*(|V|+1)*W$ méretű eloszlásvektorral jellemeztünk. A későbbiekben particionálandó hasonlósági gráf csúcsait a $|V|$ méretű szótár egy-egy eleme képezte, a csúcsok közötti élsúlyok

meghatározásában pedig a szóalakokhoz társított eloszlásvektorok játszottak szerepet.

A gráfalapú megközelítések előnye többek között az, hogy a kiugró értékek (outliers) kezelése viszonylag természetes módon kezelhető szemben például a k-közép klaszterezéssel. A nem releváns és így nem kívánt hasonlóságok kiszűrésének egy lehetséges módja a teljes gráfokról a k-legközelebbi gráfokra való áttérés lehet. Azon túl, hogy a gráfban csökkenthető a zajt okozó kapcsolatok száma, a gráf ritkításával egyúttal jótékonyan befolyásolható a gráfon végzett algoritmusok sebessége.

Éppen ezért a szóalakok egymáshoz való viszonyának reprezentálása során a teljes gráfokból $G_k = (V, E_k, w)$ k-legközelebbi szomszédságon alapuló gráfokat konstruáltunk, melyekre $E_k = \{(u, v) : n(u, k) \ni v \vee n(v, k) \ni u\}$, ahol az $n(u, k)$ és $n(v, k)$ függvények rendre az u és v csúcsokhoz tartozó k legközelebbi szomszédot adják vissza, $w(u, v)$ pedig az u és v csúcsok közötti szimmetrikus távolságot határozza meg. A csúcsok közötti távolságot a *koszinusz távolság* (1), *Jensen-Shannon divergencia* (2), illetve *Jaccard-együttható* (3) segítségével is vizsgáltuk, melyek kiszámítása a következő képletek alapján történt:

$$cos(q, r) = 1 - \frac{\sum_v q(v)r(v)}{\sqrt{\sum_v q(v)^2}\sqrt{\sum_v r(v)^2}} \tag{1}$$

$$JS(q, r) = \frac{1}{2}[D(q\|avg_{q,r}) + D(r\|avg_{q,r})] \tag{2}$$

$$jacc(q, r) = 1 - \frac{|\{v : q(v) > 0 \wedge r(v) > 0\}|}{|\{v|q(v) > 0 \vee r(v) > 0\}|} \tag{3}$$

Az előzőekben bemutatott metrikák valamelyikével a csúcsokhoz történő k legközelebbi szomszéd meghatározását követően az eddig távolságokként értelmezhető élsúlyokat hasonlósági értékekké alakítottuk át. A hasonlósági mértékre való áttérés érdekében minden (u, v) csúcs közötti súlyt a $sim(f(u, v)) = \frac{1}{1+f(u,v)}$ képletnek megfelelően alakítottuk át, ahol $f(u, v)$ az előzőekben definiált távolságfüggvények értéke u és v csúcsokra nézve. A távolság helyett a hasonlósági értékekre való áttérésnek a közösségkereső eljárás súlyozott gráfon értelmezett működése kapcsán volt fontos.

3.2. Modularitásalapú közösségkeresés

Az általunk használt, modularitás maximalizálására építő eljárás előnye, hogy a kialakuló közösségek száma a particionálandó gráf topológiája alapján kerül meghatározásra, szemben egyéb eljárásokkal (pl. k-közép klaszterezés). Egy adott gráfparticionálást jellemző modularitás kiszámításával egy jósági értéket rendelhetünk a felbontás minőségére nézve, mely figyelembe veszi a gráf topológiájából adódóan az egyes csúcspárok között elvárható élek számát, valamint egy tényleges felbontás során az egyes csoportokon belül vezető élek tapasztalt számát. Az

előzőekben elmondottak a következő képlettel számolhatók:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j) \tag{4}$$

, amelyben az összegzés minden *lehetséges* élre (minden i és j csúcsra) vonatkozik, és ahol az A_{ij} a particionálandó gráf szomszédsági mátrixának egy eleme, m a gráfban található élek száma, az összegzésben található hányados az i és j csúcsok között menő élek várható értéke, a δ függvény pedig az ún. Kronecker-delta, mely akkor veszi fel az 1 értéket, ha az i és a j csúcsok megegyező klaszterben találhatók, máskülönben 0.

Számos jó tulajdonsága miatt vonzó elgondolás lenne a gráfokhoz olyan felbontásokat keresni, amelyek a modularitás jósági mérőszámát tekintenék célfüggvényül, azt maximalizálnák. Ugyanakkor ahogy arra már rámutattak [12], ez a feladat erősen $\mathcal{NP}$-teljes. A negatív eredményből adódóan, számos közelítő eljárás látott napvilágot a probléma kezelhető időben történő minél hatékonyabb megoldására, melyek között találunk szimulált hűtéstől kezdődően spektrálmódszereken át mohó megközelítéseket is.

Ugyan a spektrálmódszereken alapuló eljárások gyakorta jobb eredményeket érnek el más megközelítésekhez képest, nagyméretű gráfok esetében sokszor nem hatékonyak, és mivel esetünkben kifejezetten nagy gráfok felbontását kíséreltük meg, így kiemelten fontos volt, hogy a maximális modularitást eredményező felbontás közelítésére alkalmazott eljárásunk számítási igénye alacsony legyen. A [3] által alkalmazott mohó optimalizáló stratégia kifejezetten nagy gráfokon is működőképesnek bizonyult, így az általuk javasolt eljárást valósítottuk meg a szóalakok gráfjának maximális modularitást elérő felosztásának meghatározására. A szerzők által javasolt eljárás egy alulról felfelé építkező klaszterező eljárás, mely kezdetén minden csúcsot egy külön klaszterbe sorolnak, majd a további lépések során a csúcsok meglátogatása során azokat a lokálisan legjobb modularitás növekményt eredményező közösséghez sorolják (esetleg egyikhez sem). Egy i csúcs C közösségbe történő mozgatása során kettős hatás figyelhető meg: egyrészt növeli a globális modularitás értékét azon élei által, amelyek immáron a C közösségbeli szomszédjaival való összeköttetést biztosítják, másrészről viszont a modularitás bizonyos mértékű csökkenése is megfigyelhető lesz azon élei kapcsán, amelyek a korábbi közösségének tagjaival való összeköttetésért voltak felelősek. Egy i csúcs C közösségbe történő átmozgatásának hatása a következők szerint összegezhető:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \tag{5}$$

, ahol $\sum_{in}$ és $\sum_{tot}$ értékek rendre a C közösségen belül, illetve a C közösséget érintő élek súlyainak összege, k_i és $k_{i,in}$ pedig rendre az i csúcsot tartalmazó, illetve az i csúcsot a C közösséggel összekötő élek súlyainak összege, m pedig a particionálandó gráfban található élek összsúlya. Miután minden csúcs besorolást

nyert az egyes közösségekbe, az algoritmus a kialakult közösségeket összevonva, és azokat egy csúcsként kezelve megismétli az előző eljárást. Egy soron következő iterációs blokk kezdetén tehát éppen annyi csúcsot tartalmazó gráfot bontunk ismét közösségekre, amennyit az előző blokkban azonosítottunk (a korábbi blokk közösségeinek megfeleltethető élsúlyok pedig a megelőző lépésben a két közösség közt menő élek összsúlyával lesz egyenlő, a közösségen belüli élek pedig hurokélként jelentkeznek.) Az iterációs blokkokat ismételhetjük fix lépésszámig, vagy addig, amíg a modularitás növekedése fenntartható. Az eljárás előnye, hogy az eredeti hasonlósági gráf csúcsai fokszámának várható értékének fix voltából adódóan az eljáráshoz elvégzendő műveletek száma nagyságrendileg a hasonlósági gráf csúcsainak lineáris függvénye lesz. További előny, hogy az iterációs blokkok mentén eltérő finomságú – de ugyanúgy a modularitás maximalizálására törekvő – felbontásait nyerhetjük ki a particionálandó gráfnak.

3.3. A legközelebbi szomszéd gráf pletykaalgoritmussal történő közelítése

Más felügyelet nélküli módszerhez hasonlóan az általunk javasolt eljárás is nagy elemszámú minta alapján próbálja a szóalakok közt fennálló szabályszerűségeket megragadni, ami azzal jár, hogy a szótár méretének növekedésével együtt a hasonlósági gráf csúcsainak száma több százezres nagyságrendben is mozoghat, ami pedig – nagyobb W kontextusablak választása esetén – akár az egyes szóalakokat leíró szókörnyezeteloszlás-vektorok milliós hosszát is eredményezheti. Jóllehet a szókörnyezeteloszlás-vektorok jellemzően igen ritkák, egy adott esetben több százezer csúcsot tartalmazó hasonlósági gráfra még így sem határozható meg igazán hatékonyan minden szögponthoz annak k legközelebbi szomszédja.

A szótárméret növekedésével együtt jelentkező hatékonysági probléma megoldására a T-Man [2] pletykaalapú peer-to-peer protokollt hívtuk segítségül, melynek eredeti célja speciális, dinamikusan változó, nagyméretű ún. overlay hálózatok topológiájának feltérképezése. Az overlay hálózatok dinamikusságából adódóan az algoritmus a hálózati topológia egy közelítését határozza csupán meg, amire esetünkben a szóalakok hasonlósági gráfjának statikusságából adódóan ugyan nem lenne szükség, ugyanakkor a szótár méretének növekedéséből adódó problémákra megoldást nyújthat sebességével. A protokoll a következők szerint jár el: minden csúcs (peer) inicializálásra kerül egy fix méretű random szomszédos csúcsokat (peereket) tartalmazó bufferrel, majd az egyes iterációk során a csúcsok (peerek) 'kommunikálnak' egymással, amely során lehetőségük nyílik a hozzájuk tartozó bufferek tartalmának frissítésére, amennyiben azzal javítani tudnak annak tartalmán. (Esetünkben az overlay hálózatok azon speciális tulajdonságával, hogy a csúcsok folyamatosan be,- illetve kiléphetnek a hálózatból, nem kellett számoljunk.)

A szerzők algoritmusuk gyors konvergenciájáról számoltak be, vizsgálataik alapján 10-15 iteráció elégségesnek bizonyult az eredeti hálózatok topológiájának közel tökéletes közelítésére. A szóalakok fölötti hasonlósági gráf k-legközelebbi szomszédságának feltérképezése kapcsán tapasztalható konvergenciával kapcsolatos eredményeinket a 4. fejezet tartalmazza.

4. Eredmények

Az előzőekben bemutatottak szerint működő közösségkeresésen alapuló szófaji egyértelműsítőt – annak felügyelet nélküli voltából adódóan – módosítások nélkül alkalmazhattuk magyar, illetőleg angol nyelvű szövegekre. Angol nyelvű vizsgálódásaink tárgyát az ACL/DCI korpuszban található Wall Street Journal 1987. évadának 1-5. fejezetei képezték, a magyar nyelvű szövegek esetében pedig – hasonló stílusú és nyelvhasználatú korpuszt keresvén – a Magyar Nemzeti Szövegtár Heti Világgazdaságot érintő részeit vizsgáltuk. Kísérleteink kitértek a szóalakok hasonlóságának meghatározásának különféle paraméterek melletti vizsgálatára: a kontextusablak mérete, akárcsak a hasonlósági gráf esetében a k legközelebbi szomszédság értékei 1 és 3 között mozogtak, továbbá megvizsgáltuk azt is, miképp befolyásolja a szóalakok csoportosításának eredményességet, ha eltérő nagyságrendű szöveg alapján hajtjuk végre mindazt. A két nyelvre elkészített eltérő nagyságrendű korpuszokkal kapcsolatos statsztikákat a 1. táblázat tartalmazza. (Mivel a Magyar Nemzeti Szövegtár esetében nem állt rendelkezésre az az információ, hogy egy szóalakra nézve melyek a szóba jöhető szófaji kódok, így ott a szóalakonkénti átlagos szófajszámot/többértelműséget nem állt módunkban kiszámolni.)

1. táblázat. Az angol és magyar nyelvű korpuszok statisztikái.

	WSJ		MNSZ	
	Szint_1	Szint_2	Szint_1	Szint_2
Mondatok száma	7053	34486	6069	30524
Tokenek száma	145002	723415	145006	723416
Szóalakok száma	13750	31686	36224	110133
Átlagos tokengyakoriság	10,55	22,83	4,00	6,57
Szóalakonkénti átlagos szófaj	$2.26 \pm 1{,}38$		-	

A nagyobb gráfok (Szint_2) esetében megvizsgáltuk a T-Man hálózatitopológia-közelítő algoritmus konvergenciájának sebességét az iterációk tükrében, ami az 1. ábrán látható. Az egyes iterációkhoz tartozó szaggatott vonalak alapján leolvasható, hogy átlagosan hány százalékkal haladta meg a közelített gráfokban szereplő élek összsúlya az etalon k-legközelebbi gráfok alapján elvárható összsúlyokat. A folytonos vonalak mentén az látható, hogy az egyes iterációk után a gráf csúcsaihoz választott legközelebbi szomszédok mekkora hányada volt megtalálható a tényleges – de csak jóval több számítás árán megkapható – k-legközelebbi szomszédságban szereplő élekhez képest. A körrel jelzett értékek a magyarra, a csillaggal jelzettek pedig az angol eredményekre vonatkoznak.

A felügyelet nélküli szófaji kódolás hatékonyságát jellemzően a kialakult klaszterek tényleges szófaji csoportokhoz való hozzárendelhetősége, valamint információelméleti szempontok szerint szokás vizsgálni. Eredményeink a megszo-

kott **V1-mérték**, illetve *'egy-az-egyhez'* (**1-1**) és *több-az-egyhez'* (**t-1**) értékek szerint kerülnek közlésre.

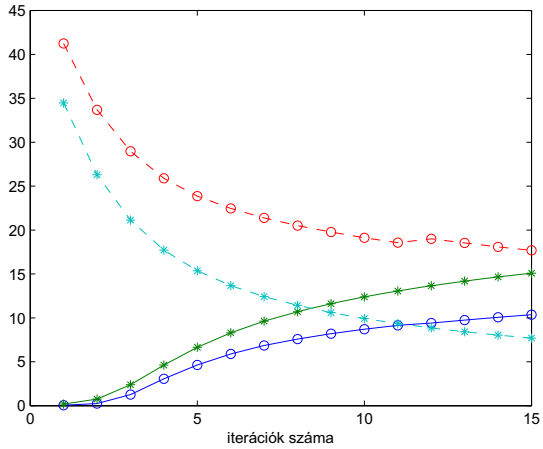
2. táblázat. A három fő paraméter (távolságszámítás módja, figyelembe veendő legközelebbi szomszédok száma, kontextusablak mérete) közül pontosan egy lefixálása mellett elért átlagos eredmények az eltérő méretű és nyelvű szövegeken.

	MNSZ						WSJ					
	$Szint_1$			$Szint_2$			$Szint_1$			$Szint_2$		
	V1	1-1	t-1	V1	1-1	t-1	V1	1-1	t-1	V1	1-1	t-1
COS	0.3336	0.2646	0.3929	0.3493	0.2793	0.4266	0.4466	0.3054	0.5501	0.4711	0.3150	0.5907
JS	0.3096	0.2260	0.3581	0.3345	0.2415	0.3800	0.4011	0.3034	0.4681	0.4631	0.3425	0.5343
JACC	0.2558	0.1880	0.2924	0.2799	0.2049	0.3142	0.3184	0.2446	0.3993	0.3204	0.2323	0.3960
k=1	0.4138	0.2510	0.4715	0.4322	0.2569	0.5212	0.4747	0.3115	0.6283	0.4932	0.3053	0.6803
k=2	0.2474	0.2164	0.2943	0.2726	0.2295	0.3013	0.3385	0.2640	0.3950	0.3875	0.3025	0.4339
k=3	0.2378	0.2111	0.2777	0.2589	0.2393	0.2982	0.3529	0.2778	0.3942	0.3740	0.2819	0.4068
w=1	0.3270	0.2316	0.3768	0.3281	0.2308	0.3838	0.3894	0.2702	0.4506	0.4258	0.2857	0.5137
w=2	0.2956	0.2342	0.3475	0.3275	0.2531	0.3820	0.3860	0.2964	0.4531	0.4380	0.3341	0.5317
w=3	0.2764	0.2127	0.3191	0.3083	0.2417	0.3549	0.3111	0.2498	0.3887	0.3909	0.26700	0.4755

3. táblázat. A nagyobb mennyiségű szövegekből készített k-legközelebbi szomszédsági gráf közelítő meghatározása segítségével elért átlagos eredmények pontosan egy paraméter lefixálása mellett.

	MNSZ			WSJ		
	V1	1-1	t-1	V1	1-1	t-1
COSINE'	0.3167	0.2645	0.3896	0.4724	0.3364	0.5859
JS'	0.2562	0.2052	0.3083	0.4029	0.2924	0.4720
JACC'	0.2135	0.1756	0.2665	0.2662	0.2090	0.3575
k'=1	0,3923	0,2494	0,4770	0,485	0,3073	0,6532
k'=2	0,2049	0,2009	0,2512	0,3399	0,2775	0,3946
k'=3	0,1883	0,1950	0,2363	0,3167	0,2530	0,3675
w'=1	0,2645	0,2087	0,3264	0,3649	0,2593	0,4632
w'=2	0,2645	0,2226	0,3248	0,4009	0,3038	0,4916
w'=3	0,2564	0,2140	0,3132	0,3758	0,2747	0,4605

A 'több-az-egyhez' kiértékelés olyan megengedő értéket határoz meg a szóalakok csoportosításához, amely a megtalált közösségeket olyan módon rendeli az etalon szófaji címkék által alkotott szóalakok csoportjaihoz, hogy a pontosság maximalizálva legyen. Ezzel szemben az 'egy-az-egyhez' kiértékelés megköveteli azt a feltételt, hogy a megtalált csoportok hozzárendelése az etalon csoportokhoz kizárólag olyan módon történhet, hogy egy etalon csoporthoz egy közösséget rendelhetünk. Jelen eredmények az 'egy-az-egyhez' hozzárendelés mohó módon



1. ábra. A k-szomszédsági gráfok pletykaalgoritmussal történő közelítésének konvergenciája a végrehajtott iterációk számának függvényében.

történő meghatározása mellett értendők (amely nem feltétlen egyezik meg a globálisan legjobb hozzárendelés értékével). Természetesen ez utóbbi kiértékelés jobban bünteti azokat a felbontásokat, amelyek az etalon szerint elvártnál jóval nagyobb számú csoportot eredményeznek.

Az információelméleti alapokon nyugvó V1-mérték [8] az egy klaszterezéshez tartozó *homogenitás* és *teljesség* értékekből számított súlyozott harmonikus átlagaként áll elő, hasonlóan az osztályozások jóságát jellemző F-mértékhez, ami a pontosság és a fedés értékeket ötvözi. A homogenitás feltételes entrópiát használva számszerűsíti, hogy a kialakuló egyes csoportok mennyire diverzek az etalon csoportokhoz képest. A teljesség számítása analóg módon történik, a különbség mindössze annyi, hogy ennek esetében az etalon címkék diverzitása kerül számszerűsítésre a megtalált klaszterek fényében. Egy tökéletes klaszterezés esetében az összes egy etalon csoportba tartozó elemet ugyanabban a megtalált klaszterben kell találjunk. Hasonlóan az F-mérték általánosításához, a V-mérték esetében is lehetőség nyílik annak két összetevőjének egymáshoz mért fontossága alapján meghatározni – $\beta = 1$ választástól különböző módokon is akár – egyéb V_β értékeket.

5. Diszkusszió

A hasonlósági gráfok segítségével leghatékonyabban a főnevek, igék, segédigék és számnevek csoportjait sikerült azonosítani: minden általunk használt módszer elfogadható mértékben azonosította őket. Ez különösen igaz a hónapnevekre és a különféle cégformák rövidített alakjaira (például *Co.* vagy *Ltd.*), hiszen ezekben az esetekben szemantikailag hasonló szavak kerültek egy csoportba. A

fenti szófajokkal szemben a legkeményebb diónak a határozószavak bizonyultak. A határozószavak elég vegyes csoportot alkotnak (morfológiai jegyekkel és mondatbeli pozícióval kevésbé megfoghatók), így megfelelő osztályba sorolásuk nehézséget jelentett mindegyik módszer számára. Érdekes módon a k legközelebbi szomszéd és a Jaccard-módszer is azonos gráfba helyezte az elöljárókat, névelőket és kötőszavakat, aminek az lehet a magyarázata, hogy hasonló környezetben fordulnak elő (például gyakran főnévi előtti pozícióban). Megjegyezzük ugyanakkor, hogy e szófajok elkülönítése problémásnak nevezhető az angol nyelvben [13]. A szomszédok számának meghatározásával és az ablakméretek rögzítésével kapcsolatban ugyanakkor azt találtuk, hogy a kisebb értékek bizonyultak hatásosabbnak, tehát elsődlegesen a szavak szűk környezete befolyásolta a csoportokba sorolást.

Az egyes módszerek összevetését tekintve a Jaccard-módszer bizonyult leghatékonyabbnak az *-ing*-es alakok (gerund) azonosításában. A k legközelebbi szomszéd módszer a melléknevek felismerésében nyújtott kitűnő eredményt, továbbá hatékonynak bizonyult az igeként és főnévként egyaránt szereplő szóalakok csoportosításában (pl. *decrease*). Szintén e módszer remekelt a névelemek osztályba sorolásában, különösen az ország- és nemzetiségnevek besorolása bizonyult sikeresnek. Ez arra utalhat, hogy e módszer a felügyelet nélküli szófaji egyértelműsítés mellett felügyelet nélküli szemantikai osztályozásra is feltehetőleg jól használható.

A közösségkereső eljárás során elnagyoltabb és részletesebb lexikai csoportok is létrejöttek. Angol nyelvre az elnagyoltabb csoportosítás esetében sikeresnek bizonyult a névmások, többes számú főnevek, tulajdonnevek és melléknevek kezelése, ugyanakkor az igei és főnévi szerepet egyaránt betölthető szóalakok is egy osztályba kerültek. Ugyanez mondható el az elöljárószavakra és határozószavakra is. Az angol nyelvű finomabb osztályozás során a szófaji osztályozáson túl szemantikai csoportok is megjelentek (például egy közösséget alkot a *TV*, *video*, *radio* szócsoport), de a helynevek osztályozása is jónak mondható. Mindemellett külön csoportokba kerültek az előbb még egy osztályba sorolt prepozíciók és névelők, determinánsok.

Magyar nyelvű kísérleteinkben a főnevek, számnevek és segédigék azonosítása volt a legeredményesebb, az igék és névutók felismerése valamivel nehezebb feladatnak bizonyult. Az angolhoz hasonlóan a funkciószavak (kötőszavak, névmások, névelők, határozószavak) itt is egy osztályba kerültek mindegyik módszer alkalmazásakor. Mindezt szintén a hasonló mondatbeli pozíció magyarázhatja: a vonatkozó névmások például a kötőszavakhoz hasonló viselkedést mutatnak. Módszereinket összehasonlítva azt találjuk, hogy a névelemek azonosításában a Jaccard-módszer felülmúlja a másik kettőt, különösen igaz ez a politikai pártokra és a személynevekre, vagyis itt is képes szemantikai alapú névelemcsoportok létrehozására.

A közösségkereső eljárás által létrehozott csoportok a magyarban kevésbé bizonyultak jónak, mint az angolban. Noha itt is megfigyelhetünk szemantikai alapú csoportosítást (hét napjai, hónapok) a részletesebb osztályozásban, általánosságban a számnevek felismerése érte el a legjobb eredményt. Érdekes

módon a főnevek és melléknevek gyakran kerültek egy csoportba, amit valószínűleg az magyarázhat, hogy a magyarban mindkét szóosztály hasonló toldalékokat vehet fel (többes szám jele, birtokos jel, esetragok).

Ha összevetjük az angolra és magyarra kapott eredményeinket, azt láthatjuk, hogy a felügyelet nélküli szófaji egyértelműsítés könnyebb feladat angolon, mint magyaron. Ezt természetesen a nyelvek közti eltérésekre vezethető vissza. Egyrészt az angolban nagyságrendekkel kevesebb szóalak tartozik egy lemmához, mint a magyarban (erre utal a lehetséges szófaji kódok száma is). Másrészt a magyarban jóval kisebb a többértelmű szóalakok (homonimák) száma, az angol ezzel szemben bővelkedik az ige/főnév/melléknév stb. szerepben egyaránt előforduló szavakban (pl. *present*). Mindebből az következik, hogy a magyarban több szóalak fordul elő, így ezek csoportosítása is nehezebb feladat. Harmadrészt az angol szórendje kötött, míg a magyar szórend a mondat információs szerkezetét tükrözi, ami azt jelenti, hogy az osztályozandó szó környezete sokkal változatosabb lehet, mint az angolban, vagyis nehezebb a kontextus felett általánosítani.

6. Összegzés

Ebben a munkában bemutattuk felügyelet nélküli szófaji egyértelműsítő módszerünket, mely közösségkeresésre épül. A szóalakok fölött értelmezett hasonlósági gráf költséges számítására való tekintettel az elosztott rendszerek területén az ún. overlay topológiák közelítésére korábban már sikeresen alkalmazott T-MAN algoritmust alkalmaztuk. Angol és magyar nyelvű eredményeink egyaránt azt igazolják, hogy sikerült átültetnünk a két különböző tudományos közösség által használt módszerek előnyeit a szófaji egyértelműsítés területére, azaz egy olyan feladatra nyújtottunk így megoldást, amelyet egy harmadik tudományos közösség tűzött ki céljául.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Halácsy, P., Kornai, A., Oravecz, C.: HunPos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, Association for Computational Linguistics (2007) 209–212
2. Jelasity, M., Montresor, A., Babaoglu, O.: T-man: Gossip-based fast overlay topology construction. Comput. Netw. **53** (2009) 2321–2339

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+
4. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. TextGraphs-1, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 73–80
5. Lamar, M., Maron, Y., Johnson, M., Bienenstock, E.: Svd and clustering for unsupervised pos tagging. In: Proceedings of the ACL 2010 Conference Short Papers. ACLShort '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 215–219
6. Gao, J., Johnson, M.: A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2008) 344–352
7. Van Gael, J., Vlachos, A., Ghahramani, Z.: The infinite HMM for unsupervised PoS tagging. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Association for Computational Linguistics (2009) 678–687
8. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 410–420
9. Christodoulopoulos, C., Goldwater, S., Steedman, M.: Two decades of unsupervised POS induction: How far have we come? In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, Association for Computational Linguistics (2010) 575–584
10. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2) (2004) 026113+
11. Biemann, C.: Unsupervised part-of-speech tagging employing efficient graph clustering. In: Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. COLING ACL '06, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 7–12
12. Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: Maximizing modularity is hard. (2006)
13. Santorini, B.: Part-of-speech tagging guidelines for the penn treebank project. Technical report, Department of Computer and Information Science, University of Pennsylvania (1990)

Szófaji kódok és névelemek együttes osztályozása

Móra György[1], Vincze Veronika[1], Zsibrita János[1]

[1] Szegedi Tudományegyetem,
Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék
6720 Szeged, Árpád tér 2.
{gymora, vinczev, zsibrita}@inf.u-szeged.hu

Kivonat: Jelen munkánkban egy, a szófaji kódok és a névelemek meghatározására szolgáló gépi tanulási modellt mutatunk be. Az általános véletlen mezőkön alapuló módszer segítségével több címkesorozat együttesen tanulható, valamint az osztályozás során a címkesorozatok legjobb kombinációját együttesen keressük. A `magyarlanc` szófaji elemző és az SZTENER névelem-felismerő jellemzőkészletét használva olyan rendszert építettünk, amely a címkék együttes osztályozásának segítségével felülmúlta a kiindulási rendszereket az általunk használt teszthalmazon. A névelem-felismerő F-mértékben mért teljesítménye 87,75-ről 89,87-re, a szófaji címkéző pontossága 97,11%-ról 97,99%-ra nőtt, úgy, hogy a kódok meghatározásának más minőségi tényezői is javultak.

1 Bevezetés

Szintaktikai szempontból a tulajdonnevek főnévként viselkednek: a *Láttad az **Interjú a vámpírral**-t?* mondatban a film címe ugyanúgy ragozható, mint bármely más magyar főnév (vö. *Láttad a filmet?*). Emiatt a tulajdonneveket gyakran a főnevek egyik alosztályának tekintik: bizonyos morfológiai kódrendszerek külön tulajdonnévi kódot tulajdonítanak nekik (például az MSD-kódrendszerben Np-s*, a PENN Treebankben pedig NNP az egyes számú tulajdonnevek kódja).

Azonban valójában nemcsak főnevek, hanem bármelyik szófajhoz tartozó elemek is lehetnek tulajdonnevek (vagy azok részei), például *Tesz-Vesz Kft*. A fenti kódrendszerek használatával a *Tesz-Vesz*-t is tulajdonnévnek kellene kódolni, ami azonban a kódok megsokszorozódásával jár, hiszen voltaképpen bármely szónak lehet tulajdonnévi kódja is. Ez egyrészt megnöveli a szófaji egyértelműsítés költségeit (sokkal több szó válik morfológiailag többértelművé), továbbá megkívánja azt is, hogy a morfológiai elemzőbe beépüljön egy tulajdonnév-felismerő rendszer. Úgy véljük azonban, hogy a tulajdonnév-felismerés nem a morfológiai elemző feladata, így az általunk alkalmazott megoldásban a két feladatot párhuzamosan hajtjuk végre. Megközelítésünkben a tulajdonnévi jelölés tehát nem a morfológiai kód része, hanem külön tulajdonnévi címkékkel látjuk el a tulajdonnév-felismerő által NE-nek ítélt elemeket, függetlenül attól, hogy milyen szófajú az adott elem.

Munkánkban megmutatjuk, hogy a szófaji címkézés és a névelem-felismerés teljesítménye kölcsönösen javítható a tanulás során a másik feladat által szolgáltatott jelö-

lésekkel. Hogy ez lehetővé váljon, olyan gépi tanuló megközelítést alkalmaztunk, amelynek segítségével a két probléma együtt, egy gépi tanulási faladatként kezelhető. Az általunk fejlesztett rendszer hatékonyan alkalmazható magyar nyelvű szövegek egyidejű szófaji címkézésére és a bennük található névelemek felismerésére, és a használt tanító és kiértékelő halmazokat figyelembe véve teljesítményében felülmúlja az eddigi különálló statisztikai alapú szófaji címkézőket, valamint névelem-felismerő rendszereket. A módszer könnyen adaptálható más nyelvekre is, amennyiben rendelkezésre áll az adott nyelven morfológiai elemző és megfelelő annotált szövegkorpusz, mivel nem alkalmaz nyelvspecifikus jellemzőket.

2 Morfológia és tulajdonnevek

A tulajdonnevek nyílt szóosztályt alkotnak, azaz nem alkotnak véges elemű halmazt, számuk állandóan bővül a nyelvben. Ez maga után vonja, hogy nem is sorolhatók fel maradéktalanul egy szótárban sem. A nyelvfeldolgozás számára azonban kiemelkedően fontos a tulajdonnevek megfelelő kezelése, így például a morfológiai elemzőkbe nagyméretű tulajdonnévszótárak épülnek be azok elemzésének megkönnyítésére. Azonban a fenti okok miatt egy morfológiai elemző sem ismerhet fel minden szóalakot, így az ismeretlen szavak (melyek nagy része tulajdonnév vagy annak származéka) kezelésére különféle, úgynevezett guessing módszereket érdemes kidolgozni [20].

A tulajdonneveket a nyelvészeti szakirodalom többnyire merev jelölőnek tekinti, mely konstans módon ugyanazt az egyedet azonosítja [7]. A fenti definícióban a „merevség” arra vonatkozik, hogy nem változik a jelölő és jelölt közti kapcsolat, azonban elgondolásunk szerint a „merevség” fogalma a tulajdonnevek morfológiájában is értelmezhető. A tulajdonnevek ugyan ragozhatók, sőt alkalmanként képzők is csatlakozhatnak hozzájuk (*New York – New York-i*), azonban a lemmájuk változatlan formában fordul elő a toldalék előtt (*Fodor – fodoros*). (A kisbetű-nagybetű változásoktól most eltekintünk.) Ez különösen akkor nyilvánvaló, amikor egy morfológiailag sajátos viselkedésű főnév fordul elő tulajdonnévi használatban. Vegyük az alábbi példákat.

Fodort Kovács, míg Bokort Szabó váltotta az elnöki székben.
Panni átugrotta a bokrot, és egy kiálló ág elszakította a szoknyája alján levő fodrot.

A *fodor* és *bokor* hangkivető főnevek, vagyis bizonyos toldalékok előtt kiesik a lemma utolsó magánhangzója. Ez a jelenség azonban nem figyelhető meg akkor, amikor személynévként használatos a két szó. E tulajdonság kihasználható a névelem-felismerésben: a morfológiai elemző a *fodrot* és *bokrot* alakokat várná *fodr+ot* és *bokr+ot* morfémákkal, ám a fenti szóalakokat csak a guesser segítségével lehet elemezni a beépített toldaléklista segítségével *fodor+t*, illetve *bokor+t* morfémákra való felbontással. Amennyiben az így kapott lemma megtalálható a morfológiai adatbázisban, viszont eltérést tapasztalunk az ott található és a guesser által adott elemzés között (vagyis jelen esetben a *fodor* és *bokor* tárgyesetű alakja nem *fodrot* és *bokrot*, hanem *fodort* és *bokort*), valószínűsíthetjük, hogy tulajdonnévről van szó.

Bizonyos tulajdonnévtípusok – műcímek, intézménynevek (különösen ha többtagúak) – gyakran tartalmaznak már eleve ragozott alakokat, például *Interjú a vámpírral*, *Bolyai Farkas Alapítvány a Magyarul Tanuló Tehetségekért*. Azonban ezek is ragozhatók:

Megnéztem az Interjú a vámpírral-t.
Köszönetet mondott a Bolyai Farkas Alapítvány a Magyarul Tanuló Tehetségekért-nek.

A helyesírási szabályok szerint ilyenkor kötőjellel kell kapcsolni az újabb toldalékot a tulajdonnévhez. Utóbbi sajátosság is kihasználható a névelem-felismerésben: a kötőjelet tartalmazó szóalakot a guesser segítségével elemezzük, majd az így kapott lemmát ismét elemezzük. Amennyiben a szóalak a második elemzés során is toldalékoltnak bizonyul, ismét valószínűsíthető, hogy tulajdonnévvel találkoztunk.

A gyakorlatban sokszor előfordul, hogy a toldalék nem kötőjellel kapcsolódik a tulajdonnévhez (akár a helyesírási szabályok ellenében). Ezekben az esetekben is a guesser nyújthat segítséget: a lehetséges végződéseket le kell vágni a szó végéről, majd a maradékot lemmaként visszaadni, és a toldaléknak megfelelő főnévi elemzést társítani a szóhoz (pl. *Agrobankhoz – Agrobank* illativusi esetű főnév).

A morfológiai elemző oldaláról nézve a vele párhuzamosan zajló tulajdonnévfelismerés abban segíthet, hogy a NER-rendszer által tulajdonnévnek minősített elemeket nem feltétlenül próbálja meg hagyományos módon elemezni, hanem egyből a beépített guessert hívja segítségül, ezzel gyorsítva a folyamatot.

3 Együttes címkézési módszerek

Hagyományosan a különböző szekvenciajelölési feladatokat (szófaji címkék, felszíni elemzés, névelemek) külön-külön gépi tanulási feladatként definiálják, és a szövegek feldolgozása során az elemzőket egymás után futtatják. Így azonban az egyes alrendszerek hibái összeadódnak, valamint csak a feldolgozási láncban hátrébb álló komponenseknek van lehetősége felhasználni az előtte állók címkéit jellemzőként.

3.1 A címketerek kombinálása

Több jelölési lépés egyesíthető a címkék kombinálásával is, de így kezelhetetlen mértékben megnőhet a címketér, illetve előfordulhat, hogy bizonyos címkekombinációk csak kevésszer fordulnak elő a tanuló adatok között, így felismerésük bizonytalan lesz. A feladatok ilyen jellegű kombinálásánál a közös jellemzőkészlet is problémát jelenthet, mert előfordulhat, hogy a különböző címkézési feladatok eltérő jellemzőkészlet mellett adnak optimális eredményt.

3.2 Gráfalapú valószínűségi modellek

Kísérleteinkben a szövegek párhuzamos címkézésére a MALLET GRMM [9][15] és a FactorIE [11] csomagban található általános feltételes véletlen mezők módszerét alkalmaztuk. A módszerek lehetővé teszik a hagyományos lineáris láncolású véletlen mezők módszeréhez képest, hogy tetszőleges valószínűségi függőségeket ábrázoló modelleket alkalmazzunk, így egy token akár egynél több címkével is rendelkezhet. A címkék közötti feltételes valószínűségi kapcsolatok modellezésével a névelem-felismerés és a szófaji címkézés egymástól független jellemzőkészlet segítségével valósítható meg, de olyan módon, hogy a szófajcímkék és a névelemcímkék együttes legjobb eloszlását tanuljuk, majd keressük a jelölés során. Természetesen a módszer kiterjeszthető más feladatokra, vagy akár kettőnél több egyidejű címkesorozat meghatározására is.

3.3 Előzetes vizsgálatok

Angol nyelvű szövegeken végzett kísérletek [10] azt mutatták, hogy a szófaji kódok és a felszíni elemzés címkéinek együttes gépi tanulásával jobb eredményt lehet elérni, mint ha ezeket a feladatokat külön tanított modellekkel egymás után szekvenciálisan végeznék el. Az általunk végzett ilyen irányú kísérletek azt mutatták, hogy a szófaji kódok meghatározásának pontossága 62,45%-ról 72,89%-ra, a felszíni elemzés pontossága pedig 83,95%-ról 85,76%-ra nőtt azonos jellemzőkészlet használata mellett, abban az esetben, ha a címkesorozatokat független osztályozása helyett azokat együttes osztályozással határozzuk meg. A két címkesorozat az osztályozás során így dinamikus jellemzőként hathat egymásra, kölcsönösen javítva a címkék meghatározásának pontosságát. A mérésekhez a CoNLL-2000 Shared Task tanító és kiértékelő halmazának ezer-ezer tokenes mintáját használtuk.

A CoNLL-2003 Shared Task [18] nyelvfüggetlen névelem-felismerési feladatán végzett kísérletek azt mutatták, hogy minimális jellemzőkészletet használva, mind a szófaji kódok címkézése, mind a névelemek felismerése javítható az együttes címkézés használatával. A verseny spanyol szövegeket tartalmazó részkorpuszából származó mintán elvégzett vizsgálatok azt mutatták, hogy míg a szófaji kódok címkézésének pontosságát csak mérsékelten 88,6%-ról 88,7%-ra, addig a névelem-felismerés F-mértékét jelentős mértékben, 39,5-ről 42,2-re növelte az együttes címkézés.

4 Névelem-felismerés

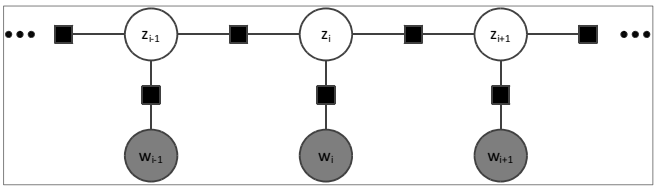
A névelem-felismerés alapvető fontosságú az információkinyerő rendszerek működése szempontjából. A felismert és különböző típusokba sorolt névelemek nem csak önmagukban érdekesek, de sok rendszerben a névelemek jelentik azokat az alapegységeket, amelyekből események épülnek fel, illetve amelyek között relációkat azonosítanak. A névelemek azonosításánál általában sokkal nagyobb kihívást jelent azok megfelelő osztályba sorolása. Az osztályozás általában környezeti jellemzők alapján lehetséges.

4.1 Kapcsolódó munkák

A névelemek felismerésének két alapvető módját különböztethetjük meg. A tokenalapú rendszerek szavankénti osztályozással döntik el, hogy az adott token része-e vagy sem egy névelemnek. Az osztályozó rendszerint szupportvektorgép [8], vagy maximum entrópia osztályozó [1][5]. Gyakran több akár különböző típusú tanulót is kombinálnak [13]. A névelem-felismerők másik, elterjedtebb csoportja a szekvenciatanulást alkalmazó módszerek. A Markov-mezőket [14] egyre inkább a feltételes véletlen mezők váltják fel a szekvenciajelölő rendszerekben. A CoNLL-2002 és a CoNLL-2003 névelem-felismerési feladatainak eredményei azt mutatták hogy a tokenenkénti osztályozást végző rendszereket többnyire felülmúlják a több token feletti címkeeloszlást tanuló megközelítések a névelem-felismerési feladatokban. [17][18]

Az általunk fejlesztett névelem-felismerő módszer az SZTENER [3] nyelvfüggetlen névelem-felismerő rendszer magyar nyelvre adaptált változatából indul ki. A szoftver a feltételes véletlen mezők módszerének MALLET [9] programcsomagban található verzióján alapszik. Elsőrendű láncolást alkalmaz, a jellemzők között ortografikus, szófrekvencia alapú, valamint szótár jellemzők találhatóak. A tanító és teszthalmaz mondataiból és szavaiból ennek a rendszernek a jellemzőkinyerő modulja segítségével készítettünk a gépi tanuló algoritmusok számára feldolgozható jellemzővektorokat.

4.2 A névelemfelismerő rendszer modellje



1. ábra: A névelemek felismeréséhez használt elsőrendű modell. A fehér körök a címkék rejtett változóit, a szürkék a jellemzők megfigyelhető változóit, a fekete négyzetek a változók közötti faktorokat jelölik.

A névelem-felismerő architektúráját megtartva a FactorIE feltételes valószínűségi programozási környezetben az [11] ábrán látható elsőrendű feltételes valószínűségi modellt definiáltunk. A modell a szó jellemzői $(w_0,w_1,...,w_n)$ és címkéi $(z_0,z_1,...,z_n)$, valamint az egymást követő címkék között definiál faktorokat. Az egyetlen különbség az eredeti és az általunk fejlesztett rendszer között, hogy a feltételes valószínűségek pontos kiszámítása helyett közelítő módszereket alkalmaztunk, ugyanis az együttes címkézési feladat során előálló bonyolult modell kiszámítása csak közelítő módszerekkel kivitelezhető elfogadható időn belül.

5 Szófaji kódok meghatározása

A szófaji kódok fontos szerepet töltenek be a szöveg további nyelvészeti elemzése során, illetve sok megközelítés közvetlenül jellemzőként is használja. A kódok hozzárendelése tokenalapon történik. Jelen munkában az MSD-kódrendszer egy egyszerűsített, gépi tanulási módszerekkel könnyebben kezelhető változatát használjuk.

5.1 Kapcsolódó munkák

Korábban több szófaji címkéző rendszer is készült a magyar nyelvre, mint például a szabály alapú RGLearn, illetve más, rejtett Markov-modellekre épülő statisztikai módszereket alkalmazó algoritmusok [4][6][12]. A szófaji címkézési feladat szerves része – különösen erősen agglutináló nyelvek esetében, mint például a magyar – a szavak morfológiai elemzése. A korábban említett magyar szófaji egyértelműsítők a HuMOR[1], illetve MetaMorpho[2] rendszereket, valamint a NooJ magyarra átültetett verzióját[3] alkalmazták.

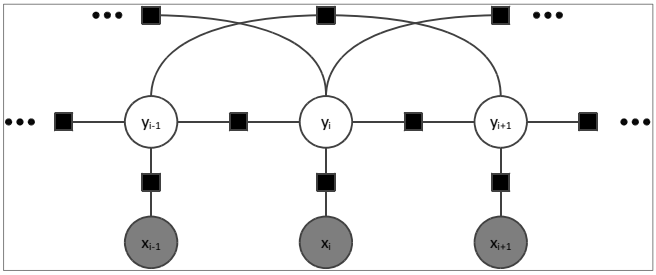
A szófaji címkéző jellemzőkészlete és felépítése a `magyarlanc` nevű [20], a Stanford POSTagger [19] módosításával létrehozott szófaji címkézőn alapszik, amely körkörös függőségű véletlen mezőket alkalmazó maximum entrópia osztályozót használ. A magyar nyelvre kifejlesztett jellemzőkészlet az 1-3 hosszú karakterprefixeket és suffixeket, a szavakat és azok szómintáját tartalmazza. Ezen kívül környezeti jellemzőként a szó előtte és utána álló szavakkal alkotott bigramjait, valamint a szavak és a környezetében található szavak szófaji címkéinek kombinációit használja. A szófaji kódok, illetve azok bi- és trigramjai a címkézés során dinamikusan állnak elő, a rendszer a lehetséges kombinációkat elemezve dönt a címkékről, így a módszer a tokenosztályozás és a szekvenciaosztályozási módszerek jegyeit is magán hordozza. Az adott szóhoz rendelhető szófaji kódokat a morfológiai elemző által megadott lehetséges kódok halmazából veszi a címkéző, ezzel is csökkentve a keresési teret [4].

5.2 A szófaji címkéző modellje

Mivel a szófaji címkéző ciklikus helyi függőségeket tartalmazó maximum entrópia osztályozót használó modellje egy az egyben nem ültethető át a FactorIE feltételes valószínűségi programozási környezetbe, a *2. ábrán* látható, az eredeti módszer ötleteit felhasználó másodrendű véletlen mezős modellt definiáltunk. A modell a névelem-felismerő szerkezetéhez hasonló, de a szó jellemzői $(x_0,x_1,...,x_n)$ és címkéi $(y_0,y_1,...,y_n)$, valamint az egymást követő címkék közötti faktorokon kívül a nem közvetlenül egymást követő címkék között is létrehoz feltételes kapcsolatokat. Ez azért

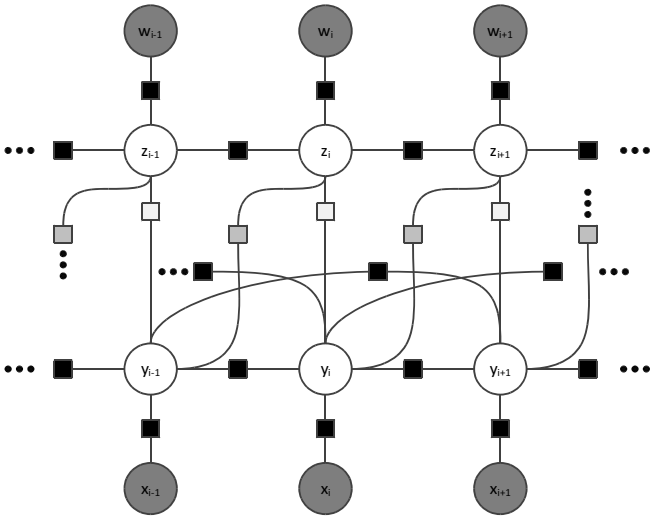
1 http://www.morphologic.hu/Morfologiai-elemzes.html
2 http://www.morphologic.hu/MetaMorpho-technologia/menuazonosito-256.html
3 http://corpus.nytud.hu/nooj/

fontos, mert a szófaji kódok erősen függenek nem csak az őket közvetlenül megelőző, hanem az azt megelőző címkétől is.



2. ábra: A szófaji címkéző által alkalmazott másodrendű modell.

A szavak felszíni jellemzői mellett a morfológiai elemző által megadott lehetséges szófaji kódok is külön vektorváltozóba kerültek. Az eredeti `magyarlanc`tól való eltérés, hogy a keresés nem korlátozódik csak ezekre a címkékre, emiatt számos esetben olyan címkéket is helyesen meghatározott, amiket a morfológiai elemző – hibásan – nem ajánlott fel.



3. ábra: A két különálló valószínűségi modell egyesítése. A világos és sötétszürke színnel jelölt faktorok a két címkesorozat közötti összefüggések leírására szolgálnak.

6 Névelemek és szófaji kódok együttes címkézése

A szófaji címkézés és a névelem-felismerés valószínűségi modelljeit a *3. ábrán* látható modellben egyesítettük. A két címkesorozat elemei között, valamint a névelem címkéjének változója és a megelőző szó szófaji kódjának változója között új faktorokat alkalmaztunk a modellek összekapcsolására. Ezen faktorok paraméterei lesznek azok, amelyek a tanulás után leírják a két címkesorozat közötti összefüggéseket.

7 Eredmények

Módszerünket a Szeged Korpusz üzleti híreket tartalmazó alkorpuszán értékeltük ki, melyben be vannak jelölve az etalon tulajdonnevek [2][16]. Az eredeti MSD-annotációban a tulajdonnevek Np-s* kóddal rendelkeztek, továbbá a többtagú tulajdonnevek össze voltak vonva. A kiértékelést megelőzően szétdaraboltuk a többtagú tulajdonneveket, és tagjaikat újraannotáltuk, a főnevek esetében pedig nem tettünk különbséget a köznévi és tulajdonnévi használat között (azaz a *köznév* és *tulajdonnév* kódokat felváltotta a *főnév* kód). Így tehát a *Magyar Nemzeti Bank* új kódja *A A N* lett. A magyar nyelven végzett kísérleteink azt mutatják, hogy – az angolhoz hasonlóan – eredményeink meghaladják a szekvenciálisan tanított modellek hatékonyságát.

A tanításhoz és a kiértékeléshez a rendelkezésre álló több mint 221 ezer tokent és 9400 mondatot tartalmazó korpuszt két részre osztottuk a mondatok véletlenszerű halmazba sorolásával. A tanító halmazba így a mondatok megközelítőleg 60%-a került, a maradékot kiértékelésre használtuk.

7.1 A névelem-felismerés kiértékelése

A jelen munkában szereplő névelem-felismerésre vonatkozó eredmények mind frázisalapú kiértékelésből származnak. Ez azt jelenti, hogy többszavas névelemek esetén csak az a jelölés számított helyesnek, ahol a névelem minden szava helyesen volt jelölve, és további szavak nem kerültek jelölésre. Az összehasonlíthatóság érdekében az összes rendszert ugyanazokon a halmazokon tanítottuk és értékeltük ki, azonos metrikákat alkalmazva. Ezt a frázisalapú F-mértéket alkalmazták a CoNLL-2003 névelem-felismerési feladat kiértékelése során is, az itt közölt eredmények azonos módszerrel lettek megállapítva.

A kiindulási rendszer teljesítménye mellett az általunk fejlesztett rendszerek eredményeit a tanuló algoritmus 2 és 5 iterációig tartó futtatása mellett is megadjuk mind a szófaji címkézéstől függetlenül tanított névelem-felismerő, mind az együttesen tanított és osztályozott névelem-felismerés esetében.

1. táblázat: Névelem-felismerés eredményei.

It.	Rendszer	Precízió	Fedés	$F_{\beta=1}$
	SZTENER névelem-felismerő	86,81	88,71	87,75
2	Független osztályozás	86,81	81,11	83,86
	Együttes osztályozás	88,57	89,27	88,93
5	Független osztályozás	84,73	81,60	83,13
	Együttes osztályozás	**89,71**	**90,04**	**89,87**

Az *1. táblázatban* található eredmények megerősítik, hogy a névelemek szófaji kódokkal való együttes osztályozása azonos jellemzőtér esetében jelentősen javítja a címkézés teljesítményét a függetlenül tanított modellhez képest. A független modell a kiindulási rendszernél is gyengébb teljesítményét 83,86-ról 88,93-ra növeli. A jellemzőtér ábrázolásának gyengeségét sejteti, hogy az eredetileg is gyengébb eredményt csak csökkenti a tanuló iterációs számának növelése, vélhetően túltanulja a jellemzőket. Ezt az információhiányt kompenzálhatja az együttes tanuláskor a szófaji kódok jelenléte.

7.2 A szófaji címkézés kiértékelése

A szófaji címkézést a csökkentett MSD szófaji kódok alapján tanítottuk és predikáltuk [20]. Ez az MSD-kódoknak egy szűkített készlete (42 kód), ahol csak azok a szófaji kódok vannak megkülönböztetve, ahol a szóalakból nem dönthető el egyértelműen a szó eredeti MSD-kódja. Erre a címketér csökkentése miatt van szükség, mert az eredeti több száz címkét tartalmazó kódrendszer gépi tanuló módszerekkel kezelhetetlen lett volna.

A csökkentett MSD-kódokat tovább redukálva csak a szófajt jelölő első karaktert megtarva is elvégeztük a szófaji címkézők kiértékelését, így láthatóvá vált, hogy a csökkentett MSD-kódokon szinte azonos eredményt elért rendszerek által hibásan jelölt MSD-kódok mennyire térnek el egymástól, azaz mennyire súlyos hibákat vét a két címkéző.

A szófaji címkézést a névelem-felismeréshez hasonlóan a kiindulási rendszerhez hasonlítottuk, és megmértük a csak szófaji címkézést végrehajtó modell és az együttes osztályozás közötti különbségeket is. A rendszerünket ebben az esetben is kettő, illetve öt iterációig tanítottuk.

A névelem-felismeréstől eltérően nem F-mértéket, hanem pontosságot alkalmaztunk a rendszerek teljesítményének elsődleges méréséhez. A pontosság mellett az egyes MSD/szófaji osztályokon elért F-mértékek átlagát (makroátlag, *1. képlet*) is megadtuk a rendszerekhez. Míg a pontosság a szöveg szavainak átlagos osztályozási pontosságát írja le, a makroátlag azt mutatja meg, hogy a ritkán előforduló címkék osztályait mennyire jól ismeri a rendszer. Ha ugyanis csak a gyakori szófajcímkéket osztályozza helyesen, akkor az osztályonkénti F-mértékek átlaga alacsony lesz a sok kis elemszámú, rosszul címkézett szófaji osztály miatt.

2. táblázat: Szófaji címkézés eredményei.

It.	Rendszer	Redukált MSD-kód Pontosság	$F_{\beta=1\ macro}$	Csak szófaj Pontosság	$F_{\beta=1\ macro}$
	magyarlanc	97,11	67,81	97,98	85,18
2	Független oszt.	97,75	71,03	98,60	84,12
	Együttes oszt.	97,78	72,48	98,68	86,32
5	Független oszt.	**98,00**	71,33	98,78	86,44
	Együttes oszt.	97,99	**73,32**	**98,81**	**88,77**

$$F_{\beta=1\ macro} = \frac{\sum F_{\beta=1}(c_i)}{|C|}, \forall c_i \in C \qquad (1)$$

A szófaji egyértelműsítés terén azt tapasztaltuk, hogy eredményeink javulása elsősorban a nagybetűvel kezdődő alakok helyes elemzésének köszönhető. Ez nem meglepő, hiszen a magyarban általában a tulajdonnevek és a mondatkezdő szavak kezdődnek nagybetűvel. A tulajdonnevek és szófaji kódok együttes jelölésével a mondatkezdő tulajdonneveket könnyebb volt azonosítani, így a „maradék" mondatkezdő elemek szófaját is nagyobb hatékonysággal lehetett megállapítani: például a *Szerinte* mondatkezdő elem főnévi kódot kapott a szekvenciális jelölésben, azonban az együttes jelölés során már a helyes határozószói kódot kapta.

Kiemelkedő javulást figyelhettünk meg a rövidítések esetében is. Noha ez a szóosztály kevés elemet tartalmaz, felismerésük 17,86%-kal javult, ami főleg a tulajdonnév részét képező *Jr*. és *Dr*. elő-, illetve utótagoknak pontosabb azonosításának volt köszönhető. Az indulatszavak kategóriájába lettek sorolva olyan tulajdonnevek is, amelyeket a morfológiai elemző – helytelenül – olyan összetételként értelmezett, amelynek utótagja indulatszó, például *Palotainé*. Ezek tulajdonnévként való felismerése javított a rendszer teljesítményén.

Összességében azt figyelhettük meg, hogy a rendszer különösen a ritkán előforduló szófajok felismerésében volt képes javulni, míg a nagyobb szóosztályok esetében minimális különbségeket vehettünk észre. Utóbbiak felismerési pontossága azonban már a szekvenciális modell esetében is kiemelkedő volt (97% feletti), így a tulajdonnevek hozzáadott értéke nem befolyásolta érdemben az eredményeket.

Az elhanyagolható pontosságbeli eltérés ellenére a jelölés minősége javult az együttes osztályozástól. A *2. táblázatban* található makroátlagok azt mutatják, hogy közel azonos pontosság mellett az együttesen tanított rendszer a kis elemszámú szófaji kódok osztályozásában jobb, ezzel összességében kiegyensúlyozottabb teljesítményt nyújt. A hibaelemzéshez alkalmazott, csak a szófajt figyelembe vevő kiértékelés pedig azt mutatja, hogy az együttesen tanított rendszer hibás címkézéskor több esetben rendel olyan szófaji kódot a szavakhoz, amelyek szófaja megegyezik a helyes szófajjal, azaz az elkövetett hibáinak kisebb hányada súlyos tévesztés, mint a függetlenül tanított szófaji kódcímkézőnek.

8 Konklúzió

Cikkünkben a szófaji kódok és a névelemek együttes címkézéséhez használható rendszert mutattunk be. Megmutattuk, hogy a hagyományos, szeparáltan tanuló módszerekhez képest mindkét címkézési feladat teljesítménye nőtt. Bár a szófaji címkézés esetében a változás nem olyan jelentős, de javultak az egyéb minőségi tulajdonságai.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Borthwick, A.: Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York University (1999)
2. Csendes D., Hatvani Cs., Alexin Z., Csirik J., Gyimóthy T., Prószéky G., Váradi T.: Kézzel annotált magyar nyelvi korpusz : a Szeged Korpusz. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szeged (2003) 238–247
3. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
4. Halácsy P., Kornai A., Oravecz Cs.: HunPos — an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
5. Chieu, H. L., Ng, H.T.: Named Entity Recognition with a Maximum Entropy Approach. In: Proceedings of CoNLL-2003 (2003)
6. Kuba A., Bakota T., Hócza A., Oravecz Cs.: A magyar nyelv néhány szófaji elemzőjének összevetése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 16–22
7. Kripke, S.: Naming and necessity. Blackwell, Oxford (1980)
8. Mayfield, J., McNamee, P., Piatko, C.: Named Entity Recognition using Hundreds of Thousands of Features. In: Proceedings of CoNLL-2003 (2003).
9. McCallum, A,. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. (2002).
10. McCallum, A., Rohanimanesh, K., Sutton,C.: Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences. In: NIPS Workshop on Syntax, Semantics and Statistics (2003)
11. McCallum, A., Schultz, K., Singh, S.: FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs. In: Advances on Neural Information Processing Systems (NIPS) (2009)
12. Novák A., Nagy V., Oravecz Cs.: Magyar ismeretlenszó-elemző program fejlesztése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 45–54

13. Radu, F., Ittycheriah, A., Jing, H., Zhang, T.: Named Entity Recognition through Classifier Combination. In: Proceedings of CoNLL-2003 (2003)
14. Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schawartz, R., Stone, R., Weischedel, R. and the Annotation Group: BBN: Description of the SIFT System as Used for MUC-7. In: MUC-7. Fairfax, Virginia (1998)
15. Sutton, C.: GRMM: GRaphical Models in Mallet..http://mallet.cs.umass.edu/grmm/.
16. Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate Named Entity corpus for Hungarian. In: Proceedings of International Conference on Language Resources and Evaluation (2006)
17. Tjong Kim Sang, E. F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of the 6th conference on Natural language learning - Volume 20 (2002)
18. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: CONLL '03 – Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (2003)
19. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL 2003 (2003) 252–259
20. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: Tanács, A., Vincze, V. (szerk.): MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 275–283

Magyar nyelvű klinikai dokumentumok előfeldolgozása

Siklósi Borbála[1], Orosz György[1], Novák Attila[2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai Kar, 1083 Budapest, Práter utca 50/a
e-mail: {siklosi.borbala, oroszgy}@itk.ppke.hu

[2] MorphoLogic Kft., 1116 Budapest, Kardhegy utca 5.
e-mail: novak@morphologic.hu

Kivonat A klinikai dokumentumok feldolgozásának első lépése azok strukturálása és normalizálása. Bemutatjuk, hogy a szerkezeti egységek hiányát hogyan tudtuk a formázási jegyek alapján automatikus transzformációkkal pótolni, illetve alapvető metainformációkat a folyó szövegből kinyerni. Ezután a korpusz szöveges részeit elválasztottuk a nem szöveges részektől, az így kapott halmazra automatikus helyesírás-javító, illetve javaslatgeneráló rendszert hoztunk létre. Módszerünk elsősorban a rendelkezésünkre álló korpusz statisztikai viselkedésére épül, de külső erőforrásokat is bevontunk a jobb minőség elérése végett. Az algoritmust két funkciója: a helyesírás-javítás, illetve a javaslatgenerálás alapján értékeltük ki. Beláttuk, hogy módszerünk a teljesen automatikus javításra pillanatnyilag önmagában nem alkalmas, azonban ez nem is volt cél, viszont minimális emberi közreműködéssel hatékonyan alkalmazható egy helyes orvosi-klinikai korpusz létrehozására.

Kulcsszavak: automatikus helyesírás-javítás, orvosi szövegfeldolgozás, szövegnormalizálás

1. Bevezetés

A legtöbb kórházban az orvosi feljegyzések tárolása csupán archiválás, illetve az egyes esetek dokumentálása céljából történik. Az így felhalmozódott adattömegek felhasználása jelenleg csupán az egyes betegek kórtörténetének visszakeresésére korlátozódik. A nyelvtechnológia, a számítógépes ontológiák és a statisztikai szövegfeldolgozó algoritmusok lehetővé tennék a folyó szövegekben rejlő összefüggések, rejtett struktúrák felfedését, a feljegyzésekben található információhalmaz elérését, abból tudás kinyerését.

Az angol nyelvterületen az ilyen irányú kutatások előrébb járnak, azonban alkalmazhatóságuk a magyar nyelv sajátosságai miatt sokszor nem egyértelmű, továbbá számos olyan nyelvi erőforrás, ami az angol nyelvre hozzáférhető, magyarra nem létezik. Az orvosi dokumentumok feldolgozása során nem csak a

magyar nyelv nyelvtani sajátosságait kell figyelembe venni, hanem az orvosi szövegekre különösen jellemző nehéz, olykor hiányos szintaktikai szerkezeteket, rövidítéseket, idegen kifejezéseket is kezelni kell.

Ezen tapasztalatok alapján fogalmazódott meg az igény, hogy a magyar nyelvű klinikai dokumentumok feldolgozását a más nyelveken már létező alkalmazások adaptálása, továbbfejlesztése és alkalmazhatóvá tétele révén aktívan kutatott területté tegyük, tekintettel a kutatás várható hasznára.

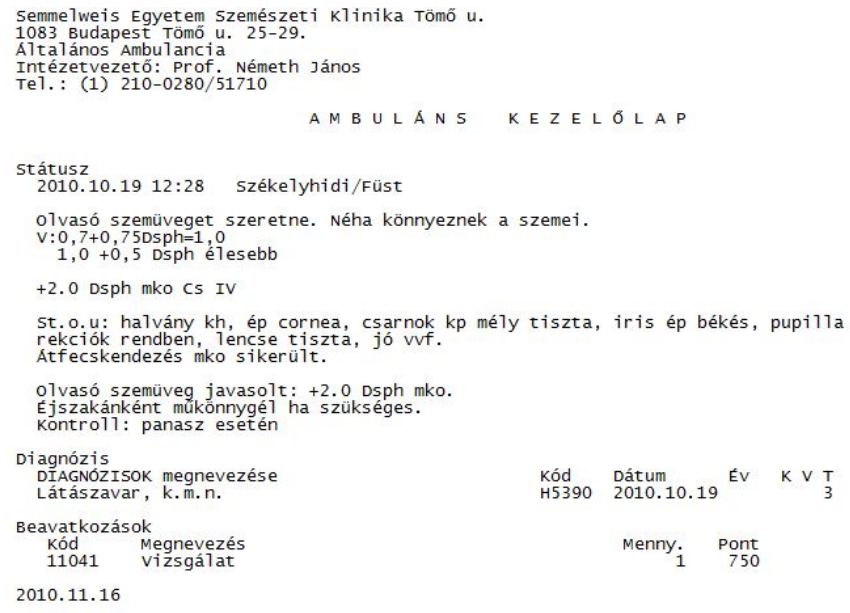
Hosszútávú célunk egy olyan keretrendszer készítése, amely orvosi dokumentumokat feldolgozva segíthet a klinikai szakembereknek új összefüggések feltárásában. Cikkünkben egy ilyen rendszer megvalósításának kezdeti lépéseit mutatjuk be. Az első probléma a rendelkezésünkre álló nyers orvosi szövegek egységes reprezentációjának kialakítása. Bár a meglévő klinikai dokumentumok láthatóan rendelkeznek struktúrával, de ezekre csak a formázás, illetve a tartalom értelmezése alapján lehet következtetni. Jelentős nehézség még a dokumentumokkal kapcsolatban, hogy készítőik nem fordítanak hangsúlyt a helyes és konzisztens fogalmazásra, tagolásra, helyesírásra. Így szükségesnek láttuk a dokumentumokban meglévő zaj (helyesírási hibák) csökkentését, ami akár orvosonként/asszisztensenként, illetve osztályonként is változó lehet.

Cikkünkben bemutatjuk a nyers orvosi dokumentumok feldolgozásakor alkalmazott algoritmusainkat, amelyekkel strukturális egységekre bontottuk a kórlapokat, és ezzel együtt a felszíni jegyekből könnyen meghatározható metainformációkat is kinyertünk, továbbá meghatároztuk az átfedő dokumentumrészeket. Ezek után bemutatjuk a szöveges és a nem szöveges részek elválasztására alkalmazott megoldásunkat, majd az automatikus helyesírás-javító rendszer első eredményeit ismertetjük.

2. A nyers dokumentumok strukturálása

Rendelkezésünkre állt a klinikai dokumentumok (kórlapok) egy rendezetlen halmaza. A szövegek struktúrájára csak a formázás, illetve a tartalom értelmezése alapján lehetett következtetni. Az alapvető tagoláson kívül – mely önmagában sem tekinthető egységesnek – nem voltak a további feldolgozás szempontjából használhatóan elkülönített egységek. Az adathalmaz jelentős része redundáns, az egyes esetek kórelőzményének minden korábbi fázisa a kórtörténet összes dokumentumában ismételten megjelenik, így a folyamat időben későbbi szakaszában készült leírások egyre hosszabbak, az összes előzmény másolása révén. Itt szintén tapasztalható volt az egységes rendszer hiánya, a folyamatok „összemásolása" többféle módon történt (időben korábbi/későbbi dokumentumok előrébb vagy hátrébb tolódása; diagnózisok elvetése/halmozása, stb.)

Mivel az eltérő szakterületek dokumentumainak felépítése eltérő, ezért elsőként a szemészeti dokumentumok feldolgozása indult el, melynek eredményei kisebb átdolgozással alkalmazhatóak lesznek más szakterületek, végül pedig általános orvosi szövegek feldolgozására.



```
Semmelweis Egyetem Szemészeti Klinika Tömő u.
1083 Budapest Tömő u. 25-29.
Általános Ambulancia
Intézetvezető: Prof. Németh János
Tel.: (1) 210-0280/51710

                          A M B U L Á N S   K E Z E L Ő L A P

Státusz
  2010.10.19 12:28   Székelyhidi/Füst

  Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
  V:0,7+0,75Dsph=1,0
    1,0 +0,5 Dsph élesebb

  +2.0 Dsph mko Cs IV

  St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla
  rekciók rendben, lencse tiszta, jó vvf.
  Átfecskendezés mko sikerült.

  Olvasó szemüveg javasolt: +2.0 Dsph mko.
  Éjszakánként műkönnygél ha szükséges.
  Kontroll: panasz esetén

Diagnózis
  DIAGNÓZISOK megnevezése                    Kód    Dátum      Év   K V T
  Látászavar, k.m.n.                         H5390  2010.10.19          3

Beavatkozások
   Kód      Megnevezés                                Menny.   Pont
   11041    Vizsgálat                                      1     750

2010.11.16
```

1. ábra. Egy eredeti dokumentum

2.1. XML-struktúra

A feldolgozás első lépéseként tehát szükséges volt a dokumentumok struktúrájának azonosítása és annak szabványos ábrázolása. Az egységek meghatározása egy egyszerű szabályalapú mintaillesztő eljárással történt, mely a rekordok szemmel is látható tagolására épül. Így a folyó szövegekben meglévő formázási elemeket transzformáltuk a szerkezetet meghatározó jellemzőkké. A kinyert struktúrák és metainformációk XML-struktúrában való tárolása során a dokumentumok felépítése a következőképpen alakult:

- Teljes eredeti: a teljes dokumentum szövegét eredeti formában is megtartottuk a későbbi megjelenítés egyszerűsítése céljából
- Tartalom: a dokumentumok szabad formájú szöveges részeit is tovább tagoltuk fejléc, diagnózisok, beavatkozások, javaslat, státusz, műtét, panasz, stb. részek megjelölésével.
- Metaadatok: a dokumentumok egyes részein alapvető automatikus módszerekkel jól felismerhető, a folyó szöveges részektől elkülönülő, adatokat tartalmazó egységeket nyertünk ki, ellátva őket az adatok típusára vonatkozó címkékkel. A következő metaadatokat nyertük ki: az adott dokumentum típusa (zárójelentés, kezelőlap stb); a dokumentumot kibocsátó osztály azonosítója; a táblázatos formában explicit módon megjelölt diagnózisok, illetve beavatkozások megnevezése és kódja.

– Egyszerű névelemek: a munkánk jelenlegi fázisában az egyszerű mintaillesztéssel kinyerhető névelemek (dátumok, orvosok, műtétek) megjelölése is megtörtént, azonban az erre alkalmazott módszerek finomítása és pontosítása még feltétlenül szükséges.
– Kórtörténet: az egyes betegek kórlefolyásának tárolása a klinikai adminisztrációs rendszer hiányosságai miatt jelenleg többféleképpen történik. Gyakori eset, hogy a kórelőzmény teljes szövege hozzáadódik az újabban keletkező dokumentumhoz, így folyamatosan egyre nagyobb dokumentumok kapcsolódnak egy pácienshez, melyek egymást tartalmazzák. Nincs egységes rendszer arra vonatkozóan sem, hogy a korábbi vizsgálatok leírása a dokumentumban előrébb vagy hátrébb – esetleg vegyesen – kerül be. Ennek ellenére megvalósult egy automatikus sorbarendezés, amelynek során minden dokumentumhoz eltároljuk az őt követő, és őt megelőző dokumentumokat – ha vannak ilyenek.

2.2. Szöveges részek elkülönítése

Az így kapott struktúra jól elkülöníti a dokumentumok egyes részeit, azonban korántsem elegendő ahhoz, hogy a szöveges részek önállóan kezelhetőek legyenek. Az általunk vizsgált szemészeti dokumentumokra különösen jellemzőek az esetek nagy részében túlnyomóan folyó szöveget tartalmazó szakaszokba ékelődő olyan nem folyó szöveg típusú részek, melyek az előfeldolgozás során zajként viselkednek. Ilyen részletek a laboreredmények, különböző számértékek, elválasztó karaktersorozatok, valamint csupán rövidítéseket, speciális jeleket tartalmazó megállapítások. Ezek kiszűrése szükséges volt ahhoz, hogy a nyelvi előfeldolgozás későbbi lépései során alkalmazott algoritmusok alapját képező korpusz előállítható legyen. Mivel azonban ezek a mintázatok önmagukban sem egységesek, különböző stílusú (feltételezhetően más-más orvos, illetve asszisztens szokásait tükröző) dokumentumok között még inkább változó módon szerepelnek, ezért szabályok, illetve mintafelismerés segítségével nem lehetett kiszűrni ezeket. A legkézenfekvőbb megoldásként klaszterezést alkalmaztunk. Mivel ezek a tartalmak sokrétűek, ezért mondatszegmentálást nem alkalmazhattunk, így a sorokra bontott dokumentumban kötöttük össze azokat, amik jó eséllyel egy egységet alkotnak. Ha egy sor nem mondatvégi írásjelre végződik, a rákövetkező sor pedig nem nagybetűvel és nem számmal kezdődik, illetve ha egy sor végén mondatközi írásjel van (vessző, pontosvessző), akkor a két sort összekötöttük.

Így megtartottuk azokat a mondattöredékeket, amik a felszíni jellemzőik alapján az elkülönítendő (nem szöveges) részekhez állnak közelebb. Az így megjelölt konkatenált sorokat K-means klaszterező algoritmussal csoportosítottuk. Célunk két diszjunkt halmaz létrehozása volt, de $k = 2$ esetén nem volt elég hatékony az elkülönítés. Mivel a jellemzőhalmaz módosításával nem sikerült célt érnünk, a klaszterek számának vizsgálata során optimális eredményt $k = 7$ esetén kaptunk, (A hét halmazból kettő tartalmazott szöveges részeket, a többi öt pedig különböző jellegű nem szöveges részeket) A klaszterezésnél használt jellemzőhalmaz, és az így létrejött tanítóanyag alkalmazásával a későbbiekben osztályozással is jól besorolhatóak lesznek a dokumentumok egyes részei. Naive Bayes-osztályozással

tesztelve a jellemzőhalmazunk hatékonyságát, 98%-os pontosságot kaptunk egy 100 sorból álló teszthalmaz esetén.

3. Helyesírás-javítás

A dokumentumok alapvető strukturálása és a szöveges tartalmak meghatározása után a következő feladat a dokumentumok normalizálása volt, amelynek első lépése a helyesírási hibák javítása. Esetünkben ez nem csupán a magyar nyelv nehézségeiből eredő problémák megoldására korlátozódott, hanem sok olyan hiba is felmerült a szövegekben, melyek a szakterület sajátosságaiból erednek. A legjellemzőbb hibák az alábbiak voltak:

- elgépelés, félreütés, betűcserék,
- központozás hiányossága (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 egy elvi szabvány létezik ezek helyesírására vonatkozóan, az orvosi szokások változatosak, és még a szakértőknek is problémát jelent az ilyen szavak helyességének megítélése,
- hiányos megfogalmazások gyakori előfordulása, melyek nem tekinthetők a hagyományos értelemben vett rövidítéseknek, azonban teljes szavaknak, kifejezéseknek sem,
- szakterületre jellemző rövidítések, melyeknek sem a jelölés módja, sem a jelentése nem általánosítható.

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 feladat másik nehézségét az jelentette, hogy egyáltalán nem állt rendelkezésünkre nagy méretű helyesen írt klinikai korpusz, ami alapján elő tudtunk volna állítani a javításhoz használható nyelvi és hibamodelleket.

Mivel munkánk jelen fázisában célunk egy kisméretű helyesen írt korpusz előállítása, így a javítási feladatot egy egyszerű lineáris modellel valósítottuk meg. Ehhez különböző nyelvi modelleket kombináltunk, melyeket részben a hibás korpusz alapján építettünk, részben külső erőforrások bevonásával jöttek létre. Az első kettőt a javítás előtti szűrőként alkalmaztuk, a többit pedig a helyes alakok előállításához.

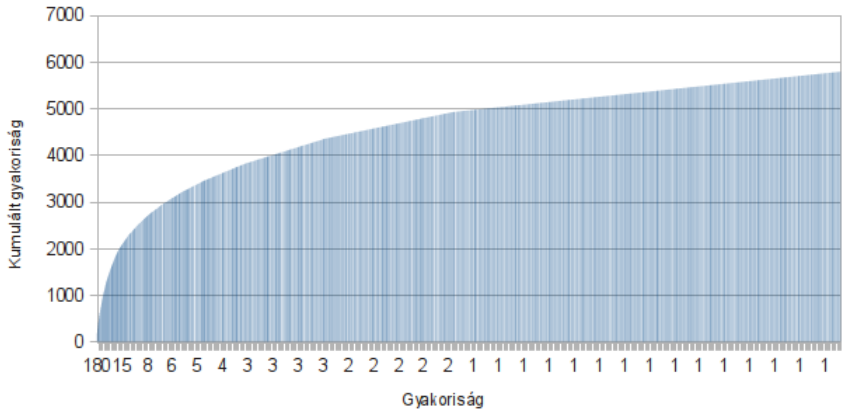
- Stopword lista: az általános stopwordöket kiegészítettük a korpuszra jellemző hasonlóan viselkedő tokenekkel, a leggyakrabban előforduló szóalakok közül kézzel válogatva ki ezeket. Ez elsősorban az írásjel-karaktereket, számokat és egyéb nem szóként vagy rövidítésként kezelendő tokeneket tartalmaz.

- Rövidítéslista: egyszerű mintaillesztéssel kiválasztottuk a potenciális rövidítéseket, majd ezt manuálisan szűrve jött létre a rendszerben használt szóhalmaz. Lehetséges rövidítésnek tekintettük azokat a tokeneket, amik nem mondatvégi szavak, rendelkeznek szó végi ponttal (és esetleg más punktuációval), morfológiai elemző számára ismeretlenek és nem hosszabbak egy előre megadott korlátnál (6 karakter).
- Morfológia által elfogadott szavak listája: kiválogattuk a korpuszból azokat a szóalakokat, amiket a HUMOR morfológiai elemző elfogadott, azaz helyesnek tekinthetőek. Ehhez a morfológiát célszerű volt kiegészítenünk a szakterületre jellemző szavakkal (gyógyszernevek, hatóanyagok, orvosi helyesírási szótár). Az így elfogadott szavak listájából unigram nyelvmodellt építettünk.
- Morfológia által el nem fogadott szavak listája: a fel nem ismert szóalakokból szintén építettünk egy gyakorisági modellt, melyet kétféle módon vettünk figyelembe a javított alakok ajánlása során. Amik kis gyakorisággal fordultak elő ebben a listában, azokat továbbra is rossznak tartottuk, amik azonban nagyon sokszor „rossz" alakban jelennek meg, azokat a morfológiának ellentmondóan, jó alakoknak tekintettük. Így azok a speciális használatú kifejezések, szakszavak, melyeket a morfológia alapján nem ismerünk fel, elfogadottá válhatnak, hiszen a használatuk elég gyakori ahhoz, hogy elfogadottnak tekintsük. A korpuszból generált kumulált előfordulási gyakoriságot reprezentáló görbe gradiensének változása alapján meghatározott küszöbértéknél (2. ábra) nagyobb gyakoriságú szavakat tekintjük helyesnek. A küszöbérték alatti frekvenciájú szavakat pedig $1 - f$ módosított gyakorisággal vettük figyelembe. (Abból a feltételezésből indultunk ki, hogy a legalább n-szer látott tokenek közt fellelhető a szóalakok legnagyobb hányada.)
- Általános és további szakszövegekből álló korpuszok: helyes alakok listájához hasonló gyakorisági modellt építettünk még a Szeged Korpusz alapján, illetve a BNO[3] betegségek listája és leírása alapján is. Itt feltételeztük, hogy csak helyes szóalakokat tartalmaznak.

A modellek létrehozása után a javítandó szöveget egy olyan nyelvfüggetlen tokenizálóval szegmentáltuk, amely képes rövidítések kezelésére a szóalakok és az írásjelek megtartásával egy tokenként, illetve hibatűrő. Érzéketlen a központozási hibákra, hiszen minden nem alfanumerikus karakter mentén – ami nem rövidítés része – új tokent hoz létre. Az fenti eszköz létrehozását az orvosi rekordok különleges nyelvezete (töredékes szerkezetek) és a központozási hibák sűrű megléte indokolta. A szegmentáló egy általános rövidítéslistát és a korábban említett szakterületi rövidítéslistát használja.

A tokenizálás után a stopword-lista és a rövidítéslista alapján kiszűrtük azokat a szavakat, amelyekre nem hajtunk végre javítást. A többi szóalak mindegyikéhez létrejön egy javaslathalmaz, mely az egy Levenshtein távolságra lévő szóalakokat, illetve a morfológia által generált lehetséges javaslatokat rangsorolva tartalmazza. A rangsorolás alapját a fenti modellek és a morfológia által együttesen meghatározott tényező képezi. Mivel minden szóalakra generálunk

[3] Betegségek Nemzetközi Osztályozása



2. ábra. A morfológia által fel nem ismert szóalakok kumulált gyakorisága.

javaslatokat, nem csak azokra, amiket a morfológia rossznak ítél, ezért azt az információt, hogy az eredeti alakot a morfológia elfogadja-e, a javaslatok rangsorolásánál kell figyelembe venni.

A rangsorolás végén a lehetőségek közül az első öt javaslatot tekintettünk lehetséges javításnak. Amennyiben az első és a második helyezett között elég nagy különbség volt, akkor az első javaslatot automatikusan elfogadtuk helyes javításnak, egyébként pedig felhasználói megerősítéssel történt meg a legjobb javaslat kiválasztása az első öt közül.

4. Eredmények

Megvizsgáljuk, hogy a kapott eljárás mint automatikus javító eszköz és mint helyesírási hibákra javaslatot nyújtó eszköz milyen eredményességgel bír. Mivel nem állt rendelkezésünkre helyesen írt szöveg, ezért a kiértékeléshez szükséges teszthalmazt kézzel kellett előállítani. Az eredeti korpusz véletlenszerűen kiválasztott 5%-át javítottuk ki (100 bekezdést). Sok szóalak esetén szembesültünk azzal, hogy gyakran az emberi javítás számára sem egyértelmű, hogy mely alakok fogadhatóak el helyesnek, különösen a vegyes latin–magyar írásmóddal írt szakkifejezéseknél. A módszer eredményeit az általánosan alkalmazott pontosság és fedés alapján értékeltük ki. A pontosság ebben az esetben azt mutatja meg, hogy az első legvalószínűbb javaslatot javításnak tekintve, mekkora a helyesen javított tokenek számának aránya az összes átírt token számához viszonyítva. A fedés értékéből pedig azt tudhatjuk meg, hogy eredeti anyagban lévő hibás tokenek mekkora részét javította a rendszer helyesen. Az F-mérték pedig ezek súlyozott harmonikus közepe. További metrikaként a helyes javaslatok rangját mérve a Mean Average Precision-t (MAP) alkalmaztuk.

1. táblázat. Eredmények az egyes modellek súlyozott kombinációira

OOV	**VOC**	**SZEGED**	**BNO**	**ISORIG**	**HUMOR**	**Pontosság**	**Fedés**	$F_{0.5}$	**MAP**
0,05	0,25	0,15	0,2	0,2	0,15	**0,5555**	**0,8769**	**0,5994**	**0,9863**
0,277	0,277	0	0,166	0,166	0,111	0,5417	0,8769	0,5865	0,9859
0,312	0,312	0	0.187	0.187	0	0.5385	0,8462	0,5807	0,9853

A kiértékelést a lineáris modellünk különböző súlyozott kombinációira vizsgáltuk:

- A morfológiai elemző által elfogadott és nem el fogadott szavak listája (VOC, OOV): Mivel a szövegeinket leginkább az eredeti korpusz jellemzi, ezért az ebből épített modelleket vettük figyelembe a legnagyobb súllyal. A sajátos stílus és szóhasználat miatt mindenképpen a korpuszon belüli előfordulás a hangsúlyosabb az általános szóhasználattal szemben.
- SZEGED, BNO: Mivel a BNO betegségek leírása sok szakkifejezést tartalmaz, viszont sokkal általánosabb formában, mint ahogy az a javítandó szövegekre jellemző, a Szeged Korpusz viszont teljesen általános, hétköznapi kifejezéseket, ezért ezeknek a súlyát kisebb mértékben szükséges figyelembe venni. Az eredményeken látszik, hogy a Szeged Korpusz figyelembevétele valamelyest javít az értékeken, azonban súlyának további növelésével nem érhető el jobb eredmény.
- ISORIG: Az eredetileg feltehetően helyesen írt kifejezések saját maguk valószínűségét erősítik, azonban ennek a tényezőnek a súlyát sem állíthattuk túl nagyra, hiszen ez a morfológia hibáját, illetve szakterületi hiányosságait erősítette volna.
- HUMOR: Jelentősen javított az eredményeken, ha a morfológia által elfogadott javaslatok súlyát megnöveltük. Ehhez szintén a szakkifejezésekkel bővített Humor-t használtuk.

A korpusz sajátos jellegének figyelembevétele miatt - az előzetes feltételezésünknek megfelelően - a meglévő korpuszra épülő modellek(OOV, VOC) magasabb súllyal való figyelembevétele, a morfológiával kiegészítve hozta a legjobb eredményt. (l. 1. táblázat)

A számszerű eredmények nem túl magas értékét több jelenség is magyarázza:

- A teszthalmaz viszonylag kis mérete nem ad teljes képet az összes hibáról, azonban egy nagyobb tesztszöveg létrehozása az emberi erőforrás igénye miatt nehéz.
- A rövidítések felismerésének hiányosságai. Sok esetben nem is értelmezhető a helyesírás-javítás a rövidítések felismerése, a tokenizálás során való helyes kezelése és a feloldás ismerete nélkül. Ilyen mondatok esetén, mint például: *„szemhéjszél idem, mérs. inj. conj, l.sin."* vagy *„Vitr. o.s. (RM) abl. ret. miatt."* a kiértékelés nem tekinthető mérvadónak, azonban a rövidítések megfelelő kezelését a későbbiekben fogjuk megvalósítani.
- Szakterületi többértelműség a latin-magyar vegyes alakok kezelése során. Az *a-á*, *c-k*, *o-ó*, stb. karakterpárok sok esetben egyenértékűek, az ilyen szavaknak sok alakja elfogadott, azonban ez nem fogalmazható meg általános

szabályként. A kiértékelés során minden szónál a gyakrabban előforduló néhány alakját tekintettük helyesnek, ez azonban enyhíthető lenne bármely alak engedélyezésével. Mivel mind az emberi olvasó számára, mind a további alkalmazás céljára alkalmas a jelenlegi módszerrel elérhető valamely forma, így csupán a számértékek növekedése lenne várható ettől, a tényleges minőség javulása nem.

2. táblázat. Példamondatok, automatikus javítással

Hibás mondat	**Automatikusan javított mondat**
A beteg intraorbitalis *implatatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonnegbeszélés* után kontrollvizsgálat javasolt.	A beteg intraorbitalis *implantatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonmegbeszélés* után kontrollvizsgálat javasolt.
Meibm *mirgy* nyílások helyenként sárgás *kupakszeráűen* elzáródtak, ezeket megint *túvel* megnyitom	Meibm *mirigy* nyílások helyenként sárgás *kupakszerűen* elzáródtak, ezeket megint *tűvel* megnyitom

A javaslatok sorrendjéről elmondható, hogy amikor nem az első eredmény tartalmazza a helyes alakot, akkor az első 5 javaslatban az esetek 99,12%-ban fellelhető a helyes szóalak. Továbbá az információ visszakeresésben használatos MAP metrikával is vizsgálva a találati listánk átlagos pontosságát, a legtöbb esetben 98% fölötti pontosságot kaptunk.

3. táblázat. Automatikus javaslatok hibás szavakhoz

Eredeti szó	**Első javaslat**	**Első öt rangsorolt javaslat**
látahtó	látható	'látható' : 0.1061, 'látahtó' : 0.0004, 'látahető' : 0.0, 'látaptó' : 0.0, 'lgtahtó' : 0.0
rajtra	rajtra	'rajtra' : 0.2631, 'rajta' : 0.1053, 'rajéra' : 0.1052, 'rajtura' : 0.1052, 'rajtja' : 0.10526
implatatumot	implantatumot	'implantatumot' : 0.1053, 'implatatumot' : 0.0009, 'implatatumít' : 0.0, 'őimplatatumot' : 0.0, 'implatáatumot' : 0.0

5. Összefoglalás

A jelenlegi algoritmus célja egy olyan helyesírás-javító alapalgoritmus megvalósítása volt, mellyel egy helyesnek tekinthető orvosi korpusz előállítását tudjuk támogatni. Ezáltal létrehozunk egy olyan szöveget, ami alapján pontosabb hibamodell építhető egy továbbfejlesztett rendszer betanításához.

A javítás egyelőre csupán szószinten történik, a környezet figyelembevétele nélkül. Ahhoz azonban, hogy a környezeteket is fel tudjuk használni az egyes szavak javítása során, egy jó minőségű n-gramokat tartalmazó nyelvmodellre is szükség lenne, aminek előállítása szintén helyes korpuszt igényel.

A javaslatok sorrendjének meghatározásához és azok generálásához, továbbá a modellek felépítéséhez jelenleg csupán teljes szavakat veszünk figyelembe, egy megfelelő hatékonyságú guesser segítségével azonban lemmaszinten is meg lehetne vizsgálni a javaslatok értékét. Ez minden olyan helyzetben segítene, ahol a helyesírási hiba a szótőben fordul elő.

A magyar nyelv agglutináló jellegéből és az összetett szavak írásmódjából adódóan a lehetséges szóalakok kvázi-végtelen száma miatt kézenfekvő volna súlyozott véges állapotú transzducerrel megoldani a javaslatgenerálási feladatot, ami tartalmazná mind a morfológiát, mind az előfordulási gyakoriságokat és a hibamodellt is.

Az elért eredmények alapján bemutattuk, hogy a hosszú távú célként megfogalmazott rendszer kezdeti állapotában is olyan alkalmazásokat tesz lehetővé, amelyek az eredeti dokumentumok kereshetőségében, alkalmazhatóságában, áttekinthetőségében jelentős előrelépést jelentenek. Bemutattuk, hogy egy átfogó, klinikai dokumentumokat elemző rendszer felépítése során a kiindulási állapot létrehozása sem triviális feladat, számtalan nehézséggel kell megküzdeni, ami különösen a kezdeti lépések során mindenképpen igényel emberi munkát is. Az így elérhető egyre nagyobb és egyre pontosabb korpusz javítása azonban fokozatosan teljesen automatikussá válhat.

Hivatkozások

1. Levenshtein, V.: Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission **1**(1) (1965) 8–17.
2. Contractor, D., Faruquie, T., Subramaniam, L.: Unsupervised cleansing of noisy text. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 189–196
3. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts., Stanford, California (2005) 150–157.
4. Pirinen, T.A., Lindén, K.: Finite-State Spell-Checking with Weighted Language and Error Models – Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In: Xth SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010. (2010) 13–18.
5. Patrick, J., Sabbagh, M., Jain, S., Zheng, H.: Spelling correction in Clinical Notes with Emphasis on First Suggestion Accuracy. In: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining. (2010) 2–8.
6. Farkas, R., Szarvas, G.: Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics **9** (2008)

IV. Beszédtechnológia

Nyelvimodell-adaptáció ügyfélszolgálati beszélgetések gépi leiratozásához

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

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

Kivonat: A folyamatos nagyszótáras gépi beszédfelismerés kritikus eleme a statisztikai nyelvi modell, melynek betanításához feladatspecifikus (in-domain) tanítóadatra van szükség. Ilyen tanítóadat azonban a gyakorlatban csak korlátozott mennyiségben áll rendelkezésre, mely felveti a feladattól független vagy ellenőrizetlen (out-of-domain) tanítószövegek felhasználását is. Formálisan nyelvi modell adaptáció révén építhető be az addicionális tanítószövegben tárolt tudás a feladatspecifikus nyelvi modellekbe. Cikkünkben azt vizsgáltuk, hogy telefonos ügyfélszolgálati hanganyagok felismerési pontossága javítható-e a különféle nyelvimodell-adaptációs technikákkal. Kísérleteink szerint mind felügyelt, mind felügyelet nélküli nyelvimodell-adaptációval szignifikánsan növelhető a valós beszélgetéseket leiratozó rendszerek pontossága.

1 Bevezetés

A jelenleg elterjedt nagyszótáras beszédfelismerők statisztikai úton tanított **nyelvi modell**t használnak, így a modell pontosságát döntően befolyásolja, hogy milyen mennyiségű és minőségű tanítószöveg áll rendelkezésünkre. Jó minőségű tanítószöveg általában a felismerési feladathoz illeszkedő hanganyagok kézi leirataiból állítható elő (**in-domain tanítószöveg**). A gyakorlatban azonban a begyűjthető hanganyagok mennyisége és a kézi leiratozás költségei határt szabnak az ilyen úton nyerhető tanítószöveg méretének. Éppen ezért a tudományos közösséget régóta foglalkoztatja, hogyan lehet az akusztikus modellek adaptációjához hasonlóan egy feladattól független (**out-of-domain**), de robosztus nyelvi modellt egy in-domain, de elégtelen mennyiségű adaton tanított modellhez adaptálni.

Cikkünkben különböző méretű és feladatunkhoz különböző mértékben illeszkedő tanítószövegek alapján készült nyelvi modelleket kísérelünk meg adaptálni ügyfélszolgálati beszélgetések felismerésre készített rendszerünkhöz. Megmutatjuk, hogy milyen módon célszerű eljárni, ha kisméretű, de a feladathoz jól illeszkedő kiegészítő szöveghez jutunk, illetve ha egy több tízmillió szót tartalmazó webkorpuszt szeretnénk felhasználni az in-domain modell javítására. **Felügyelt** adaptáció mellett **felügyelet nélküli** adaptációs kísérleteket is végzünk, azaz megvizsgáljuk, hogyan

használhatóak fel a felismerés korábbi kimenetei a nyelvi modell további pontosítására.

A nyelvimodell-adaptációs technikáknak alapvetően két nagy ágát kell megkülönböztetnünk [2]. Az első módszer az ún. maximum a posteriori (MAP) becslésen alapszik [4], és a célja, hogy úgy változtassa meg az out-of-domain modell paramétereit, hogy azok az in-domain modell paramétereinek eloszlását kövessék. A másik adaptációs megközelítésnél az objektív cél az, hogy az out-of-domain nyelvi modell minél kevesebb felismerési hibát vétsen egy kijelölt in-domain tesztanyagon. Itt a paraméterek hangolása diszkriminatív tanítás útján történik. A két megközelítés közül a MAP-adaptáció sok esetben jobban teljesít [2], mint a diszkriminatív tanítás, emellett a megvalósítása is egyszerűbb, így kísérleteinkben ezt módszert alkalmaztuk. A felügyelet nélküli adaptáció hatékonyabbá tehető, ha konfidenciaadatok alapján súlyozzuk vagy szűrjük a felismerési kimeneteket [5], azonban a rendelkezésünkre álló felismerési leiratok nem tartalmaztak megbízhatósági mértéket, így a felügyelet nélküli adaptáció esetén is csakúgy, mint a felügyelt esetben egy más típusú válogatási eljárást alkalmaztuk, melyet a cikkünk későbbi részében ismertetünk.

A következőkben először a kísérletekhez használt tanító és tesztadatbázisokat ismertetjük, majd kitérünk a modellek tanításnál és adaptálásánál alkalmazott módszerekre. A felismerési feladat és módszertan bemutatása után ismertetjük a különböző adaptációs megközelítésekkel kapott eredményeket, míg végül összefoglalását adjuk kísérleteink legfontosabb következményeinek.

2 Tanító és tesztadatbázisok

2.1 Tanító adatbázisok

Két ügyfélszolgálati rendszer in-domain nyelvi modelljének javítását tűztük ki kísérleteink céljaként, melyekre a továbbiakban **MTUBA** (Magyar Telefonos Ügyfélszolgálati Beszédadatbázis) I., illetve II. néven fogunk hivatkozni. Az **MTUBA I.** rendszernél az in-domain modell tanításához egy összesen 380 ezer szavas, kézi leiratokat tartalmazó tanítószöveg állt rendelkezésünkre. Az **MTUBA II.** feladatnál valamivel kisebb, összesen 280 ezer szavas kézi leiratot használhattunk. A felügyelet nélküli adaptációs kísérletekhez további két korpuszt gyűjtöttünk, melyek az egyes rendszerek felismerési kimeneteit tartalmazzák.

Az adaptációs kísérletekhez szükségünk volt egy a feladatokhoz semmilyen módon nem kötődő, out-of-domain korpuszra is. Ideális választásnak tűnt erre a célra a **Magyar Webkorpusz** [6]. Óriási mérete miatt csak a webkorpusz egy tizedét használtuk, mely önmagában 100 millió szót jelent, így elegendően nagynak bizonyult vizsgálatainkhoz. Az eredmények könnyebb értelmezhetősége érdekében egy mind méretében, mind illeszkedésében az in-domain és az out-of-domain korpuszok között elhelyezkedő kiegészítő tanítószöveget is szerettünk volna találni. Erre a megoldást egy ügyfélszolgálati levelezéseket tartalmazó, összesen 1,8 millió szavas korpusz jelentette. Ez az **e-mail korpusz** az in-domain szövegekhez hasonlóan ügyfélszolgálati témájú, így a webkorpusznál jobban illeszkedik a feladathoz, azonban szigorúan véve nem tekinthető in-domain tanítóanyagnak sem, ugyanis a

valódi beszélgetések leiratai sokkal több spontán elemet tartalmaznak, mint az elektronikus levelezés.

1. táblázat: A szöveges tanító adatbázisok méretei

	In-domain		**Felismerési kimenet**		**Kiegészítő korpusz**	
	MTUBA I.	MTUBA II.	MTUBA I.	MTUBA II.	E-mail korpusz	Web-korpusz
Méret [millió szó]	0,38	0,28	32	5,3	1,8	100

2.2 Tesztadatbázisok

A változatos nyelvimodell-konfigurációk kiértékeléséhez minden esetben a tanítóanyagoktól független tesztfelvételeket használtunk. Az MTUBA II. adatbázison több mint 5 órányi felvételt tudtunk tesztelési célokra elkülöníteni, mely megbízható kiértékelést tesz lehetővé, így tesztjeink többségét ezen végeztük. Annak érdekében, hogy minden esetben garantáljuk a független tanítást és tesztelést, egy másik, összesen 2 órás tesztanyagot is definiálnunk kellett az MTUBA II. adatbázison, melynek részletes okaira az 4.2.1 fejezetben térünk ki. Az MTUBA I. adatbázison egy kb. 1 órás tesztanyagot jelöltünk ki, melyen felügyelet nélküli adaptációval kapcsolatos kísérletet végeztünk.

2. táblázat: A teszt adatbázisok jellemzői

	Hossz [min]	Szavak száma [ezer szó]
MTUBA I.	56	5,7
MTUBA II.-5h	300	35
MTUBA II.-2h	120	14

3 Módszertan

3.1 Nyelvimodell-adaptáció

Kísérleteinkben a MAP becslésen alapuló nyelvimodell-adaptáció egy-egy speciális esetét jelentő **korpuszegyesítéses** (count merging) és **nyelvimodell-interpolációs** eljárásokat alkalmaztuk [1]. Két szöveges tudásforrás egyesítésének legegyszerűbb módja, ha n-gram statisztikájukat egyesítjük, és ez alapján készítjük el az n-gram nyelvi modellt. Gyakorlatban ez a két tanítószöveg összemásolásával vitelezhető ki a legegyszerűbben. Ez az eljárás jól működhet, ha hasonló mértékben illeszkedő tanítószövegeket egyesítünk. Abban az esetben azonban, ha egy out-of-domain tanítószöveget szeretnénk egy in-domain tanítószöveghez adaptálni, a korpuszegyesítéssel aránytalanul nagy súllyal kerülhetnek az egyesített modellbe a feladathoz rosszul illeszkedő tanítószöveg n-gram becslései [11]. Ilyenkor

jelenthetnek megoldást az interpolációs eljárások, melyekkel különböző nyelvi modellek n-gram becslései egyesíthetőek tetszőlegesen megválasztott súlyozó tényezővel. Mi az ún. lineáris interpolációt alkalmaztuk [7].

3.2 Perplexitásalapú előválogatás

Nyelvimodell-interpolációval hatékonyan orvosolhatóak az adaptáció során a modellek illeszkedési különbségeiből fakadó problémák. Önmagában használva az adaptáció azonban nem feltétlenül elegendően hatékony. Egy nagyméretű kiegészítő korpusz egyszerre tartalmaz olyan szövegrészeket, melyek a feladatunk szempontjából hasznos n-gramokat hordoznak és olyanokat is, melyek nyugodtan elhagyhatóak lennének. Ha valóban el tudjuk hagyni az adaptáció előtt az adaptálandó nyelvi modellből azokat az n-gramokat, melyek nem illeszkednek a feladatunkhoz, két ponton is nyerhetünk. Egyrészt csökkenthető a nyelvi modell mérete, másrészt a szükségtelen tanítóadatok elhagyásával a modell pontossága is nőhet.

A kiegészítő tanítószövegek sorainak előválogatására egy perplexitásalapú eljárást alkalmazunk. Ennek az egyszerű, de hatékony eljárásnak a lényege abban áll, hogy az in-domain nyelvi modell segítségével kiszámítjuk a kiegészítő korpusz minden sorához az illeszkedési mértéket (**perplexitást**). Ezek után kijelölünk egy küszöböt, amely alatti perplexitással rendelkező sorokat megtartjuk, míg a többit eldobjuk. Tehát az eljárás lényegében arra a feltétezésre épít, hogy azok a sorok, melyeket nagy pontossággal képes megjósolni az in-domain modell, potenciálisan tovább erősítik a modellt, míg azon sorok, melyek rosszul jósolhatóak, nem tartoznak szorosan a felismerési témához, így elhagyhatóak a modellből.

A perplexitást kétféle módon szokás számolni. A hagyományos eljárás szerint, az (**1**)-es képletben w_0-al jelölt mondatkezdő szimbólumot és a w_{K+1} mondatzáró szimbólumot is figyelembe vesszük a *P(s)* mondatvalószínűségek számításakor. Az ez alapján számított perplexitást szokás **PPL**-el jelölni.

$$P(s)_{PPL} = \prod_{i=0}^{K+1} P\left(w_i \mid w_{K-1},..., w_{K-(N-1)}\right) \quad (\mathbf{1})$$

Ezzel szemben a **PPL1**-gyel jelölt metrika a mondatvalószínűségek kiszámításakor nem veszi számításba mondatkezdő és mondatzáró karaktereket (**2**). Vizsgálataink során mindkét mérőszámot kipróbáltuk a gyakorlatban. Az erre vonatkozó eredményeket az 4.1.1 fejezet foglalja össze.

$$P(s)_{PPL1} = \prod_{i=1}^{K} P\left(w_i \mid w_{K-1},..., w_{K-(N-1)}\right) \quad (\mathbf{2})$$

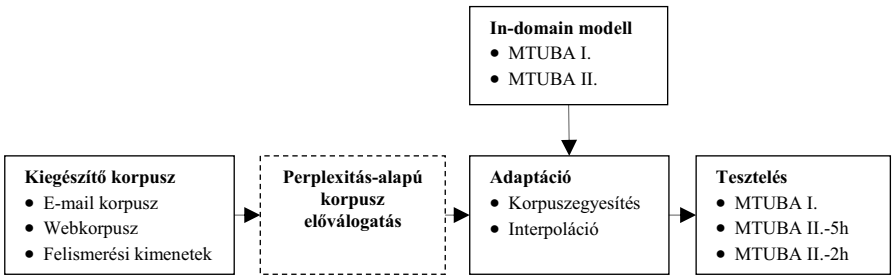
3.3 Tanítás és dekódolás

A vizsgált nyelvi modellek módosított Kneser-Ney simítás [3] használatával készültek az SRI Language Modeling Toolkit (**SRILM**) [10] segítségével. A létrehozott 3-gram, szóalapú modellekben entrópiaalapú metszést egyetlen esetben sem

alkalmaztuk. Interpolált nyelvi modellek készítéséhez és optimalizálásához az SRILM beépített lineáris interpolációs és perplexitásszámító eljárásait használtuk.

Az MTUBA I. feladathoz tartozó akusztikus modell tanításához az erre a célra elkülönített 27 óra, míg az MTUBA II. akusztikus modellhez 38 óra hanganyagot használtuk fel. Az annotált felvételek felhasználásával háromállapotú, balról-jobbra struktúrájú, környezetfüggő rejtett Markov-modelleket tanítottunk a Hidden Markov Model Toolkit [13] eszközeinek segítségével. A létrejött akusztikus modell 4048 egyenként 13 Gauss-függvényből álló állapotot tartalmaz az MTUBA I. modell esetén és 3535 egyenként 16 Gauss-függvényből álló állapotot az MTUBA II. modell esetén. Minden kísérletben a felismerési feladathoz illeszkedő akusztikus modellt használtuk.

A 8 kHz-en mintavételezett, telefonos tesztfelvételek lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre, és ún. vak csatornakiegyenlítő eljárást [8] is alkalmaztunk. A súlyozott véges állapotú átalakítókra (**WFST** – Weighted Finite State Transducer) [9] épülő felismerő hálózatok generálását és optimalizálását az Mtool keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztéshez a VOXerver [12] nevű WFST dekódert használtuk. A felismerő rendszerek teljesítményének értékeléséhez szóhibaarányt (**WER** – Word Error Rate) és karakterhiba-arányt (**LER** – Letter Error Rate) számoltunk, utóbbi gyakran pontosabb képet ad egy felismerő rendszer megbízhatóságáról morfémákban gazdag nyelvek esetén.



1. ábra. Kísérleteink általános módszertani lépései
(a szaggatott vonal opcionális lépést jelöl).

4 Kísérleti eredmények

Ebben a fejezetben a már bemutatott tanító- és tesztadatok felhasználásával, az előző fejezetben ismertetett módszerekkel elért eredményeinket mutatjuk be. Vizsgálataink első felében az MTUBA II. feladat nyelvi modelljéhez kísérelljük meg adaptálni a külső tudásforrásokat, majd a fejezet második felében a felismerési kimenetekkel visszacsatolt felügyelet nélküli adaptációban rejlő lehetőségeket mutatjuk be. Kísérleteink általános módszertani lépéseit az **1. ábra** foglalja össze.

4.1 Felügyelt adaptáció az MTUBA II. nyelvi modellhez

A fejezet során három tudásforrást próbálunk meg adaptálni az MTUBA II. in-domain nyelvi modellhez: nagyméretű, általános tematikájú webkorpuszt, a kisebb méretű, jobban illeszkedő e-mail szövegadatbázist és az MTUBA I. feladat tanítószövegét.

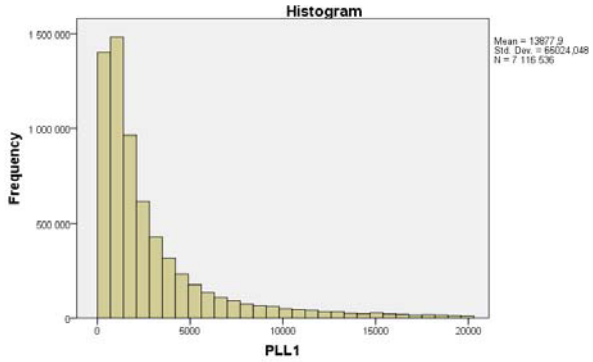
4.1.1 PPL és PPL1 metrika összehasonlítása

Annak eldöntésére, hogy a tanítószövegek sorainak előválogatásához melyik perplexitás-mérőszámot érdemes alkalmazni, terveztünk egy kísérletsorozatot. Első lépésként kerestünk olyan PPL és PPL1 értékpárokat, melyeknél a webkorpuszon végrehajtva a válogatást egyforma méretű tanítószöveget kapunk. A kérdés ezek után úgy módosult, hogy melyik ilyen módon kapott előválogatott tanítószöveggel érhetünk el nagyobb pontosságnövekedést az MTUBA II. felismerési feladaton. Ennek meghatározásához egyesítettük az előválogatott webkorpuszokat az MTUBA II. tanítószövegével, majd az egyesített tanítószövegeken tanítottunk új nyelvi modelleket. Ezután az új nyelvi modellekkel perplexitás- és szótáron kívüli szóarány (OOV – Out of Vocabulary) méréseket hajtottunk végre az MTUBA II.-5h tesztanyagon. A kísérletsorozat eredményeit a **3. táblázatban** foglaltuk össze.

3. táblázat: MTUBA II. in-domain modell és a PPL, valamint PPL1 alapján előválogatott webkorpusz korpuszegyesítéses adaptációjával kapott eredmények az MTUBA II.-5h teszthalmazon kiértékelve.

Válogatási módszer / határ	MTUBA II. tanítószöveg [millió szó]	Kiegészítő webkorpusz [+millió szó]	OOV arány (MTUBA II.-5h) [%]	PPL (MTUBA II.-5h) [-]
PLL-400	0,28	22	1,7	580
PLL1-750			**1,7**	**550**
PPL-200	0,28	7,5	2,1	501
PLL1-400			**2,1**	**454**
PPL-100	0,28	3	2,5	423
PPL1-260			**2,6**	**373**
PPL-50	0,28	1,5	2,9	357
PPL1-200			**2,9**	**320**

A 3. táblázat alapján azt mondhatjuk, hogy azonos kiegészítő korpusz méret mellett a PPL1 metrika segítségével előválogatott webkorpusz nagyobb mértékben járul hozzá az in-domain modell pontosításához. Ez abból olvasható ki, hogy az MTUBA II.-5h tesztanyagon mindkét megközelítés páronként nagyjából megegyező OOV-arány ért el, azonban a PPL1 válogatással kapható perplexitások minden korpuszméret mellett alacsonyabbak. Ennek oka az lehet, hogy a rövid, sok szótáron kívüli szót tartalmazó soroknál a PPL1 metrika reálisabb képet fest az illeszkedés mértékéről. A továbbiakban minden esetben PPL1 alapján végezzük a kiegészítő korpuszok sorainak előválogatását.

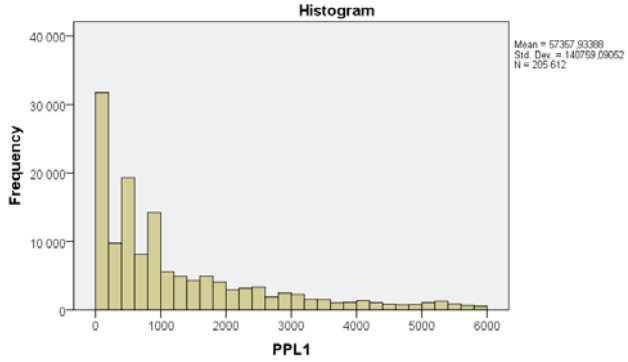


2 ábra. A webkorpusz sorainak PPL1 eloszlása az MTUBA II. in-domain modell alapján, [0;20000] tartományon ábrázolva.

4.1.2 Adaptációs paraméterek

Annak érdekében, hogy megfelelő válogatási küszöböt tudjunk beállítani a **webkorpuszon**, ismerni kell a sorainak PPL1 eloszlását (**2. ábra**). Az adaptációs kísérletekhez a már előző pontban is vizsgált „PPL1-400" illetve „PPL1-260" előválogatási határokat választottunk. 400-nál nagyobb határt megengedve, nagyon megnőtt volna az adaptált modell memóriaigénye, míg 260-nál kisebb határt beállítva már túl sok értékes sort veszítettünk volna. Az interpolációs súly optimalizálásakor mindkét korpuszméret mellett a webkorpuszok 0,1-es súlyozású figyelembevételével kaptuk a legalacsonyabb perplexitásokat az MTUBA II.-5h tesztanyagon.

Az **e-mail korpuszon** a webkorpusz esetében már bemutatott eljárást követtük. Először megvizsgáltuk a korpusz sorainak MTUBA II. in-domain modellel számított PPL1 eloszlását (**3. ábra**), majd ez alapján válogatási küszöbértékeket határoztunk meg. A két kiválasztott küszöbérték az eloszlás első csúcsának határához (1000), illetve a még számottevő mintával rendelkező tartomány határához (6000) illeszkedik. Az e-mail korpusz azonban a webkorpusznál két nagyságrenddel kevesebb szót tartalmaz, ezért a korpusz előválogatás mellett a válogatás nélkül kapható



3. ábra. Az e-mail korpusz sorainak PPL1 eloszlása az MTUBA II. in-domain modell alapján, [0;6000] tartományon ábrázolva.

eredményekre is kíváncsiak voltunk. A perplexitás minimalizálását célzó kísérleteink eredményeként a webkorpuszhoz hasonlóan itt is a 0,1-es kiegészítő modell súly adódott optimálisnak minden esetben.

A kísérletsorozat utolsó állomásaként az **MTUBA I.** modellt adaptáltuk az MTUBA II. modellhez. Mivel a két ügyfélszolgálati feladat szóhasználatában és fordulataiban nagyon hasonlít egymáshoz, az MTUBA I. közel in-domain tanítószövegnek tekinthető, így itt a korpuszegyesítéses eljárást is kiértékeltünk. Az MTUBA I. korpusz kis mérete miatt korpusz-előválogatást nem alkalmaztunk. Az interpoláció során az ideális kiegészítő modell súly 0,2-nek adódott.

4.1.3 Felügyelt adaptációs felismerési eredmények

A MTUBA II.-5h felismerési feladaton kiértékelt felügyelt nyelvimodell-adaptációs eredményeket a **4. táblázatban** foglaltuk össze.

4. táblázat: MTUBA II.-5h tesztanyagon mért felismerési eredmények felügyelten adaptált nyelvi modellek használatával.

Nyelvi modell	Szótár-méret [ezer szó]	OOV arány [%]	PPL [-]	WER [%]	LER [%]
MTUBA II. in-domain	21	4,3	167	46,4	25,0
+0,1 Webkorp. PPL1-400	386	2,1	208	**45,2**	24,6
+0,1 Webkorp. PPL1-260	228	2,6	201	45,5	24,7
+0,1 E-mail korpusz	70	3,3	181	45,4	24,6
+0,1 E-mail korpusz PPL1-6000	55	3,4	178	45,3	24,6
+0,1 E-mail korpusz PPL1-1000	40	3,7	176	45,6	24,7
+MTUBA I. (korpuszegyesítés)	37	3,1	189	45,4	24,6
+0,2 MTUBA I. (interpoláció)	37	3,1	176	**45,2**	**24,5**

A felismerési eredmények alapján látható, hogy a felügyelt adaptációval készült modellek használatával szignifikánsan alacsonyabb felismerési hibát érhetünk el, mint az in-domain MTUBA II. modellel. Bár a kisméretű in-domain nyelvi modellel mérhető a legkisebb perplexitás MTUBA II.-5h tesztanyagon, az adaptált nyelvi modellek ellensúlyozni tudják ezt nagyobb szótárméretükkel, melynek segítségével le tudják szorítani a tesztanyagon mérhető OOV arányukat.

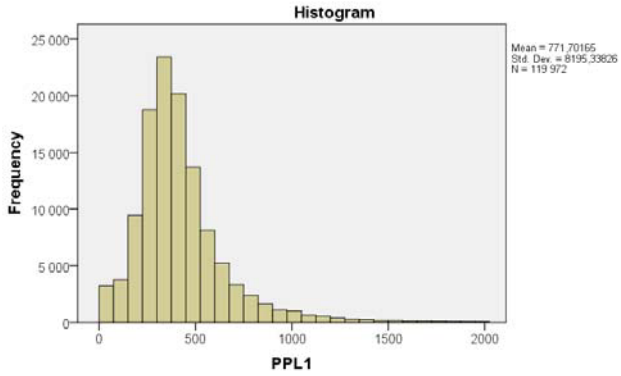
A legalacsonyabb felismerési hibát mind LER mind WER értelemben az MTUBA I. adaptációjával értük el, ráadásul az adaptált modellek közül ehhez tartozott a legkisebb szótárméret is. Igaz tehát, hogy a feladathoz jól illeszkedő tanítóanyagok a legnehezebben hozzáférhetőek és esetenként a legköltségesebbek is, azonban ezekkel lehet a leghatékonyabban végrehajtani az adaptációt. Megfigyelhető továbbá, hogy hasonló mértékben illeszkedő tanítószövegek esetén is eredményesebb eljárás a modell-interpoláció, mint a korpuszegyesítés.

Az MTUBA I.-től nagyon kicsit elmaradva, meglepően jól teljesített a webkorpuszos adaptáció. Igaz, hogy ugyanakkora WER eléréséhez itt tízszer akkora szótárra volt szükség, azonban az MTUBA I.-el ellentétben a webkorpuszt hatékonyan lehet adaptálni más felismerési feladathoz is, így egyfajta univerzális kiegészítő modellnek tekinthető. Az e-mail korpusszal mért eredmények is csak kis

mértékben maradnak el a két korábbi csoport eredményeitől. Itt a valódi érdekességet az adja, hogy összevethetőek a teljes és válogatott kiegészítő korpusszal kapott eredmények. Ez alapján azt mondhatjuk, hogy a túlzott metszés ronthatja az adaptáció hatásfokát (PPL1-1000), azonban az sem igaz, hogy a teljes out-of-domain korpusz alkalmazása jó megoldás. Optimális eredmény akkor született, amikor bár szűrtük a korpuszt, de nem túlzottan nagy mértékben. Mindez arra is utalhat, hogy akár pontosabb felismerési eredmény is elérhető lenne a webkorpusz használatával, ha az adaptáció előtt nagyobb előválogatási küszöböt alkalmaznánk, azonban ilyen nagy szótárméretű felismerő hálózatot szóalapon nem tudunk létrehozni a hálózatépítés nagy memóriaigénye miatt.

4.2 Felügyelet nélküli adaptáció

Felügyelet nélküli adaptációs kísérleteket az MTUBA I. és MTUBA II. feladaton is végeztünk. Vizsgálataink központi kérdése az volt, hogy a felismerő rendszer nyelvi modellje vajon milyen mértékben képes profitálni abból, ha az általa generált korábbi kimenetekkel adaptálunk.



4. ábra. Az MTUBA I. felismerési kimeneteit tartalmazó korpusz sorainak PPL1 eloszlása az MTUBA I. in-domain nyelvi modell alapján, [0;2000] tartományon

4.2.1 Adaptációs paraméterek

Felügyelet nélküli adaptáció esetén egyből adódik a kérdés, hogy vajon szükség van-e perplexitásalapú korpusz előválogatásra. A kérdés megválaszolásához felvettük a 32 millió szavas MTUBA I. felismerési kimenet korpusz PPL1 eloszlását **MTUBA I.** in-domain modell alapján (**4.ábra**). Míg a webkorpusz esetén egy nagyon vegyes szöveggel álltunk szemben, ezért jól különválaszthatóak voltak a jól és kevésbé jól illeszkedő sorok, addig a felismerési kimeneteket tartalmazó korpusznál sokkal egyenletesebb az eloszlás, és az illeszkedés mértéke is átlagosan nagyobb. Ez alapján az feltételezhető, hogy nagymértékű méretcsökkentés csak jól illeszkedő sorok elhagyásának árán valósítható meg. Éppen ezért az eredeti, válogatás nélküli korpusszal is végzünk adaptációt. Az ideális kiegészítő modellsúly 0,9-nek adódott az előválogatott és az eredeti korpusz használatakor egyaránt.

Az MTUBA I. mellett az **MTUBA II.** feladaton is szerettünk volna felügyelet nélküli adaptációs kísérleteket végezni. Ehhez azonban nem használhattuk az MTUBA II.-5h tesztanyagot, ugyanis az MTUBA II. rendszerrel előálló felismerési kimenetek a felismerő egy olyan konfigurációjából származtak, ahol az in-domain nyelvi modell az 5 órás tesztanyag leiratait is tartalmazta. Ez további 2 óra MTUBA II. hanganyag kézi átírását tette szükségessé, melyből megszületett a tanítástól már független MTUBA II.-2h tesztanyag. MTUBA II. esetén csak a teljes, válogatás nélküli kiegészítő korpusszal végeztünk kísérletet. A kiegészítő modellsúly értékét 0,8-nál mértük optimálisnak.

4.2.2 Felügyelet nélküli adaptációs eredmények

A felügyelet nélküli adaptációval készült felismerési eredményeket az **5. táblázatban** foglaltuk össze.

5. táblázat: Felügyelet nélküli adaptációs eredmények az MTUBA I. és MTUBA II.-2h teszthalmazon.

Nyelvi modell	OOV arány [%]	PPL [-]	WER [%]	LER [%]
MTUBA I. in-domain	5,7	310	48,0	25,9
+ 0,9 MTUBA I. felism. PPL1-300	5,7	207	47,5	25,5
+ 0,9 MTUBA I. felism.	5,7	192	**46,8**	**25,1**
MTUBA II. in-domain	5,6	255	50,9	27,5
+ 0,8 MTUBA II. felism.	5,6	173	**49,7**	**26,9**

Megfigyelhető, hogy felügyelet nélküli adaptációval az OOV arányt nem lehet csökkenteni, ami nem meglepő, hiszen ennél az eljárásnál az in-domain nyelvi modell által szolgáltatott felismerési kimeneteket integráljuk, azaz a rendszer szótára elvileg sem bővülhet. Érdekes eredmény azonban, hogy a korábbi kimenetek figyelembevételével jelentősen sikerült csökkenteni a perplexitást és így a szó-, illetve karakter-hibaarányt is. Azaz egy működő rendszerben érdemes lehet a felismerési eredményeket időről-időre adaptálni a nyelvi modellhez, ugyanis ezzel további költségek nélkül pontosabbá tehető a felismerés. A kiegészítő korpusz méretét itt azonban nem érdemes csökkenteni, mert mint az már a perplexitáseloszlás alapján is sejthető volt (**4. ábra**), nehéz olyan vágási határt találni, mely még jelentősen csökkenti a modellméretet, viszont nincs jelentős hatással a felismerési hibára.

5 Összefoglalás

Cikkünkben azt vizsgáltuk, hogy milyen módszerekkel és milyen mértékben lehet felügyelt és felügyelet nélküli adaptációs technikákkal telefonos ügyfélszolgálati hanganyagok felismerésére készített rendszerek in-domain nyelvi modelljeinek pontosságát javítani. Eredményeink alapján azt a következtetést vonhatjuk le, hogy amennyiben a nyelvi modell méretének az alacsonyan tartását tűzzük ki célul, akkor a legjobb eredményt a felismerési feladathoz jól illeszkedő nyelvi modellek

felhasználásával érhetjük el. Ilyen tanítóadatok azonban nem minden esetben állnak rendelkezésre korlátlan mennyiségben, illetve előállításuk a költségek miatt esetenként már nem gazdaságos. Ebben az esetben további pontosságnövekedés érhető el out-of-domain tanítókorpusz felhasználásával is, ha a cikkünkben ismertetett módon kinyerjük a feladathoz jól illeszkedő részeket a korpuszból. El kell azonban fogadni, hogy a nem feladatspecifikus tanítóadatok felhasználása óhatatlanul a modell méretének növekedésével jár.

Különösen értékes és a gyakorlatban jól hasznosítható eredmény továbbá, hogy két már működő ügyfélszolgálati felismerő rendszerben átlagosan 2,4%-os relatív WER-csökkenést sikerült elérni a felismerési kimenetek felügyelet nélküli adaptálásával. Felügyelet nélküli adaptációnál az OOV arány nem csökken, hiszen felismerő rendszer szótára nem bővül, így a javulás egyedül a nyelvi modell jobb előrejelző képességre vezethető vissza, mely a nagy mennyiségű in-domain hanganyag gépi leiratában rejlő tudás felhasználásának köszönhető.

Köszönetnyilvánítás

Kutatásunkat a TÁMOP-4.2.1/B-09/1/KMR-2010-0002-es, a KMOP-1.1.1-07/1-2008-0034-es, a GOP-1.1.1-09/1-2009-0068-as, a KMOP-1.1.3-08/A-2009-0006-os és a NAP-1-2005-0010-es projektek keretében az NFÜ és az NIH támogatta.

Bibliográfia

1. Bacchiani, M., Roark, B.: Unsupervised language model adaptation. In: Proc. of Acoustics, Speech, and Signal Processing (ICASSP ‘03) (2003) 224–227
2. Bacchiani, M., Roark, B., Saraclar,M.: Language model adaptation with MAP estimation and the perceptron algorithm. In: Proc. of HLT-NAACL 2004 (2004) 21–24
3. Chen, S. F., Goodman, J.: An Empirical Study of Smooting Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University (1998)
4. Gauvain, J.-L., Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. In: IEEE Transactions on Speech and Audio Processing Vol.2, No.2 (1994) 291–298
5. Gretter, R., Riccardi, G.: On-line learning of language models with word error probability distributions. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’01) (2001) 557–560
6. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proc. of the 4th international conference on Language Resources and Evaluation (LREC2004) (2004)
7. Jelinek, F., Mercer, R. L.: Interpolated estimation of Markov source parameters from sparse data. In: Proc.Workshop on Pattern Recognition in Practice (1980)
8. Mauuary, L.: Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition. In: Proc. of EUSPICO’98, Vol.1 (1998) 359–363
9. Mohri, M., Pereira, F., Riley, M.: Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language Vol.16, No.1 (2002) 69–88

10. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing. Denver (2002) 901–904
11. Tarján B., Mihajlik P.: Magyar nyelvű nagyszótáras beszédfelismerési feladatok adatelégtelenségi problémáinak csökkentése nyelvi modell interpoláció alkalmazásával. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Magyarország (2010). 216–223
12. Tarján, B., Mihajlik, P., Balog, A., Fegyó, T.: Evaluation of Lexical Models for Hungarian Broadcast Speech Transcription and Spoken Term Detection. In: CogInfoCom 2011: 2nd International Conference on Cognitive Infocommunications. Budapest, Hungary (2011) 1–5
13. Young, S., Ollason, D., Valtchev, V., Woodland, P.: The HTK book. (for HTK version 3.2.) (2002)

Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval

Csapó Tamás Gábor[1], Németh Géza[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{csapot, nemeth}@tmit.bme.hu

Kivonat: A prozódiai változatossággal kiegészített szövegfelolvasó rendszer olyan alkalmazásokban lehet hasznos, ahol hasonló jellegű, ismétlődő mondatok szintetizálására van szükség. A cikkben bemutatunk egy új módszert, amellyel egy adott szöveghez különböző prozódiával rendelkező mondatváltozatokat lehet szintetizálni. A prozódia komponensei közül a dallammal és hangsúllyal foglalkozunk az alapfrekvencia (F0) változtatásán keresztül. Ehhez egy statisztikai F0-modellt használunk fel rejtett Markov-modell alapú beszédszintetizátorban. A betanításhoz használt eredeti beszédkorpuszt a SOFM (Self Organizing Feature Map) módszerrel felbontjuk több részkorpuszra. A különböző beszédkorpuszokból betanult modellekkel eltérő dallamú mondatváltozatokat szintetizálunk azonos szöveghez. A mondatváltozatok közötti különbségeket megvizsgálva a szubjektív kísérletek azt mutatják, hogy az alapfrekvencia eltérése sok esetben elég jelentős ahhoz, hogy ez az emberi fül számára is észlelhető legyen.

1 Bevezetés

A szövegfelolvasó rendszerek érthetősége elérte a megfelelő szintet, viszont más tulajdonságokban még hiányosságok fedezhetőek fel. Ezek közé tartozik az emberi beszéd változatossága, amelyet ritkán modelleznek beszédszintetizátor rendszerekben. Az emberi beszédben a prozódia (dallam, hangsúly, ritmus) rendkívül változékony jellemző. Egy-egy mondatot még akarattal sem tudunk többször ugyanúgy elmondani, a mindennapi beszédben pedig nagy különbségek tapasztalhatóak mindegyik fenti jellemzőben. A legtöbb szövegfelolvasó rendszer ezzel szemben determinisztikusan állítja elő a prozódiát, azaz egy-egy bemeneti szöveghez ismételt szintéziskor mindig ugyanaz a prozódia tartozik. Ez sokszor ismétlődő, monoton minták túlzott előfordulásához vezet, ami zavaró lehet a szintetizált beszédben. A prozódiai minták ismétlődése azért fordulhat elő a szövegfelolvasó rendszerekben, mert a beszédszintetizátor mindig a legjobb prozódiát próbálja egy-egy mondathoz rendelni. Így az emberi beszéd változatossága lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, ami a változékonysághoz szokott, könnyen felismerhető, és hosszabb szintetizált beszédrészlet hallgatása során zavaró lehet.

1.1 Prozódiai változatosság

Az a cél, hogy a szövegfelolvasó egy-egy bemeneti mondatához ne mindig ugyanolyan prozódiájú mondatot szintetizáljunk, úgy valósítható meg, ha a bemeneti szöveghez többféle dallammenetet és ritmusszerkezetet tudunk generálni, és ezek közül a rendszer szintéziskor egyet kiválaszt. Ekkor ugyanis csökken a monotonitás, hiszen nem-determinisztikussá válik a mondatokhoz történő dallammenet- és ritmushozzárendelés. Ezen elv segítségével a hasonló szerkezetű egymás után előforduló mondatokhoz is eltérő prozódiát tudunk kialakítani. A cikk további részében a prozódia dallam és hangsúly részével foglalkozunk, az alapfrekvencia (F0) megfelelő beállításán keresztül.

Korábbi kutatásaink során a fenti célt korpuszalapú prozódiai modellel kíséreltük meg elérni. Egy nagyméretű beszédkorpuszból kigyűjtöttük a jellemző mondatdallam-mintázatokat, majd ezeket rendeltük a szintetizálandó szöveghez, hasonlósági mértékként a mondatrészek szótagszámát felhasználva. Ezeket a vizsgálatokat egy diádos beszédszintetizátorral végeztük el [2, 8]. Jelen cikkben a korábbiakhoz hasonló kísérleteket végzünk, statisztikai alapú prozódiai modellt felhasználva.

A nemzetközi szakirodalomban Díaz és Banga foglalkozott a prozódiai változatosság témájával egy korpuszos, elemkiválasztásos beszédszintetizátoron végzett kísérletek keretében [3, 4]. A módszer megőrzi az eredeti beszélő intonációjának változatosságát, mivel az összefűzendő elemek kiválasztásakor több lehetséges sorozatot megtart, melyek mindegyike hasonló minőségű szintetizált beszédet eredményez.

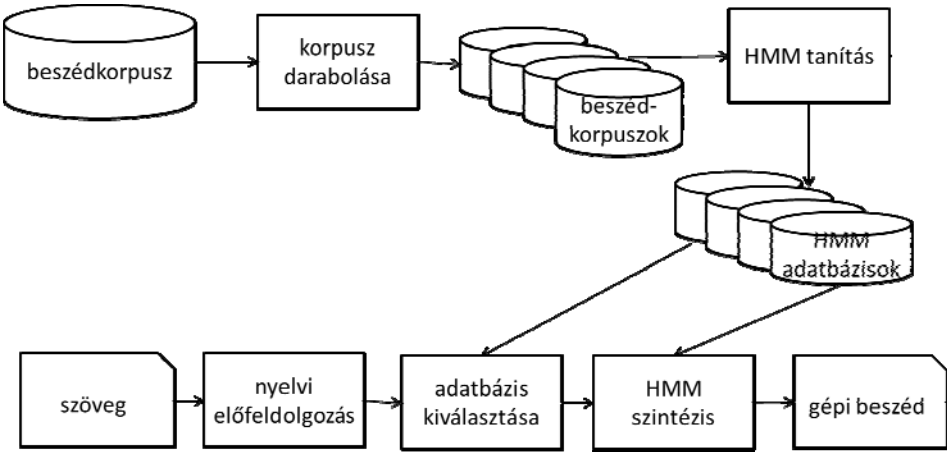
1.2 Rejtett Markov-modell alapú beszédszintézis

A szövegfelolvasó technológiák közül az elmúlt években a rejtett Markov-modell (Hidden Markov Model, HMM) alapú beszédszintetizátorral foglalkozott sokat a szakirodalom, melynek előnye a korábbi megoldásokhoz képest az alacsonyabb erőforrásigény és a statisztikai alapú parametrikus működés. A statisztikai beszédszintézisben a rendszer a tanulási fázis során kinyeri a tanító beszédadatbázisból a beszélő hangjára jellemző tulajdonságokat, és ezek alapján határozza meg később a szintézis során a beszéd generálásához szükséges paramétereket, majd egy beszédkódoló eljárás ez alapján létrehozza a szintetizált beszédet. Ezen paraméterek közé tartoznak például a beszéd alapfrekvenciája, hang- és szünetidőtartamai, illetve spektrális együtthatói.

A kutatás során a HTS [13] nyílt forráskódú HMM-alapú beszédszintetizátor magyar nyelvre adaptált változatát alkalmaztuk [12]. A kísérletekhez egy professzionális női bemondóval készült fonetikailag gazdag beszédadatbázist használtunk fel, amely 2 órányi 16 kHz-en mintavételezett, 16 bites kvantálású beszédet tartalmaz összesen 1940 kijelentő mondatban.

2 Módszerek

Amennyiben a HMM-alapú beszédszintézisben az eredeti tanító adatbázist több részre bontjuk, és ezekre külön-külön elvégezzük a statisztikai alapú tanítást, akkor ez alapján különböző paraméterértékeket tanul be a rendszer. A különböző résztanítóadatbázisok paramétereit egy beszédszintézisre épülő alkalmazásban párhuzamosan felhasználva (azaz felváltva használva az eltérő paraméterhalmazokat) elérhető, hogy egy adott mondathoz ne mindig ugyanaz a prozódia tartozzon. Ha a résztanítóadatbázisok mondatai elég különbözőek voltak, akkor a generált ismétlődő mondat tulajdonságai is eltérőek lesznek ismételt szintézis során, illetve azt várjuk, hogy hasonló szerkezetű mondatok is lényegesen eltérő prozódiával fognak rendelkezni. A HTS rendszerrel végzett betanítási és szintetizálási, valamint adatbázis feldarabolási lépéseket az 1. ábra mutatja be.



1. ábra: A beszédkorpusz feldarabolása, majd HMM tanítási fázis (felső rész). A bemeneti szöveghez HMM adatbázis kiválasztása, majd szintézis fázis (alsó rész).

2.1 Prozódiai távolságmértékek

Két mondat prozódiájának objektív összehasonlítására számos módszer található a szakirodalomban. Amennyiben csak a mondatok alapfrekvencia-menetét akarjuk összehasonlítani, többek között az átlagos négyzetes közép távolság (Root Mean Square Error, RMSE) [6], a Hermes-korreláció [5], vagy ez utóbbinak DTW-vel (Dynamic Time Warping) kiegészített változata [10] használható.

Az RMSE a következő módon számítható két mondat dallama között [6]:

$$RMSE_{f1,f2} = \sqrt{(\frac{1}{n}\sum_{i=1}^{n}(f_1(i) - f_2(i))^2)}$$

ahol f_1 és f_2 jelöli a két összehasonlítandó mondat F0 értékeit, n pedig a mérőpontok száma.

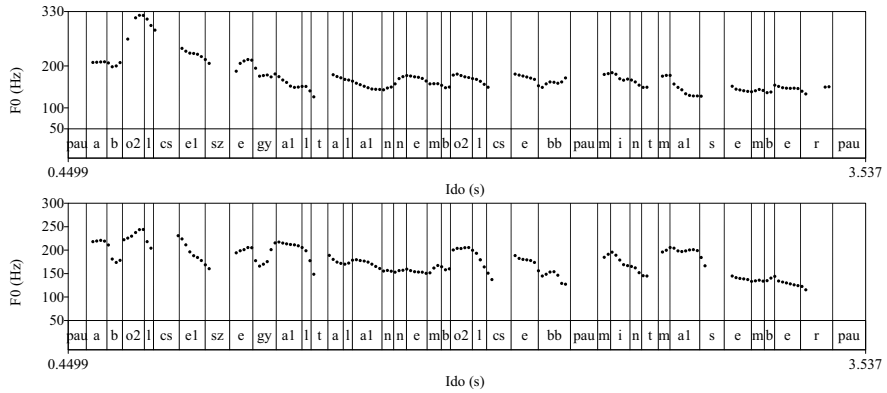
A Hermes-korreláció számítása [10] alapján:

$$Hermes_{f1,f2} = \frac{\sum_i w(i)(f_1(i)-m_1)(f_2(i)-m_2)}{\sqrt{\sum_i w(i)(f_1(i)-m_1)^2 \sum_i w(i)(f_2(i)-m_2)^2}}$$

ahol f_1 és f_2 jelöli a két összehasonlítandó mondat F0 értékeit, m_1 és m_2 ezeknek az átlagos F0-ja, ezen kívül a $w(i)$ egy súlyozó faktor az adott jelszakasz intenzitásának függvényében. Az alapfrekvenciát sok esetben nem közvetlenül Hz-ben, hanem logaritmizálva alkalmazzák ezen képletekben [10].

A DTW alapú Hermes-korreláció akkor lehet hasznos, ha olyan mondatok alapfrekvenciájának összehasonlítására van szükség, amelyeknek időszerkezete jelentősen eltérő.

A 2. ábra egy példát mutat két mondat F0-menete közötti RMSE távolság és Hermes-korrelácó értékére. A továbbiakban a Hermes-korrelációt használtuk fel prozódiai távolságmértéknek, mert a szakirodalom alapján ez alkalmasabb az alapfrekvencia-különbségek kimutatására, mint az RMSE távolság [5].



2. ábra: Egy mondat két különböző F0-menettel rendelkező változatának összehasonlítása (amennyiben a mondatok időszerkezete megegyezik). A szótagonkénti átlagos F0 értékek alapján számolva az RMSE távolság 0,1619; a Hermes-korreláció pedig 0,6337.

2.2 Tanító adatbázis felbontása

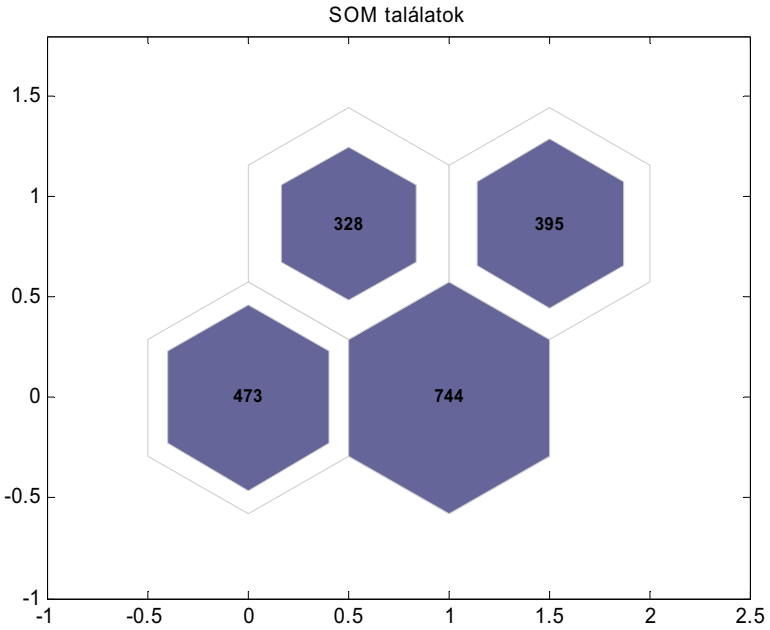
A kutatás során megvizsgáljuk, hogy egy adott beszélőtől származó különböző részt-anítóadatbázisokkal mennyire különböző prozódiájú mondatok állíthatóak elő a dallam, illetve alapfrekvencia tekintetében.

Az eredeti 1940 mondatból álló beszédkorpuszt több eltérő módon választottuk külön csoportokba. Első kísérletként véletlenszerűen szétválogattuk a mondatokat 2, 4, 8, illetve 16 csoportra, majd mindegyik rész-tanítóadatbázis segítségével elvégeztünk egy tanítást a HTS beszédszintetizátorral, majd leszintetizáltunk 40 mondatot. A szintetizálás során csak a betanult dallam modellt változtattuk (a gerjesztési, hangidőtartam és egyéb paramétereket változatlanul hagyva).

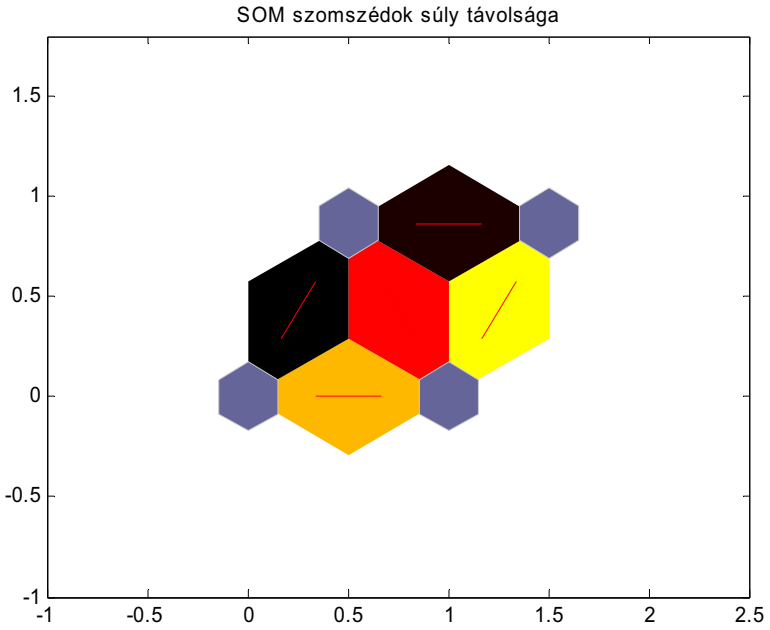
Ezután a 2.1 szakaszban ismertetett Hermes-korreláció objektív távolságmértéket felhasználva ellenőriztük, hogy egy adott szöveghez tartozó szintetizált változatok mennyire különböznek egymástól a mondat F0-menetének szempontjából. Ehhez a szótagonkénti átlagos F0 érték alapján számoltuk a Hermes-korrelációt. A véletlen szétválasztás esetén a mondatváltozatok közötti Hermes-korreláció magas volt (a legtöbb esetben 0,95 fölötti érték), azaz olyan mondatokat sikerült így szintetizálni, melyeknek F0-menetében nem fordult elő ezen mérték szerint jelentős különbség.

A véletlen választás mellett a továbbiakban azt vizsgáltuk, hogyan lehet gépi tanuló algoritmussal célzottan szétválasztani az eredeti beszédkorpuszt több klaszterre. Ehhez a választásunk a felügyelet nélküli tanításon alapuló Self-Organizing Feature Map (SOFM) eljárásra esett. A Kohonen által bemutatott megoldást [7] használtuk fel egy Matlab-alapú implementációban [1]. A SOFM-ot korábban sikeresen alkalmazták hangoskönyvek beszédanyagának expresszivitás szerinti szétválasztására [11]. A SOFM alkalmasnak látszik az alapfrekvencia szerinti szétválasztás feladatára, mivel felügyelet nélküli gépi tanulási módszer. A betanítás során azt kell beállítanunk, hogy hány részre bontsa szét a korpuszt az algoritmus. A SOFM bemeneteként felhasznált tulajdonságoknak az F0 bizonyos statisztikáit választottuk (minimum, maximum, átlag, szórás 1-1 mondaton belül), azaz mondatonként ezek a paraméterek álltak rendelkezésre a felügyelet nélküli tanításhoz.

A SOFM további előnye, hogy a többdimenziós adat kétdimenziós térképen ábrázolható. A 3. ábrán a klaszterezés eredményeként kapott 4 csoport látható, melynek során az 1940 mondat egy nagyobb és három kisebb részkorpuszra lett felbontva. A 4. ábra a szomszédos klaszterek közötti távolságok térképét mutatja. A hexagonok a bemeneti változókon (vagyis az F0 paraméterei) elvégzett felügyelet nélküli tanításból származó klaszterek. Azok a kapcsolatok, amelyek nagyobb távolságot mutatnak a klaszterek között, sötétebb színnel vannak jelölve. Az ábráról az látható, hogy a bal felső csoport távolsága nagy a többi csoporttól, míg a többi távolság ehhez képest alacsonyabb. Ez alapján azt várjuk, hogy azok a szintetizált mondatok, amelyek a bal felső mondatokkal mint tanító adatbázissal készülnek, dallam szempontjából nagyobb távolságra lesznek a többi tanító adatbázissal készült szintetizált mondatoktól, mint azok egymástól.



3. ábra: A SOFM alapú klaszterezés eredményeként felbontás után kapott négy tanítóadatbázis mondatainak elemszáma.



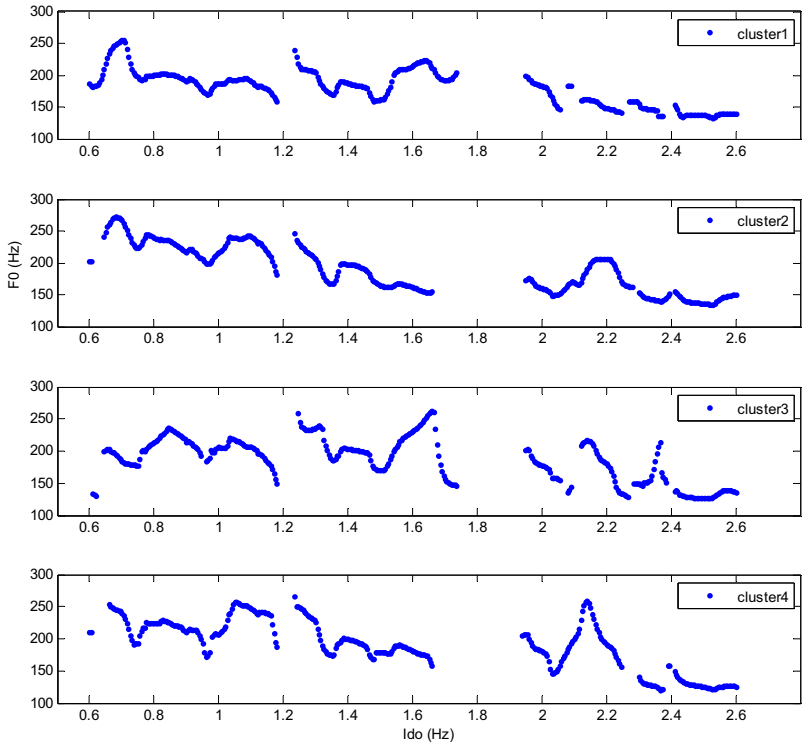
4. ábra: A SOFM alapú klaszterezés eredményeként felbontás után kapott négy tanító adatbázis egymástól mért távolsága. A világosabb szín kisebb, a sötétebb szín nagyobb távolságot jelöl.

3 Eredmények

A SOFM alapú klaszterezés eredményességét objektív és szubjektív vizsgálatokkal is ellenőriztük. 2000 kiválasztott mondatot leszintetizáltunk a 4 tanító adatbázisból származó F0-modellel külön-külön (a gerjesztési és időtartam paramétereket a teljes tanító adatbázisból származó modellből felhasználva).

3.1 Objektív különbségek

A mondatváltozatok közötti dallambeli különbség vizsgálatára a 2.1 szakaszban ismertetett Hermes-korrelációt használtuk fel. A szintetizált mondatok 4 változatát páronként összehasonlítottuk, majd kiszámoltuk az egyes mondatváltozatok közötti Hermes-korrelációt, melyre egy példát az 5. ábra és az 1. táblázat #1625 része mutat.



5. ábra: A #1625 mondat („Zsigmond nem tagadja, hogy ő zsidó.”) négy szintetizált változata, különböző tanító adatbázisokból kiindulva. Az alapfrekvencia-menet (és így a mondatdallam, illetve a hangsúlyok helye és erőssége) eltérő a különböző változatokban.

Ezután a 2000 mondatból kiválasztottunk 10 mondatot, melyeknél a változatok közötti F0 szerinti Hermes-korreláció a legalacsonyabb volt (így várhatóan ezek között észlelhető a legnagyobb különbség a mondatdallamban).

3.2 Szubjektív különbségek

A 10 legnagyobb objektív különbséggel rendelkező mondat 4-4 változatát választottuk ki a szubjektív teszt hanganyagához páros összehasonlítás keretében, így összesen 60 mondatpár állt rendelkezésre. A meghallgatásos teszt célja az volt, hogy ellenőrizzük, a Hermes-korreláció milyen mértékben mutatja meg a mondatdallambeli különbséget egy percepciós vizsgálathoz képest. Hasonló vizsgálatot végeztek korábban például német mondatokon [9].

A meghallgatásos tesztet internetes tesztfelületen végeztük. A mondatokat páronként kellett meghallgatniuk a tesztelőknek, és arra a kérdésre válaszolniuk, hogy „Hallasz-e különbséget a két mondat dallama között? Igen – Nem". Ezután ha „Igen"-nel válaszoltak, egy második kérdést is meg kellett válaszolniuk: „Ha hallottál különbséget, akkor milyen mértékű? Kicsi – Közepes – Nagy".

A mondatpárok meghallgatását 9 tesztelő végezte el. A tesztelők mindannyian ép hallású, magyar anyanyelvű emberek voltak, a 23-60 év közötti korosztályból (átlagosan 33 év). Egy részük a témához értő beszédtechnológiai szakértő vagy fonetikus volt, míg a többiek egyetemi hallgatók köréből kerültek ki. A teszt átlagos meghallgatási ideje 12 perc volt.

Az 1. táblázatban hasonlítjuk össze a mondatváltozatok között mért Hermes-korrelációt, és a tesztelők „Igen" válaszainak arányát. A szubjektív teszt 2. kérdését, (azaz a dallambeli különbség mértékét) itt nem vettük figyelembe, de az észrevehető volt a válaszok között, hogy a tesztelők leggyakrabban „kicsi" és „közepes" különbséget jelöltek csak be. A táblázatban a Hermes-korrelációnál az alacsonyabb érték jelent nagyobb F0 eltérést, míg az „Igen" aránynál a nagyobb szám jelenti azt, hogy többen észleltek különbséget a mondatváltozatok dallamában. Az eredmények alapján az objektív és a szubjektív mérték között nem található erős összefüggés ($R^2 = 0{,}115$).

A 60 mondatpárból összesen 35 esetben válaszolta a tesztelők legalább 65%-a, hogy hall különbséget a változatok között. A maradék 25 mondatpárt megvizsgálva az derült ki, hogy ezekben az esetekben a mondatváltozatok közötti szótagonkénti átlagos F0 különbsége legfeljebb 10-20 Hz volt. Azoknál a mondatpároknál, ahol hallottak különbséget a tesztelők, a legnagyobb F0 különbség akár a 70 Hz-et is elérte, és több helyen előfordult, hogy a mondat hangsúlya (az ereszkedő jellegű alapfrekvencia-menetből lényegesen kiugró rész) is másik szóra került. A #0074-es mondat (*„A bölcsész egyáltalán nem bölcsebb, mint más ember."*) esetén például a négy változatban különböző pozíciókra helyeződött a mondathangsúly: *„bölcsész"*; *„egyáltalán"*; *„bölcsebb"*; *„más"*. Ezek közül nem minden változat megfelelő, a *„más"* szóra helyezett hangsúly például helytelen hangsúlyozást jelent.

1. táblázat: A 10 kiválasztott mondat 4-4 változata közötti Hermes-korreláció és a szubjektív teszt alapján számolt különbség.

Mondat	v1	v2	Hermes-korreláció	Szubjektív „Igen”
#0044	1	2	0,7833	88,89%
#0044	1	3	0,7416	66,67%
#0044	1	4	0,8271	55,56%
#0044	2	3	0,9408	55,56%
#0044	2	4	0,9071	33,33%
#0044	3	4	0,9385	33,33%
#0046	1	2	0,7697	44,44%
#0046	1	3	0,7410	44,44%
#0046	1	4	0,7185	77,78%
#0046	2	3	0,9356	22,22%
#0046	2	4	0,9158	66,67%
#0046	3	4	0,9644	88,89%
#0069	1	2	0,7663	77,78%
#0069	1	3	0,8016	66,67%
#0069	1	4	0,8260	77,78%
#0069	2	3	0,9273	22,22%
#0069	2	4	0,8608	55,56%
#0069	3	4	0,9381	77,78%
#0074	1	2	0,6337	88,89%
#0074	1	3	0,8452	77,78%
#0074	1	4	0,8101	77,78%
#0074	2	3	0,7819	44,44%
#0074	2	4	0,7759	66,67%
#0074	3	4	0,8971	77,78%
#0091	1	2	0,9034	66,67%
#0091	1	3	0,6437	66,67%
#0091	1	4	0,9006	66,67%
#0091	2	3	0,8481	44,44%
#0091	2	4	0,9777	0,00%
#0091	3	4	0,8189	55,56%

Mondat	v1	v2	Hermes-korreláció	Szubjektív „Igen”
#0186	1	2	0,8515	44,44%
#0186	1	3	0,7416	77,78%
#0186	1	4	0,7650	66,67%
#0186	2	3	0,8877	66,67%
#0186	2	4	0,9575	33,33%
#0186	3	4	0,9108	66,67%
#0849	1	2	0,6929	77,78%
#0849	1	3	0,7921	44,44%
#0849	1	4	0,8694	55,56%
#0849	2	3	0,9327	55,56%
#0849	2	4	0,8991	22,22%
#0849	3	4	0,9406	66,67%
#1342	1	2	0,9205	55,56%
#1342	1	3	0,7346	77,78%
#1342	1	4	0,9032	55,56%
#1342	2	3	0,8172	55,56%
#1342	2	4	0,9127	77,78%
#1342	3	4	0,7591	66,67%
#1425	1	2	0,8240	66,67%
#1425	1	3	0,8310	66,67%
#1425	1	4	0,7815	77,78%
#1425	2	3	0,9546	11,11%
#1425	2	4	0,8546	88,89%
#1425	3	4	0,9040	66,67%
#1625	1	2	0,7812	44,44%
#1625	1	3	0,8299	44,44%
#1625	1	4	0,8523	77,78%
#1625	2	3	0,6547	77,78%
#1625	2	4	0,9233	66,67%
#1625	3	4	0,8081	66,67%

A kísérletet végighallgatóknak a teszt végén megjegyzések hozzáfűzésére is volt lehetőségük. Az egyik tesztelő a mondatdallambeli különbséget jóval nagyobbnak érezte azokban az esetekben, amikor a hangsúly is másik szóra került (esetleg olyan szóra, amit valójában nem is kellett volna hangsúlyozni), mint amikor a hangsúly pozíciója azonos volt a két változatban, de az alapfrekvenciában mégis jelentős különbség volt.

4 Összefoglalás

A kutatás során bemutattunk egy egyszerű módszert, amivel egy adott szöveghez különböző dallammal rendelkező mondatokat lehet szintetizálni. Ehhez egy statisztikai F0-modellt használtunk fel HMM-alapú beszédszintetizátorban. Az eredeti beszédkorpuszt az SOFM módszerrel bontottuk fel négy részre. A különböző beszédkorpuszokból betanult modellekkel eltérő dallamú mondatváltozatokat szintetizáltunk (azonos szöveghez). Ezután megvizsgáltuk a mondatváltozatok közötti különbségeket. A szubjektív kísérletek azt mutatják, hogy az alapfrekvencia eltérése a vizsgált mondatpárok felében annyira jelentős volt, hogy ez az emberi fül számára is észlelhető (azonban ez nem áll szoros összefüggésben az objektív távolságmértékkel). Ahhoz, hogy percepciós szempontból eltérő prozódiájú mondatokat tudjunk létrehozni, az szükséges, hogy az eredeti beszédkorpusz felbontása minél jobban eltérő részekre történjen, melyre a SOFM módszer alkalmasnak látszik.

A változatosabb prozódiával kiegészített beszédszintézis azokban a rendszerekben jelenthet javulást a felhasználók számára, ahol hosszabb szövegek felolvasása történik, illetve gyakran előfordulnak ismétlődő, hasonló szerkezetű mondatok. Ezek közé tartozik a könyv és az e-levél felolvasás.

A kutatást részben a TÁMOP-4.2.1/B-09/1/KMR-2010-0002 projekt támogatta.

Bibliográfia

1. Bealen, M.H., Hagan, M.T., Demuth, H.B.: Neural Network Toolbox, Revised for Version 7.0, Release 2010b, http://www.mathworks.com/help/toolbox/nnet/ (2010)
2. Csapó, T.G., Zainkó, Cs., Németh, G.: A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System. Infocommunications Journal, Vol. LXV, No.1 (2010) 32–37
3. Campillo Díaz, F., Rodríguez Banga, E.: A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. Speech Communication Vol. 48 (2006) 941–956
4. Campillo Díaz, F., van Santen, J., Rodríguez Banga, E.: Integrating phrasing and intonation modelling using syntactic and morphosyntactic information. Speech Communication, Vol. 51, No.5 (2009) 452-465
5. Hermes, D.J.: Measuring the perceptual similarity of pitch contours. Journal of Speech Language Hearing Research Vol. 41 (1998) 73–82
6. Klabbers, E., van Santen, J., Wouters, J.: Prosodic factors for predicting local pitch shape. In Proceedings 2002 IEEE Workshop on Speech Synthesis. Santa Monica, CA (2002)

7. Kohonen, T., Kaski, S., Lappalainen, H.: Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. Neural Computation Vol. 9, No. 6 (1997) 1321–1344
8. Németh, G., Fék, M., Csapó, T.G.: Increasing Prosodic Variability of Text-To-Speech Synthesizers. In: Proc. of Interspeech (2007) 474–477
9. Reichel, U.D., Kleber, F., Winkelmann, R.: Modelling similarity perception of intonation. In: Proc. of Interspeech (2009) 1711–1714
10. Rilliard, A., Allauzen, A., Boula de Mareüil, P.: Using Dynamic Time Warping to compute prosodic similarity measures. In: Proc. of Interspeech (2011) 2021–2024
11. Székely, E., Cabral, J. P., Cahill, P., Carson-Berndsen, J.: Clustering expressive speech styles in audiobooks using glottal source parameters. In: Proc. of Interspeech, (2011) 2409–2412
12. Tóth B.P., Németh G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 246–256
13. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system version 2.0. In: Proc. of ISCA SSW6 (2007)

A szintaktikai szerkezet automatikus feltérképezése a beszédjel prozódiai elemzése alapján

Szaszák György[1], Beke András[2]

[1] BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
[2] MTA Nyelvtudományi Intézet, Fonetikai Osztály
E-mail: szaszak@tmit.bme.hu; beke.andras@gmail.com

Kivonat A prozódia és a szintaktikai szerkezet közötti összefüggés aligha kérdéses, hiszen számos kutatás foglalkozott már kapcsolatukkal, illetve ezt az összefüggést számos beszédtechnológiai – elsősorban beszéd szintézisét célzó - alkalmazásban ki is használják. Az általánosan elfogadott álláspont szerint a prozódiai és a szintaktikai szerkezet szorosan összefügg ugyan, közöttük a kapcsolat azonban nem egy-egyértelműen meghatározott. Mindenesetre gyakorlati alkalmazások bizonyítják, hogy a szintaktikai elemzés alapján a prozódia jól előrejelezhető és kiválóan előállítható beszédszintetizátor alkalmazásokban. A prozódia és a szintaxis közötti összefüggés másik irányát azonban – nevezetesen a szintaxis visszakövethetőségét prozódiai jegyek alapján – eddig kevesen vizsgálták, illetve ha mégis, ezen vizsgálatok jellemzően minimál mondatpárok prozódia alapján történő elkülöníthetőségére vonatkoztak. Bár e vizsgálatok értékét nem szeretnénk alábecsülni, hiszen fontos elméleti jelentőségük van, eredményeik a gyakorlati alkalmazásokat tekintve azonban csak elvétve, nem igazán életszerű körülmények között lennének felhasználhatók. Cikkünkben ezért arra keressük a választ, hogy lehetséges-e a prozódiai szerkezet feltárása alapján szintaktikai szerkezetre vonatkozó információ kinyerése általánosabb, a mindennapi élethez jobban köthető tematika esetében. Miután a kutatás célja az automatikus elemezhetőség vizsgálata, ezért a prozódiai szerkezet elemzését is automatikus eszközökkel valósítjuk meg. Eredményeink tanúsága szerint a beszédben a szintaktikai frázisok jelentős része jól beazonosítható, ráadásul, a szintaktikai hierarchia magasabb szintjein jól el is helyezhető. Mélyebb szinteken - többszörös beágyazásban - pontos szintaktikai szintbeli elhelyezést nem várhatunk a prozódiától, a határok jelzése azonban megmaradhat.

Kulcsszavak: prozódia, szintaktikai elemzés, prozódiai szegmentálás, szintaktikai hierarchia, prozódiai hierarchia, szintaxis-fonológia interfész

1. Bevezetés

A prozódia és a szintaktikai szerkezet közötti összefüggést számos megközelítésben vizsgálták már, a szintaktikai és a fonológiai reprezentáció közötti interfészt

azonban eddig nem sikerült egységesen leírni. Ez nem meglepő, hiszen összetett jelenséggel állunk szemben, így az egységes modell megalkotása nem is feltétlenül volna megvalósítható elképzelés. Mindenesetre az eddigi kutatások néhány főbb ponton összecsengenek, így a szintaktikai és prozódiai szerkezetek közötti összefüggés általánosan elfogadott, természetét tekintve azonban nem teljesen feltárt. Az egyik legismertebb hipotézis Selkirk nevéhez fűződik (*prosodic structure hypothesis*), mely szerint egy-egy mondat prozódiai szerkezete nagyban - de nem teljes mértékben - függ a felszíni szintaktikai szerkezettől [11]. Más szerzők viszont amellett érvelnek, hogy a prozódiát közvetlenül és többnyire egyértelműen a szintaktikai szerkezet határozza meg [5]. A szerzők tapasztalatai alapján ez utóbbi megállapítás túlzottnak tűnik, ugyanakkor az idézett elméletek nem térnek ki arra, hogy a prozódiai, illetve szintaktikai hierarchiában magasabban elhelyezkedő szintek sokkal biztosabban, míg a mélyebbek esetlegesebben feleltethetők meg egymásnak.

A prozódiai szerkezet az általánosan elfogadott hipotézisek szerint ([11], [4]) felülről lefelé haladva az alábbiak szerint alakul: a megnyilatkozás (utterance) *intonációs frázis*okból áll (IF), amelyek tovább bonthatók az ún. *fonológiai frázis*okra (FF). A fonológiai frázisokat pedig *fonológiai szavak* (FSz) építik fel, ezeket gyakran prozódiai szónak is hívják [11]. A hierarchia tovább finomítható egészen a szótag szintig, de a fonológiai frázisnál mélyebb egységeket a cikkben nem fogjuk használni, így a további ismertetéstől eltekintünk. A prozódiai szerkezet jól szemléltethető fával vagy a hierarchiát tükröző zárójelezéssel.

A mondatok szintaktikai elemzésekor hasonló hierarchiában gondolkodunk, amely az alapvető építőelemeket (pl. szavak) kapcsolja össze mondatokká: az egyes szavak szószerkezeteket alkotnak, ezek a szintaktikai frázisok (SzF). Az egyes frázisokba további frázisok ékelődhetnek (embedding), létrehozva a szintenként reprezentálható hierarchiát. A szintaktikai frázist általában domináns eleme (ún. fej) után nevezik el. A domináns elem az az elem, amely a frázis viselkedését az eggyel magasabb szintaktikai szinten meghatározza. Ily módon beszélhetünk névszói frázisokról (a fej névszó), igei és határozói stb. frázisokról. A szintaktikai elemzés során elterjedt a fareprezentáció.

A beszédtechnológiában az írott mondatok szintaktikai elemzése beszédszintézis előtt elterjedt technológia [6]. Az első ilyen irányú próbálkozások egészen az 1980-as évekig nyúlnak vissza. A módszer alapja az a feltételezés, hogy a szintaktikai elemzés alapján az előállítandó beszéd prozódiai jellegzetességei igen jól előrejelezhetők. Ez tehát azt jelenti, hogy a felszíni szintaktikai szerkezet leképezhető a prozódiai szerkezetre, ráadásul a gyakorlati tapasztalatok alapján igen biztosan. Teljes leképezhetőségről azonban a beszédszintézis esetén sem beszélhetünk, részben éppen ezzel magyarázható, hogy a beszédszintézis alkalmazásokat miért érdemes egy-egy behatárolt tématerületre szűkíteni a minőség javítása érdekében [12].

A fordított irányú leképezés, azaz a prozódia alapján a szintaktikai viszonyokra való következtetés jóval kevésbé elterjedt, néhány – igaz, leginkább kutatási, kísérleti, de kevésbé gyakorlati – alkalmazásban azonban találkozhatunk vele. Több kutatásban is vizsgálták például egymástól jelentésben és/vagy ta-

golásban, írásjelezésben különböző, de a felépítő szavakat tekintve megegyező, ún. minimál mondatpárok elkülöníthetőségét prozódia alapján [9] (lényegét tekintve tehát jelentés-egyértelműsítés céljából). Az idézett tanulmányban Price és munkatársai következtetései alapján a prozódia alapján többségében jól elkülöníthetőek voltak a minimál párok, néhány kivételtől eltekintve. Munkájukban javaslatot is tettek olyan automatikus prozódiacímkézőre, amely normalizált időtartamadatok alapján szünetek osztályozására volt alkalmas. A prozódia alapján végzett egyértelműsítést vizsgálták már beszédfelismerésben is, leginkább itt is minimál párok elkülöníthetőségét célozva.

A beszédalapú egyértelműsítési feladatokban az előbbiekben bemutatott minimál páros szemléltetés a legelterjedtebb, pedig az ily módon konstruált mondathalmazok gyakran mesterkéltek, gyakorlati alkalmazásban ritkán, de semmiképp sem univerzálisan használhatók. Ezért jelen kutatásban arra helyeztük a hangsúlyt, hogy amennyire lehetséges, általános célú és általánosan felhasználható eszközt dolgozzunk ki. Az alkalmazott megközelítés az automatikus szintaktikai és a prozódiai elemzések összevetése lesz, általános, relatíve nagy méretű beszédkorpuszon. A vizsgálat arra keresi a választ, hogy lehetséges-e a szintaktikai szerkezet legalább részleges, illetve minél teljesebb feltárása a beszédjel prozódiai elemzése alapján. Ha igen, mennyire megbízható ez az elemzés, lehetséges-e a szintaktikai hierarchia felállítása is? A kísérlethez automatikus prozódiai elemzőt használunk [13], így a lehetőségeket azonnal az automatikus elemezhetőség jelentette korlátok között értékeljük.

Cikkünk felépítése az alábbiak szerint alakul: elsőként bemutatjuk a prozódiai elemzést és a szintaktikai elemzést, a beszédkorpuszt. Ezt követi a kísérleti feltételek részletes leírása, a kiértékeléshez használt mérőszámok bemutatása, az eredmények ismertetése és a következtetések származtatása.

2. Beszédjel automatikus prozódiai szegmentálása

A prozódiai szerkezet feltérképezésére a beszédjelen *prozódiai szegmentálás*t végzünk. Az eljárást részletesen bemutattuk már [14], [13], így itt csak a lényegesebb jellemzőit foglaljuk össze. A prozódiai szegmentáló feladata fonológiai frázisok (FF) illesztése a beszédjelhez. Ehhez a szegmentáló 7 beépített fonológiai frázismodellt tárol rejtett Markov-modell formájában (lásd 1. táblázat). Az illesztés a hangsúlyok és a dallammenetek együttes figyelembevételével történik. A felhasznált akusztikai jellemzők az alapfrekvencia- és az energiamenet, kinyerésüket a következő, 2.1 alfejezetben röviden áttekintjük. A fonológiai frázisokra úgy tekintünk, mint a legkisebb, önálló hangsúllyal és dallammenettel jellemezhető egységre [4]. A magyar nyelvben kijelentő módban a tipikus FF elején a hangsúlynak megfelelő kiemelést tapasztalunk, amelyet lassan ereszkedő dallammenet követ a következő hangsúlyos egységig. Ezt tekintjük a FF prototípusának (*fs*). Mivel azonban a fonológiai frázisok intonációs frázisokba, illetve megnyilatkozásegységekbe - olvasott beszédben mondatokba, spontán beszédben virtuális mondatokba - szerveződnek, magasabb szintű tényezők is befolyásolják a hangsúlyozást és a dallammenetek alakulását. Emiatt az osztályozáshoz/illesztéshez

további FF-ok elkülönítése szükséges: a tagmondat eleje (*me*) és a tagmondat vége (*mv*) jellemzően befolyásolja a FF prototípusát, akárcsak a fókusz (*fe*) és a folytatást jelző dallammenet-emelkedés (*fv*). Ez utóbbi a következő fonológiai frázist olykor inverz hangsúlyba fordítja, azaz kiemelkedés helyett a prozódiai jellemzők lokális minimumot adnak (*s*). A prozódiai szegmentáló kimenetén tehát az illesztett fonológiai frázisok jelennek meg kezdő- és végidőpontjaikkal.

1. táblázat. A prozódiai szegmentáláshoz modellezett fonológiai frázistípusok.

Címke	FF típus
me	Tagmondat eleje
fe	Erős hangsúly
fs	Prototípus
mv	Tagmondat vége
fv	Folytatást jelző
s	Inverz hangsúly
sil	Csend

A prozódiai szegmentálás során a fonológiai frázisok egymáshoz kapcsolódási szabályszerűségeit leíró, prozódiai-nyelvi jellegű modellt is használunk. Ez a modell teszi lehetővé egyrészt az illesztést (milyen FF milyen FF után milyen valószínűséggel következhet), másrészt előkészíti a szintaktikai szerkezetre való leképezést, hiszen a prozódiai szegmentáló FF-modelljei a mondatokban, tagmondatokban elfoglalt helyük, szerepük szerint lettek kialakítva. A használt modell éppen a mondatok, virtuális mondatok (idealizált) felépítését adja meg: minden mondat tagmondat eleje frázissal (*me*) indít és tagmondat vége frázissal (*mv*) zár. Közben erősen (*fe*) és közepesen hangsúlyos (*fs*, prototípus) fonológiai frázisok tetszőleges sorrendben váltakoznak, esetleges folytatást jelző frázisokkal (*fv*). Ez utóbbit tagmondat eleje frázis (*me*) vagy inverz hangsúlyt tartalmazó frázis (*s*) követheti. Kivételes esetben mondat vége is lehet (pl. kérdés esetén). A mondatok között szünetet feltételezünk (*sil*). Fontosnak tartjuk megvilágítani, hogy az alkalmazott illesztési eljárás nem pusztán egyes prozódiaeseményhez köthető jelölők (pl. szünetjelölők, hangsúlyjelölők) detektálásán alapul (vö. ToBI, [12]), hanem a prozódiai, illetve a hozzá társított akusztikai jellemzők folyamatos követését biztosítja, ily módon véleményünk szerint rugalmasabb és egységesebb prozódiai szegmentálást tesz lehetővé, lényegében az egyes detektálandó eseményeket a fonológiai frázisok modelljei inkorporálják.

2.1. Akusztikai-prozódiai előfeldolgozás

Az akusztikai-prozódiai előfeldolgozás a [13] irodalomban ismertetettek alapján történik, de az egyes jellemzők kinyerésénél használt konstansok értékeit az alábbiak szerint állítottuk be: az alapfrekvencia (F_0) kinyerése ESPS algoritmussal történik 25 ms hosszúságú, csúsztatott ablakolással. Az energia kinyeréséhez használt ablak is 25 ms. A keretidő mindkét jellemzőre 10 ms. A nyert

alapfrekvencia-menetet ezután oktávugrásoktól szűrjük, majd 5 pontos átlagoló szűrővel simítjuk. Ezután az alapfrekvenciát logaritmikus tartományban lineárisan extrapoláljuk a zöngétlen helyeken, de csak akkor, ha a zöngétlen szakasz nem hosszabb 150 ms-nál és ha a zöngétlen szakasz után az alapfrekvencia nem indul túl magasról (nem emelkedhet többet 10%-nál a zöngétlen szakasz előttihez képest. Erre azért van szükség, hogy a frázisok közötti, levegővétellel nem társuló szünetet nehogy zöngétlen beszédhangszakasznak vegyük. Az így előfeldolgozott jellemzőkhöz delta és delta-delta együtthatókat fűztünk. Az előfeldolgozás minden egyéb tekintetben azonos a [13] irodalomban bemutatottal.

2.2. Prozódiai szegmentálás és szóhatárok detektálása

Korábbi munkákban [13] [14] [2] vizsgáltuk már a szóhatárok detektálhatóságát prozódiai jellemzők segítségével. Ennek egyik útja szintén a fonológiai frázisok illesztése volt, amely magyar nyelvre a kötött hangsúlyozást kihasználva szóhatárok detektálását tette lehetővé, 77% körüli pontossággal és 57% körüli hatékonysággal magyar nyelvre, 69% körüli pontossággal és 76% körüli hatékonysággal pedig finn nyelvre. A szóhatár-detektálás vizsgálatakor nem végeztünk szintaktikai elemzést, viszont hipotézisünk, hogy a szintaktikai frázisok határa eső szóhatárok jobban detektálhatók, mint a frázisok belsejébe esők (igaz, a szóhatárok jelentős részén szintaktikai frázis határa is van). A szóhatár-detektálás elsősorban a gépi beszédfelismerést segítette, míg a szintaktikai elemzés – ha lehetséges a prozódia alapján – a beszéd gépi elemzését teheti lehetővé, amely kiemelt fontosságú az átfogóbb, gépi beszédértést/-elemzést is igénylő rendszerekben (pl. gépi tolmácsolás).

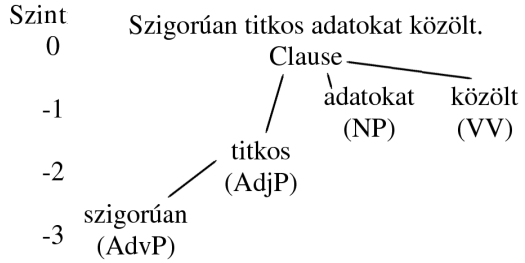
3. Szintaktikai elemzés (szövegalapú)

A szintaktikai elemzéshez a szabadon elérhető HunPars eszköz állt rendelkezésünkre [1]. Ez a szintaktikai elemző belső erőforrásokként ún. frázisstruktúra nyelvtant és lexikai adatbázisokat és a HunMorph morfológiai elemzőt használja fel. A morfológiai elemző használata nagyban emeli a komplexitást, de a magyar nyelv sajátosságai miatt aligha megkerülhető. A szintaktikai elemző kimenetén az elemzett mondat tagekkel ellátva és a szintaktikai hierarchiában elfoglalt helyzetet tükröző zárójelezéssel jelenik meg, amelyből fastruktúrájú reprezentáció is generálható. Az elemző minden lehetséges hipotézist megad, ez hosszabb mondatokra több tíz, kirívó esetben néhány száz lehetséges elemzés is lehet. Miután a prozódiát ezúttal nem egyértelműsítésre kívánjuk felhasználni, az elemzéseket szakértő kézzel egyértelműsítette. Az egyértelműsítés egyébként leginkább egyes lexikai elemek több jelentéséből adódóan vált szükségessé (pl. az 'egy' szót mindig háromféle elemzésben kapjuk meg (határozó, névelő, számnév), ha több nem egyértelmű elem is van a mondatban, akkor a hipotézisek száma összeszorzódik).

4. Anyag és módszer

A kísérleti anyagot a BABEL magyar nyelvű adatbázis [10] szolgáltatta, amely 5-7 mondatból álló bekezdéseket is tartalmaz. Ebből 330 mondatot elemeztünk (az ismétlődések miatt 155 különböző mondatot kellett csak szintaktikailag elemezni) 60 beszélőtől (30 férfi, 30 nő). Elsőként a 155 különböző mondat szintaktikai elemzését végeztük el. Ezután az egyes felvételeket beszédhang szinten szegmentáltuk kényszerített illesztéssel. A beszédhang szintű szegmentálásból kinyertük az egyes szintaktikai egységek határához köthető időpontokat. Ezeket fogjuk a prozódiai szegmentálás eredményeként előálló fonológiai frázisok határaival összevetni. Az összevetést szintaktikai szintenként végezzük elkülönítve, mivel a hipotézisünk az, hogy a magasabb szintaktikai szinteket a prozódia jobban tükrözi. Az elkülönített szintaktikai szinteket számmal jelöltük, felülről lefelé haladva: 0, −1, −2, −3, −4 (vö. 1. ábra). A mondatokat tagmondatokra bontjuk, így kapjuk a 0. szintet. A tagmondatokat szintaktikailag tovább elemezve egymásba ágyazott szintaktikai frázisokat találunk, ezeket reprezentálják a negatív számmal jelölt szintek. Míg a beágyazást nem tartalmazó szintaktikai frázisok (−1. szint) és az egyszeres beágyazást tartalmazók (−2. szint, a legtöbb jelzős szerkezet ilyen) igen gyakoriak, addig kétszeres vagy többszörös beágyazás már viszonylag ritka (lásd a 2. és 3. táblázatokban).

A szintaktikai és a fonológiai frázishatárokat akkor tekintettük egybeesőnek, ha közöttük kezdőidőpontjukat (végidőpontjukat) tekintve 150 ms-ot meghaladó időbeli eltérés nem volt. Ezt a tűrésintervallumot az alábbi megfontolások alapján jelöltük ki: (i) az intervallumnak lehetővé kell tennie kb. fél szótag nagyságrendű eltérést, mert a prozódiai szegmentáló pontossága ilyen nagyságrendű (vö. [13]), illetve (ii) mert a referenciaként vizsgált szintaktikai határokat automatikus szegmentálással határoztuk meg, ami pontatlanabb szegmentálást jelent a kézzel végzettnél. A választott tűréshatáron belül így még biztosított, hogy (iii) a prozódiai szegmentáló által illesztett fonológiai frázisok várható hossza jóval nagyobb 150 ms-nál (a vizsgált korpuszon az átlagos FF-hossz 618 ms, 211 ms szórás mellett). A fonológiai frázisok kezdetét mindig a szintaktikai frázisok kezdetével, a FF-ok végeit mindig a SzF-ok végével vetettük össze.



1. ábra. Szintaktikai szintek hierarchikus reprezentációban

5. Eredmények és értékelés

5.1. Szintaktikai frázisok behatárolása

Az első kísérlet arra irányult, hogy megvizsgáljuk, a szintaktikai frázisok mennyire határolhatók be a prozódia alapján. Mérőszámként a *recall* értéket használjuk, mely definíció szerint:

$$Recall = \frac{tp}{tp + fn}, \tag{1}$$

itt tp a helyesen azonosított szintaktikai határok száma (true positives), fn pedig a nem azonosított szintaktikai határok száma (false negatives). Az eredmények a 2. táblázatban láthatók, külön a frázisok kezdetére és a végére. Már említettük, hogy a kiértékelést szintaktikai szintekre bontva külön-külön végezzük. Egyes esetekben (nem is ritkán) több, különböző szintű szintaktikai határt találunk egy helyen (pl. az "ügyes ember" frázisban egyszerre indul a -1. és a -2. szintű szintaktikai frázis, a -2. szintű az "ügyes", az -1. szintű az "ember" után ér véget). A kiértékelést emiatt két szálra bontottuk: az egyik szálon egy helyen egyetlen, a legmagasabb szintű szintaktikai határt számoljuk csak (erre az 1B/W jelöléssel utalunk), míg a másik szálon az egy helyen található valamennyi szintaktikai határt egyszerre figyelembe vesszük (tehát utóbbi esetben helyes detektálás esetén valamennyi szinten egy helyes detektálást számítunk, holott "több legyet ütöttünk egy csapásra". Ugyanakkor ha elvétjük a határt, akkor azt természetesen valamennyi szinten hibaként vesszük figyelembe. Erre a számítási módra a MB/W jelöléssel utalunk).

2. táblázat. Szintaktikai frázisok határainak detektálása (recall). 1B/L= egy (a legmagasabb szintű) szintaktikai határ egy helyen; MB/W= több szintaktikai határ is lehet egy helyen.

Szintaktikai szint	Kezdet		Vég		Előf. száma (MB/W)
	1B/W	MB/W	**1B/W**	MB/W	
0	**0,85**	0,85	**0,79**	0,79	3124
-1	**0,45**	0,70	**0,48**	0,68	10339
-2	**0,42**	0,70	**0,48**	0,69	5763
-3	**0,44**	0,74	**0,45**	0,65	814
-4	**0,48**	0,70	**0,50**	0,67	187
Összes szint	**0,54**	0,72	**0,55**	0,69	20227

Az átlagos recall érték 71% (MB/W), illetve 55% (1B/W), amely a tagmondatok szintjén jelentősen magasabb: 85% (fráziskezdet) és 79% (frázisvég). Az eredmények statisztikai alátámasztására Kruskal-Wallis próbát végeztünk, amely igazolta, hogy a fonológiai és a szintaktikai frázisok között szignifikáns összefüggés van ($\chi^2 = 6430,606; p < 0,000$).

A megfelelő SzF kezdő- és végidőpontokat párba állítva és a recall értékeit vizsgálva Mann-Whitney és Wilcoxon W tesztekkel a tagmondatok esetén a tagmondat kezdetét szignifikánsan jobban lehet detektálni, mint a végét ($Z = -7,807; p < 0,000$). Mélyebb szintaktikai szinteken azonban megszűnik a szignifikáns különbség a kezdő és végidőpontok tekintetében (-1. szint: $Z = -0,407; p > 0,1$; -2. szint: $Z = -0,016; p > 0,1$; hasonlóan a mélyebb szintekre is).

A tagmondat szintnél mélyebb szinteken a recall értékek szinte azonosak, ebből arra következtethetünk, hogy a prozódia a szintaktikai hierarchiában elfoglalt helyzettől függetlenül jelez szintaktikai frázishatár-információt: nincs szignifikáns különbség a recall értékek között a szintaktikai szint függvényében a tagmondatnál mélyebben: ($\chi^2 = 0,224; p > 0,1$). Tehát minden SzF önálló entitásként viselkedik, függetlenül a szintaktikai hierarchiában elfoglalt helyétől.

5.2. Szintaktikai szintek elkülönítése a prozódia alapján

A következő lépésben azt vizsgáltuk, mennyire különíthetők el az egyes szintaktikai szintek a fonológiai frázisokra történő szegmentálás alapján, illetve van-e olyan FF, amely valamely szintaktikai szinthez társítható (a frázistípusok elkülönítésénél használt metodika alapján hipotézisünk, hogy kell lennie). Ha a FF típusa alapján különbséget tudunk tenni a szintaktikai szintek között, az nagyban emelné a prozódiai szegmentálás értékét az elemzésben. Azt is jó lenne tudnunk, mennyire megbízható a detektálás az egyes fonológiai frázisok típusától függően (ha van közöttük különbség). A választott mértékünk a precision:

$$Precision = \frac{tp}{tp + fp}, \tag{2}$$

ahol tp ismét a FF-ok által helyesen (150 ms-on belül) jelzett SzF határ, míg fp a beszúrt FF határok száma (amelyek tehát nem esnek egybe SzF-sal). A precision mérőszám mellett specificitás jelleggel azt is vizsgáljuk, hogy fonológiai frázistípusokra bontva hogyan alakulnak a szintenkénti relatív gyakoriságok (milyen típusú FF milyen szintű SzF-nak felel meg leggyakrabban/tipikusan). Az eredményeket a 3. és a 4. táblázatokban mutatjuk be, külön frázisok elejének és végének összehasonlítására. A relatív gyakoriságok mellett az utolsó oszlopban a FF-hoz tartozó precision értéke is megtalálható.

A 3. táblázat eredményei szerint a *me* FF 86% relatív gyakorisággal tagmondat kezdetét jelöli. A -1. szintű szintaktikai frázis kezdetére a *fe*, *fs*, *mv*, illetve kisebb mértékben a *fv* fonológiai frázisok utalnak. Az *s* típusú frázis kezdete nem egyértelmű szintaktikai utalás szempontjából. A -2. szintaktikai szinttől mélyebben a FF-ok eloszlása lényegében egyenletes az egyes szintek között, így a FF típusa nem utal a szintaktikai szintre. Az eredmények összességében tehát azt jelentik, hogy a tagmondatok kezdete igen biztosan előrejelezhető a FF típusa alapján (0. szint), illetve hogy a -1. szint ettől és a mélyebben fekvő szintektől még jól elkülöníthető. Tehát a szintaktikai hierarchia prozódiai szemszögből 3 szintre tagolódik, a 0. szintaktikai szintre, a -1. szintaktikai szintre és

3. táblázat. SzF szintek és FF-ok típusának kapcsolata frázisok elején (relatív gyakoriságok) és precision.

FF típusa	Szintaktikai szint 0	−1	−2	−3	Előfordulások száma (összes)	Precision
me	**0,86**	0,07	0,04	0,02	1736	**0,84**
fe	0,12	**0,78**	0,07	0,02	2517	**0,58**
fs	0,09	**0,83**	0,06	0,01	1399	**0,55**
mv	0,14	**0,80**	0,04	0,02	2094	**0,46**
fv	**0,22**	**0,72**	0,04	0,01	1326	**0,51**
s	**0,50**	**0,41**	0,07	0,02	1456	**0,57**
Összes FF	0,36	0,56	0,05	0,02	10539	**0,58**

az összevont $-2.-N$. mélyebb szintekre. Arra is tekintettel, hogy a szintaktikai hierarchiában a mélyebb szintek felé haladva a SzF előfordulások gyakorisága radikálisan csökken, tehát igen ritkák a kettőnél többször beágyazott frázisok (vö. 2. táblázat), a fonológiai frázis segítségével behatárolt szintaktikai frázisok jelentős hányadáról tehát eldönthető, hogy nagy valószínűséggel milyen szinthez tartoznak. Az összes −2. szintű és mélyebben elhelyezkedő frázis valójában több mint 85%-ban −2. szintű frázisnak felel meg, csak a fennmaradó szűk 15% az ennél mélyebb szinten elhelyezkedő. Közöttük viszont a prozódia alapján különbséget nem tudtunk tenni.

4. táblázat. SzF szintek és FF-ok típusának kapcsolata frázisok végén (relatív gyakoriságok) és precision.

FF típusa	Szintaktikai szint 0	−1	−2	−3	Előfordulások száma (összes)	Precision
me	0,05	**0,74**	0,11	0,08	1736	**0,58**
fe	0,09	**0,68**	**0,20**	0,03	2517	**0,64**
fs	0,08	**0,68**	**0,18**	0,04	1399	**0,60**
mv	**0,83**	0,11	0,04	0,02	2094	**0,80**
fv	**0,60**	**0,28**	0,09	0,03	1326	**0,73**
s	0,13	**0,64**	**0,17**	0,06	1467	**0,57**
Összes típus	0,34	0,49	0,13	0,04	10593	**0,66**

A 4. táblázat eredményei szerint a frázisok végét vizsgálva a detektált *mv* típusú FF 83% relatív gyakorisággal jelezte a 0. szintű tagmondat végét. Az *fv* típusú FF gyakran (60%) szintén tagmondat végét jelzi (0. szint), azonban viszonylag gyakran jelezheti −1. szintű szintaktikai frázis végét is (28%). Az *me* típusú FF vége egyértelműbben a −1. szinthez kapcsolható 74% gyakorisággal, míg az *fe*, *fs* és *s* típusú FF-ok vége −1. vagy −2. szinten jelzi a SzF-ok végét. Ellentétben a frázisok elejére végzett vizsgálatokkal, a frázisok végét vizsgálva

már a −1. és a −2, illetve mélyebb szintek sem különíthetők el az illesztett FF típusa alapján a relatív gyakoriságok vizsgálatával. Ehhez tehát a frázisok elejét kell vizsgálnunk. A gyakorlatban természetesen a frázisok elejét és végét együttesen tudjuk vizsgálni az esetek döntő többségében, hiszen a frázisok végén rendszerint újabb frázisok kezdődnek (kivéve a megnyilatkozás végén és hosszabb csend előtt, bár ez utóbbi szintén informatív elem, hiszen előtte – legalábbis olvasott beszédben – a szintaktikai frázis, sőt a tagmondat is általában lezárt.

A precision és recall mérőszámok értékeit redukált FF elemhalmazzal is számítottuk annak vizsgálatára, hogy ily módon esetleg egyértelműbben lehetségese a szintaktikai szintek elkülönítése. A redukált FF halmazzal történő vizsgálat során a prozódiai szegmentáló nem illesztheti az *fs* és az *s* FF-okat. Utóbbit azért zárjuk ki, mert a frázisok elejére végzett vizsgálatkor nem jelezte egyértelműen a szintaktikai szintet, előbbit pedig azért, mert szerepét várhatóan az erősebben hangsúlyos, de dallammenetben nem különböző *fe* típusú FF modellje részben átveheti. A redukált FF elemhalmazzal végzett vizsgálatok eredményei a frázisok elejét vizsgálva az 5. táblázatban láthatók. A recall értéke visszaesik (átlagosan 48%-ra, 1B/W esetben), tehát a redukált elemhalmazzal kevesebb szintaktikai frázis kezdetét tudjuk meghatározni, ugyanakkor a precision értéke szignifikánsan nem változik. Ami miatt mégis érdemes lehet a vizsgálatot elvégezni, hogy a 0., tagmondat szintet sokkal biztosabban kiemeli. A frázisok végét vizsgálva hasonló eredményeket kaptunk: gyengébb recall mellett szignifikánsan nem jobb precision, a 0. és a −1. szintek elkülöníthetősége javul, a −2. szintet pedig érdemben már nem detektálja a rendszer.

5. táblázat. SzF szintek és FF-ok típusának kapcsolata frázisok elején redukált FF elemhalmazzal (relatív gyakoriságok); precision és 1B/W recall az egyes szintaktikai szintekre.

FF típusa	Szintaktikai szint 0	−1	−2	−3	Előfordulások száma (összes)	Precision
me	0,88	0,07	0,02	0,02	1835	**0,92**
fe	0,13	0,77	0,07	0,02	3455	**0,58**
mv	0,26	0,67	0,04	0,02	1914	**0,53**
fv	0,37	0,58	0,04	0,01	1782	**0,57**
Összes típus	0,42	0,51	0,05	0,02	8986	**0,64**
Recall	0,80	0,39	0,34	0,37	Átl. recall: 0,48	

5.3. Összefüggés a fonológiai és a szintaktikai frázis típusa között

Végezetül azt is vizsgáltuk, hogy felfedezhető-e valamiféle összefüggés a fonológiai frázis típusa (me, fe, fs, mv, fv, s), illetve a szintaktikai frázis típusa között (NP, AdjP, AdvP, NumP, VV, VV-Inf, PostpP). Az eredmények tanúsága szerint ilyen összefüggés a magyar nyelvben nem mutatható ki ($\chi^2 = 0,349; p > 0,1$),

a fonológiai frázisok véletlenszerűen kombinálódnak a szintaktikai frázisokkal. A frázistípusok össze nem függése a magyar nyelvben a kötetlen szórend miatt nem meglepő, a vizsgálatot érdemes lenne más, a szemantikai összefüggéseket szórenddel érzékeltető nyelven is elvégezni.

6. Összefoglalás és kitekintés

Cikkünkben a szintaktikai szerkezet feltérképezhetőségét vizsgáltuk olvasott beszédben. Egy prozódiai szegmentáló kimenete alapján a szintaktikai frázisok határait azonosítottuk, és vizsgáltuk a szintaktikai hierarchiához rendelt szintek visszakövethetőségét is pusztán a beszédjel prozódiája alapján. A tagmondathatárok akár 92%-a, a tagmondatban elhelyezkedő, akár egymásba is ágyazott szintaktikai frázisok határainak 50-70%-a volt automatikusan meghatározható. A tagmondathatárok detektálásában a pontosságot jellemző precision mérőszám maximális értéke 84% volt, a beágyazott szintaktikai frázisokra 46 és 58% között alakult. Végkövetkeztetéseink az alábbiak: a prozódia olvasott beszédben (i) a szintaktikai határokat jól jelzi, (ii) többnyire világosan elkülöníti a tagmondathatárokat a szószerkezetek határaitól, (iii) a FF-ok/SzF-ok elejét összevetve az egyszeres beágyazódások még esetenként megkülönböztethetők (-1. és -2. szintek elkülönítése), a mélyebb szintaktikai szintek viszont egybeolvadnak, határaik azonban esetenként továbbra is detektálhatók. Ezek alapján a prozódia ütemező, szinkronizáló szerepe feltételezhető a humán beszédpercepcióban, amelyet szerényebb rétegző szerep egészít ki (0., -1. és -2. és mélyebb szintek elkülönítése).

A prozódiai és szintaktikai szerkezet összefüggéseit spontán beszédben is vizsgáljuk, ezek a kísérletek azonban még folyamatban vannak – reményeink szerint előadásunkban már az eredményekből is ízelítőt adhatunk. Spontán beszéd esetében a prozódiai szegmentálás nagyjából elvégezhető, ugyanakkor számolni kell az elemzést megnehezítő elemek megjelenésével: érzelmi töltet, amely a prozódiát is befolyásolja; nagyobb dinamikatartomány (ez az előfeldolgozásban - oktávugrás elleni szűrésben és interpolálásban - okozhat nehézségeket; a hangsúlyozási-hanglejtési "szokásjog" gyakori megszegése, dinamikus változása). A spontán beszéd szintaktikai elemzése igen nehéz feladatnak bizonyul, mivel nem tartalmaz jól körülhatárolható, egyértelműen meghatározható mondatokat. Áthidaló megoldásként ún. virtuális mondatok elemzését fogjuk elvégezni (ez alatt a spontán beszédbeli megnyilatkozások olvasott beszédhez hasonló mondatszerű formára konvertált alakját értjük - vö. [3], [7]). Továbbra is problémát jelentenek azonban a megakadásjelenségek, befejezetlen gondolatok stb., amelyek a prozódiai és virtualizált szintaktikai szerkezet egymásra képezését jelentősen nehezíthetik.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki Nagy Katalinnak, a BME villamosmérnök hallgatójának a bemutatott munkában nyújtott segítségéért.

Hivatkozások

1. Babarczy A., Bálint G., Hamp G., Kárpáti A., Rung A., Szakadát I.: Hunpars: mondattani elemző alkalmazás, III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Magyarország, 2005. pp. 20-28.
2. Beke András, Szaszák György: Szótagok automatikus osztályozása spontán beszédben spektrális és prozódiai jellemzők alapján, VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Magyarország, 2010. pp. 236-248.
3. Gósy Mária: Virtuális mondatok a spontán beszédben, Beszédkutatás 2003, MTA Nyelvtudományi Intézet, Budapest, 2003. pp. 19-43.
4. Hunyadi László: Hungarian Sentence Prosody and Universal Grammar, Peter Lang, 2002.
5. Kaisse, Ellen M.: Connected Speech: The Interaction of Syntax and Phonology, Academic Press, San Diego, 1985.
6. Koutny Ilona: Parsing Hungarian Sentences in order to Determine their Prosodic Structures in a Multilingual TTS system, Proc. of the Eurospeech'99 International Conference on Speech Communication and Technology, pp. 2091-2094, Budapest, Hungary, 1999.
7. Markó Alexandra: A spontán beszéd néhány szupraszegmentális jellegzetessége: Monologikus és dialogikus szövegek összevetése, valamint a hümmögés vizsgálata, PhD értekezés, ELTE, Budapest, 2005.
8. Olaszy Gábor, Németh Géza, Olaszi Péter: Automatic Prosody Generation - a Model for Hungarian, In: European Conference on Speech Communication and Technology (Eurospeech 2001). Aalborg, Dánia, 2001. pp. 525-528.
9. Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., Fong. C.: The use of prosody for syntactic disambiguation, Journal of the Acoustical Society of America 90(6):2956-2970, 1991.
10. Roach, P. et al.: BABEL: An Eastern European multi-language database, Proc. of the 4^{th} International Conference on Speech and Language Processing, Philadelphia, USA, Vol 3. pp. 1892-1893, 1996.
11. Selkirk, Elisabeth: The Syntax-Phonology Interface, in Smelser, N.J. and Baltes, Paul B. [Eds], International Encyclopaedia of the Social and Behavioural Sciences, 15407-15412, Oxford: Pergamon, 2001.
12. Silverman, K.: On costumizing prosody in speech synthesis: names and addresses as a case in point, in Proc. ARPA Workshop on Human Language Technology, pp. 317-322, 1993.
13. Szaszák György: A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben, PhD értekezés. Budapesti Műszaki és Gazdaságtudományi Egyetem, 2008.
14. Vicsi Klára, Szaszák György: Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján, II. rész: Statisztikai eljárás, finn-magyar nyelvű összehasonlító vizsgálat, III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Magyarország, 2005. pp. 360-370.

A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei. Automatikus prozódiai annotáció

Szekrényes István[1], Csipkés László[1], Oravecz Csaba[2]

[1] Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
H-4032, Debrecen, Egyetem tér 1.
xepenator@gmail.com, laszlo.csipkes@freemail.hu
[2] Magyar Tudományos Akadémia, Nyelvtudományi Intézet
H-1394, Budapest, Pf. 360
oravecz@nytud.hu

Kivonat: A különböző kommunikációs események számítógépes elemzése során nélkülözhetetlen támpontot jelent, hogy gépileg feldolgozható formában elérhetők legyenek az azokat kísérő és általánosságban jellemző fizikai jegyek, mint amilyen a gyorsuló beszédtempó vagy az eltérő hanghordozás. A jelen tanulmányban bemutatásra kerülő, a HuComTech-korpusz és -adatbázis bővítéseként tervezett automatikus prozódiai annotáció ezeknek az információknak a feltérképezését szolgálja abból a célból, hogy a lehetővé tegye a korpusz annotációiban rögzítésre került kommunikációs jelenségek akusztikai jellemzését. A tanulmány a korpusz általános bemutatása után ennek céljait, módszereit és lehetőségeit kívánja részletezni.

1 Bevezetés

A HuComTech projekt[1] keretében létrehozott multimodális élőnyelvi korpusz és adatbázis számtalan feldolgozási és kutatási lehetőséget rejt magában. A kommunikációelméleti szakemberek, digitális képfeldolgozók és számítógépes nyelvészek közreműködésével, 113 beszélő részvételével gyűjtött, 50 órányi annotált anyag azzal a céllal készült, hogy egy egységes elméleti kerethez igazodva létrejöjjön egy olyan empirikus erőforrás, amely különféle kutatásokra, adatbányászatra, gépi betanításra alkalmas alapanyagot jelent a projektben együttműködő, illetve külső kutatók számára [4]. Jelen tanulmány a jelenlegi specifikációk rövid ismertetése után az adatbázis bővítéseként tervezett automatikus prozódiai annotációt, annak módszereit és lehetőségeit kívánja bemutatni.

[1] A kutatás alapjait *Az ember–gép kommunikáció technológiájának elméleti alapjai.* című, TÁMOP-4.2.2-08/1/2008-0009 projekt azonosítójú program keretei között teremtettük meg. Jelen tanulmány *A felsőoktatás minőségének javítása a kutatás-fejlesztés-innováció-oktatás fejlesztésén keresztül a Debreceni Egyetemen* című, TÁMOP-4.2.1/B-09/1/KONV-2010-0007 projektazonosítójú program keretein belül jött létre.

1.1 A HuComTech-korpusz és -adatbázis bemutatása

A HuComtech-korpusz magját egy összességében 50, beszélőnként fél óra hosszú audio- és videófelvétel alkotja. A felvételek mindegyike két személy (egy interjúztató és egy interjúalany) részvételével került rögzítésre, egy formális és egy informális társalgási szcenárió felhasználásával. Az első (formális) rész egy szimulált állásinterjú formájában, a második egy irányított beszélgetés szabadabb keretei között valósult meg, amelyek során az interjúztató különféle módszerekkel igyekezett az interjúalanyból spontán reakciókat kiváltani.

1. ábra: pillanatfelvétel a HuComTech korpuszból. Az interjúalany oldala.

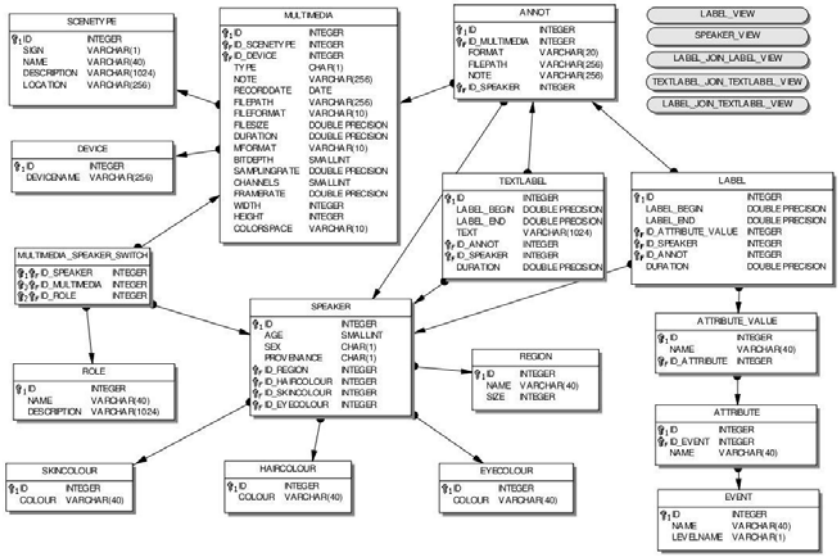
A korpusz számítógépes feldolgozhatóságát a felvételekhez készült annotációk biztosítják, amelyek elkészítésre az akusztikus és a vizuális csatornán párhuzamosan, többféle megközelítésben (fizikai jelek, nyelvi egységek és kommunikációs jelenségek megfigyelése), azokon belül is több elemzési szempont alapján történt.

A vizuális annotáció a képi anyagon megfigyelhető, a kommunikációs eseményeket kísérő, azok lehetséges jellemzőit képező fizikai jeleket rögzíti (fejmozgás, gesztikuláció, tekintetirány stb.), illetve interpretálja (arckifejezés jellege stb.). Az audioanyag szegmentálása során a beszédfolyam szintaktikai egységekre bomlik, amelyek mentén az annotáció a beszédfolyam szöveges átiratán kívül további információként tartalmazza annak hallás alapján meghatározott érzelmi töltését (a szemantikai tartalom figyelmen kívül hagyásával). Az így kinyerhető adatok a vizuális és akusztikus csatorna összefüggéseinek vizsgálatán túl a pragmatikai szempontú annotáció címkéivel összevetve válnak igazán informatívvá, ahol az annotátorok már nem nyelvi egységeket vagy fizikai jeleket, hanem kommunikációs eseményeket rögzítenek, vizuális, akusztikus és audiovizuális jegyek alapján.

Technológiai szempontból az audio- és a videócsatorna annotációja különböző számítógépes eszközökkel[2] és eltérő szegmentálási módszerekkel valósult meg, nem kizárva ezzel az utólagos konverziók, a modalitások egyesítése révén megvalósítható multimodális lekérdezéseket sem. Az annotációk tartalmazta adatok a feldolgozás során egy SQL-alapú adatbázis részeivé válnak, amely a felvételekkel kapcsolatos

[2] A videófelvételek rögzítésére a digitáliskép-feldolgozó csoport által fejlesztett Qannot, az audiofelvételek feldolgozására pedig a Praat beszédfeldolgozó szoftver szolgált [2].

különféle metainformációkat (beszélő neme, életkora stb.) is magában foglalja, az annotációs címkéket pedig a modellben elfoglalt helyük és tulajdonságtípusaik (arckifejezés, érzelmi töltés stb.) alapján rendszerezi (2. ábra).



2. ábra: A HuComTech adatbázisséma.

Az SQL lekérdezéseken kívül, a nyers adatokon (felvételek és annotációk) folytatott munka a feldolgozás azon részét képezi, amely egyúttal a korpusz bővítését is magával vonja az automatikusan generált új annotációk vagy metaadatok formájában. Az automatizált adatgyűjtés és címkézés ilyen számítógépes nyelvészeti irányú részét képezi a különféle akusztikai információk kinyerése és annotálása a már meglévő manuális annotációk felhasználásával.

1.2 Az automatikus prozódiai annotáció szerepe az adatbázisban

A prozódiai annotációval ellátott beszélt nyelvi korpuszok rendkívül értékes nyelvi erőforrást képviselnek, ám előállításuk igen munkaigényes. További problémát okoz, hogy a nemzetközi gyakorlatban nincs egyértelmű megállapodás arra vonatkozóan, hogy pontosan mit is tartalmazzon egy prozódiai annotáció.

Saját annotációs eljárásunk megtervezése során a távlati célok figyelembevételével azokat az elemzési megközelítéseket tekintettük megfelelőnek, amelyek az adatbázisban jelölésre került kommunikációs események gépi detektálásához szolgáltathatnak releváns információkat. Ennek megfelelően a kommunikációs eseményeket kísérő, általánosságban jellemző és valós időben is feldolgozható fizikai jegyeket szükséges

elemezhetővé tenni, amelyek együttesei, meghatározott irányú progressziója alapján amazok felismerhetővé válnak.

A pragmatikai annotációkban jelölt kommunikációs események ilyen értelemben vett potenciális kísérőjegyei vizuális oldalon részben manuálisan, részben automatikusan (pl. a szájmozgás) rögzítésre kerültek, detektálásuk pedig a digitális képfeldolgozás feladatkörébe esik, a kapcsolódó prozódiai információk viszont az adatbázis jelenlegi állapotában egyáltalán nem elérhetők. Az automatikus prozódiai annotáció célja pótolni ezt a hiányt, hogy a nyers adatok (F0 és intenzitásértékek) az adatbázisban közvetlenül, illetve a különféle címkézési eljárások révén feldolgozott formában is lekérdezhetővé váljanak. A feldolgozás eredményeként kapott címkesorokból aztán tágabb körű elemzések útján további metainformációk nyerhetők ki az interakciók beszéddinamikai mintázatairól, amelyek feltérképezése által a kommunikációs események felismerését segítő tudás birtokába juthatunk. Például arról, hogyan változik egy dialógus intenzitása az abba bekerülő új információk, témaváltások hatására.

2 A prozódiai annotáció lépései

2.1 F0- és intenzitásadatok kinyerése és integrálása az adatbázisba

A beszédfolyam akusztikai karakterizálásához leginkább felhasználható F0 és intenzitás adatok kinyerésére a Praat beszédfeldolgozó szoftver [2] e célra kidolgozott, beépített szkript nyelve által könnyen automatizálható lekérdező funkciói mellett saját fejlesztésű, valós időben is működő, jelenleg tesztelés alatt álló algoritmusokat kívánunk a későbbiekben felhasználni. Ezek tetszőleges formára hozható kimenete a korpusz részeként további elemzések bemenetéül szolgál, illetve feltöltésük után az eredmények az adatbázis-lekérdezések során is felhasználhatóvá válnak.

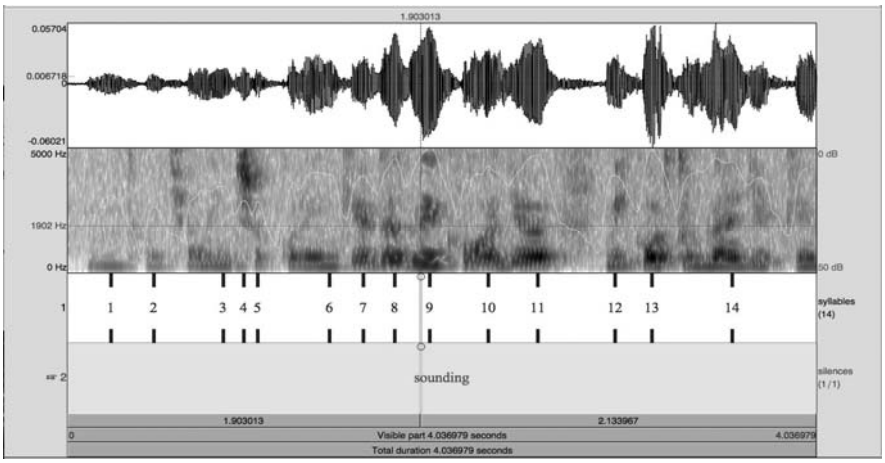
A HuComTech projekt jelenlegi adatbázissémája egyetlen relációs táblában tárolja a különböző típusú annotációk címkéit a címkekezdet, címkevég, címkeérték oszlopokban rögzítve az azokat jellemző legfontosabb információkat (lásd 1. ábra). Az olyan típusú akusztikai adatok, mint az egy adott időpillanathoz tartozó F0- és intenzitásértékek tárolására ez a tábla nem alkalmas, így a többi annotációs címkétől szeparáltan, külön táblában kerülnek tárolásra, amely később alkalmas egyéb, megegyező struktúrájú (idő → érték) fizikai adatok tárolására is. Ezek az adatok a lekérdezések során természetesen csak bizonyos kalkulációk, például bizonyos címkeszakaszokra vagy az egész fájlra számolt átlagértékek után válnak kellően informatívvá.

2.2 A beszédtempó annotációja

A feldolgozási eljárás egyik fontos komponensét a beszédtempó mérése és címkézése jelenti, melynek során a beszéd sebességének változásairól kívánunk számot adni.

A beszédtempó mérésének kivitelezéséhez elsősorban egy olyan mérési objektum meghatározására van szükségünk, amelynek egy adott időegységre mért gyakorisága,

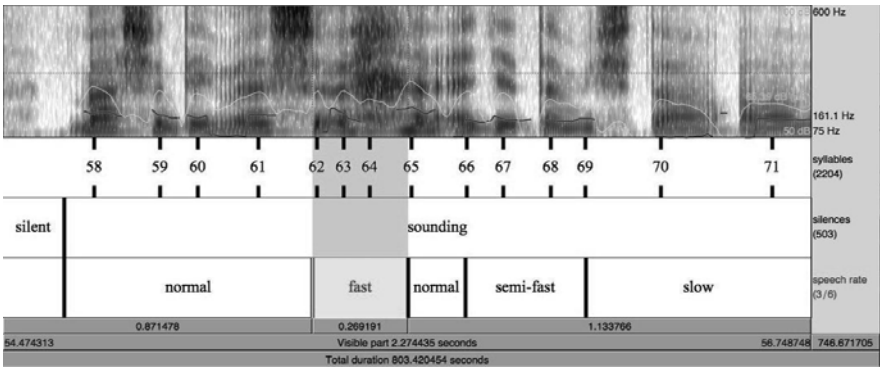
sűrűsége megragadhatóvá teszi azt. A létező megoldások után kutatva találtunk rá Nivja H. de Jong és Ton Wempe tanulmányára [3]. A szerzők a beszédtempó vizsgálatához a szótagmagokat választották mérési objektumként, amelyek detektálására egy jól működő módszert is kidolgoztak. Az eljárás Praat beszédfeldolgozó program beépített szkript nyelvét, függvényeit és mérési algoritmusait használja. A szótagmagok detektálása az intenzitás görbe csúcsainak meghatározott küszöbértékek (csúcsok közötti minimális értékbeli különbség stb.) szerinti szűrése által történik a beszédfolyam nem hangzós részeinek kizárásával. Az eredményül kapott intenzitáscsúcsok időpillanatai a Praat TextGrid formátumú annotációs fájljaiban kerülnek tárolásra, amelyek a program szerkesztőfelületén jeleníthetők meg (2. ábra), illetve egyéb szoftveres megoldásokkal is könnyen feldolgozhatók.



3. ábra: A szótagmagok detektálása.

A beszéd sebességének ingadozása így a szótagmagok helyét reprezentáló intenzitáscsúcsok közötti távolság változásain keresztül válik megragadhatóvá.[3] Ehhez természetesen figyelembe kell vennünk a beszéd sebességének az adott beszélő egyedi beszédtempójából következő relatív viszonyait, amely a teljes beszédfolyamra számolt előzetes statisztikák segítségével valósítható meg. A hangzós részekre számolt csúcsok közötti távolság átlagértékének megadásával meghatározhatjuk az adott beszélő normál beszédtempóját. Az eljárás során az átlagolást először minden hangzós szakaszra külön-külön végezzünk el, majd ezeket az eredményeket átlagoljuk újra. A normál beszédtempó meghatározása után relatív küszöbértékek kiszámításával további kategóriákat állíthatunk fel, amelyek már az adott szakaszokra történő címkézési eljárás során kerülnek felhasználásra (3. ábra).

[3] A különböző magánhangzók eltérő ejtési idejéből fakadóan ez az eljárás könnyen vezethet megtévesztő eredményekhez. Az algoritmus tökéletesítéséhez tehát plusz információként figyelembe kell venni a csúcsok által reprezentált szótagmag időbeli terjedelmét is, amely az F0- és az intenzitásgörbe további vizsgálata révén lesz megvalósítható.



4. ábra: A beszédtempó címkézése.

A beszéd aktuális tempóját tehát az adott szegmensen belül fellelt szótagmagok átlagsűrűségének az adott beszélőre jellemző normál átlagsűrűséghez viszonyított különbsége fogja meghatározni a beszéd aktuális tempóját. A eljárás lépéseit összefoglalva:

- szótagmagok detektálása (de Jong és Wempe munkája [3] nyomán)
- normál beszédtempó meghatározása a szótagmagok hangzós részekre számolt átlagsűrűsége alapján (beszélőspecifikus tulajdonság)
- az adott beszédszegmens átlagsűrűségének kiszámítása
- az adott beszédszegmens tempójának kategorizálása a normál beszédtempótól való eltérés foka alapján

A címkézés esetében problematikus kérdés, hogy milyen egységekre, a beszédfolyam mely szakaszaira történjen az aktuális beszédtempó kategorizálása. Lehetséges utat jelent a korábban már manálisan annotált szegmentumok, illetve a szünettől szünetig tartó hangzós részek tempójának címkézése. Az eljárásnál problémát jelent, hogy egy folytonos (szünettől szünetig tartó) beszédszakaszon, vagy akár egy szintaktikai egységet reprezentáló annotált szegmentumon belül is számítanunk kell a tempó ingadozására. Hogy ezeket az információkat ne veszítsük el, az adott egységen belül is vizsgálnunk a beszédtempó alakulását, a beszélőt és az egységet jellemző adatokból számolt küszöbértékek felhasználásával.

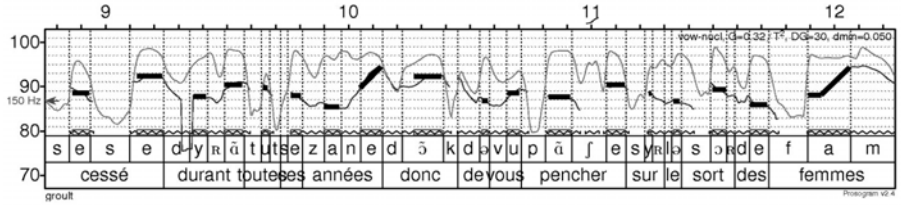
2. 3 Az alapfrekvencia progressziójának annotálása

A prozódiai annotáció következő lépését az alapfrekvencia progressziójának elemzése jelenti, amelynek eredményeként a beszédfolyam meghatározott szegmentumaihoz valamilyen egzakt tonális karaktert jelölő annotációs címkét (emelkedő, ereszkedő, eső stb.) vagy címkekombinációt rendelünk. Ennek megvalósítása érdekében a kimért F0-értékekre számolt trendvonalak formájában előbb feldolgozható formában stilizálnunk kell az alapfrekvencia változásait.

Az eljárás megvalósítására Piet Mertens kapcsolódó munkáját [5] terveztük felhasználni. Mertens előzetesen számos fontos feltételt fogalmaz meg, amelyeket a prozódiai annotáció során nem szabad figyelmen kívül hagyni:

- Az annotácónak alapvetően az érzékelhető intonációt kell reprezentálnia ojektív és könnyen értelmezhető módon,
- Az alapfrekvencia változását hosszabb beszédfolyamon keresztül is tükröznie kell, a szélesebb tartományokra kiterjedő változások rögzítése érdekében,
- A fizikai jelek időbeli szerveződését meg kell őriznie a szünetek, hezitációk, beszédtempó és a ritmus azonosíthatósága érdekében,
- Az annotációnak automatikusnak vagy félautomatikusnak kell lennie,
- Az annotáció elméletsemleges kell, hogy legyen, a széleskörű használhatóság érdekében,
- Az annotáció lehetőleg időben illesztett fonetikai és szöveges átírást tartalmazzon az olvashatóság és szöveges keresés lehetőségének biztosítása érdekében.

Mertens [5] kifejlesztett egy, a fenti feltételeknek megfelelő transkripciós rendszert, amely a vokális szótagmag alapfrekvenciájának stilizált kontúrját felhasználva félautomatikusan rendel prozódiai annotációt fonetikai transkripcióhoz. A stilizálás [1] alapján a tonális érzékelés pszichoakusztikai modelljére épül. Az annotáció megőrzi az akusztikai jel temporális jellemzőit, és beépíti a szöveges, illetve a fonetikai transkripciót is, ahol ez utóbbi a vokális szótagmag azonosításában játszik szerepet. A rendszer többféle részletességű információt tartalmazó kimenetet képes generálni: a kompakt változat a stilizált beszéddallam szöveges és fonetikai átírással kiegészített annotációját tartalmazza (lásd 5. ábra).



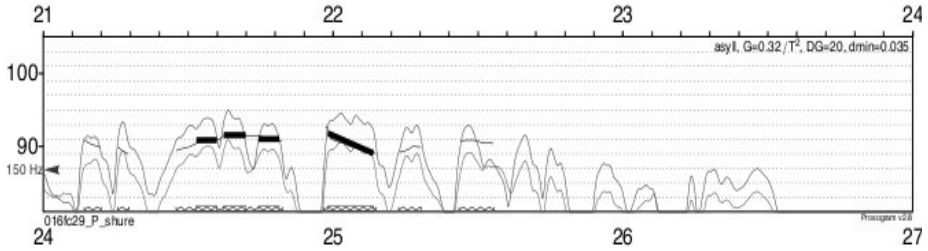
5. ábra: A Mertens-féle transkripciós rendszer kimenete.

A módszer implementációja a Praat beszédfeldolgozó program felhasználásával történt. A transkripciókat generáló Praat szkript a hozzá tartozó dokumentációval együtt Prosogram (v2.8) néven szabadon hozzáférhető[4], többféle beállítással és üzemmódban futtatható, lehetőséget biztosítva például meglévő, a megfelelő formátumban tárolt manuális szegmentációk használatára. A HuComTech-korpuszban hozzáférhető szöveges transkripciók tagmondatszintű annotációkat takarnak, így az alapfrekvencia félautomatikus stilizációjához ezek nem

[4] http://bach.arts.kuleuven.be/pmertens/prosogram/

felhasználhatók, viszont a program lehetőséget kínál a hanganyag szótagokra és szótagmagokra történő automatikus szegmentálására is.[5]

Az eredményül kapott stilizációknak[6] a felhasználásával további elemzésével lehetővé válik a beszédfolyam szegmentumainak egzakt kategorizációja. Problémát jelent viszont, hogy a stilizációkat tartalmazó kimenet csak grafikus formában elérhető. A általunk tervezett, a HuComTech adatbázisba integrálható prozódiai annotáció megvalósításához így a stilizációk megjelenítésért felelős algoritmust előbb vissza kell fejtenünk és át kell alakítanunk, hogy a célnak megfelelő, a további számításokhoz felhasználható numerikus kimeneteket (a stilizációk kezdő és végpontja) tudjunk produkálni. A program saját anyagunkon végzett tesztelésének grafikus kimenetét az 5. ábra szemlélteti.



6. ábra: A Prosogram grafikus kimenete.

A további elemzések bemenetét tehát az alapfrekvencia stilizált progressziója adja, amely a dallamgörbe normalizált darabjainak hosszában, a kezdő és végpontok frekvenciaértékének különbségében ragadható meg. Ezeknek az értékeknek a felhasználásával történik a beszédfolyam tonális egységeinek címkézése, ahol minden címke az adott egység dallamának karakteréről próbál feldolgozható leírást adni.

Mint ahogyan a beszédtempónál, az alapfrekvencia annotálásánál is problémát jelent, hogy a beszédfolyamnak melyek azok az egységei, amelyek kiértékelése révén az alapfrekvencia változásairól a számunkra megfelelő léptékű képet kapjuk. A jelenlegi tervek szerint ezek az egységek a korpuszban már manuálisan annotált, potenciális intonációs frázisokat jelentő tagmondatok lesznek, nem kizárva a dallammenet tágabb léptékű, különféle kommunikációs események mentén történő elemzését. Ezekhez a vizsgálatokhoz célszerű a tagmondatszintű progresszió kategorizálása mellett számot adni a beszéddallam aktuális tartományáról, annak relatív magasságának függvényében.[7]

[5] Ennek megbízhatósága saját anyagunkon jelenleg tesztelés alatt áll.

[6] Amelyeket a továbbiakban az alapfrekvencia normalizált progressziójának tekintünk.

[7] Ennek a relatív magasságnak a meghatározásához az adott beszélőre jellemző hangterjedelem szolgáltat információkat.

3 Összegzés

A HuComTech-korpusz és -adatbázis jelenlegi állapotában számos vizsgálati lehetőséget biztosít kommunikációelméleti kutatások folytatására. Az automatikus prozódiai annotáció sikeres implementációja jelentős mértékben kitágítja ezeket a vizsgálati lehetőségeket az akusztikai információk feldolgozható formában történő bekapcsolásával, olyan további kutatásokat alapozva meg, melyek egy adott kommunikációs esemény valós időben történő detektálásának vagy predikciójának algoritmizálhatóságát célozzák.

Bibliográfia

1. Alessandro, P., Mertens., P.: Automatic pitch contour stylization using a model of tonal perception. Computer Speech & Language Vol. 9, No. 3 (1995) 257-288
2. Boersma, P., Weenink, D. (2010): Praat: doing phonetics by computer 5.1.43. Institute of Phonetic Sciences, University of Amsterdam. http://www.praat.org
3. de Jong, N. H., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods Vol. 41, No. 2 (2009) 385-390.
4. Hunyadi, L.: Multimodal human– computer interaction technologies. Theoretical modeling and application in speech processing. Argumentum. Megjelenés alatt (2011)
5. Mertens, P.: The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In: Bel, B., Marlien, I. (eds.): Proceedings of Speech Prosody 2004i, Nara (Japan), 23-26 March (ISBN 2-9518233-1-2) (2004)
6. Pápay, K., Szeghalmy, Sz., Szekrényes, I.: HuComTech Multimodal Corpus Annotation. Argumentum. Megjelenés alatt (2011)

A HuComTech audio adatbázis szintaktikai szintjének elvei és szabályrendszerének újdonságai

Kiss Hermina[1]

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

Kivonat: A HuComTech multimodális adatbázis egyik annotációs szintje a szintaktikai szint. Az annotációs szempontrendszer kialakítása során újbóli átgondolásra került a mondat fogalma, a tagmondatok hierarchiájának elemzési módszere és az implicit nyelvi elemek kimutatásának módszertana. Ennek tükrében létrehoztunk egy új típusú mondatelemzési módszert, aminek szintaktikai alapegysége a tagmondat. Az adatbázis legfontosabb alapelvei: az adatbázis legyen preteoretikus, tükrözze a különböző tudományos megközelítések közötti konszenzust, valamint legyen alulspecifikált. 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. Ez az előadás erre vállalkozik.

1 Bevezetés

A munkánk alapját a HuComTech spontánbeszéd-korpusz és adatbázis képezi. Az adatbázis a kommunikáció számos multimodális jegye mellett nyelvi, ezen belül a beszédre is vonatkozó adatot tartalmaz. Külön kihívás a folyamatában megszülető, a kommunikáció során kialakuló spontánbeszéd mondattani elemzése, hiszen az gyakran ellenáll a hagyományos mondatelemzésnek. Elemzése és annotálása számos problémát vet föl egyrészt azért, mert a beszélő még nem tudja, hogy az általa kifejezendő információ milyen szerkezetben fog megjelenni, másrészt pedig az élőszó spontaneitásának gyakori következménye a pongyola nyelvhasználat, ami egy nem kellőképpen átgondolt és nem megfelelően létrehozott szintaxist hoz létre. Első és legfontosabb dolog a spontán beszéd annotációs szabályainak kialakításához, hogy meghatározzuk a használandó alapfogalmakat. 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. A szintaktikai struktúra alapjának a tagmondatot tekintjük (mélyebb bontásra már csak azért sem vállalkozunk, mert ezt a beszélt nyelvi produkció gyakran nem is teszi lehetővé) és ezt szerkezeti sajátosságai alapján határozzuk meg. Az elemzés és az annotáció

[1] A jelen tanulmány alapjául szolgáló kutatásban a szerzőt *A felsőoktatás minőségének javítása a kutatás-fejlesztés-innováció-oktatás fejlesztésén keresztül a Debreceni Egyetemen* című, TÁMOP-4.2.1/B-09/1/KONV-2010-0007 projektazonosítójú program támogatta.

egységes strukturális szempontok alapján azt ígéri, hogy az elemzés jól tükrözi a nyelv beszédben kialakuló szerkesztését, ugyanakkor kellően alulspecifikált ahhoz, hogy különböző elméleti megközelítésekben is jól használható legyen.

2 A mondat és a tagmondat fogalma

A mondat fogalmának definícióját olyan szempontból közelítjük meg, hogy érvényesüljön az az alapvető célunk, miszerint az általunk kidolgozott szintaktikai modell preteoretikusan működtethető, tehát az ember-gép közötti kommunikáció tanulmányozására létrejött adatbázisban szinkronba hozható a nyelvészeti szakirodalom mondatfogalmának többféle szempontú megközelítése és ennek megfelelően többféle meghatározása.

Ennek a célnak az egyik velejárója az, hogy elemzési szempontrendszerünk alulspecifikált, hiszen a tagmondatok közötti viszonyok meghatározása után nem bontjuk tovább az elemzési szempontrendszert úgy, hogy az alá- és mellérendelő mondattípusok megnevezését is lehetővé tegyük.

A *Strukturális magyar nyelvtan* mondattanról szóló kötetében az alárendelő mellékmondat vonzatnak minősül, ezért nem érvényes az a szerkezeti meghatározás, miszerint a mondat szerkezete egyszerű és összetett mondatokból áll össze [1]. Mi viszont elfogadjuk, hogy a kifejtett mondatrész külön tagmondat, hogy minél részletesebben és érzékletesebben kimutassuk a mondat implicit elemeit. Nem mondjuk azt tehát, hogy az alárendelt tagmondat egy vonzat, és nem hiányzik semmi a mondatból, hanem külön tagmondatként értelmezve felszínre hozzuk az így kimutatható implicit nyelvi elemeket.

Ebbe a rendszerbe beilleszthető a vonzatról való felfogásunk, amit a *Strukturális magyar nyelvtan*, illetve a *Magyar grammatika* [2] is elfogad: vonzatnak az elhagyhatatlan bővítményeket tartjuk, ami azt jelenti, hogy a vonzat a grammatikai struktúra sérülése nélkül nem hagyható el a nyelvi egység mellől, amihez tartozik. Az alanyt viszont nem tekintjük vonzatnak.

Ennek megfelelően az elemzésünk alapegysége a tagmondat. A tagmondat szerkezetileg nem más, mint szavak kapcsolódása egy hierarchikus rendben. Egy tagmondat szerkezeti határát az képezi, amikor egy adott szót már nem tudunk az addig (az azt lineárisan megelőző és/vagy követő szavakból) felépült hierarchikus rendben elhelyezni. Funkcionálisan egy hiánytalan tagmondat a régensből (állítmány) és kötelező vonzataiból, valamint az alanyból áll. Számunkra az állítmány az igét és annak vonzatait jelenti együttesen, tehát nem csupán a leíró nyelvtan szerinti egyszerű és összetett állítmányt, hanem azzal együtt a vonzatokat is magába foglalja.

A szerkezetek láncszerű grammatikai kapcsolata tagmondatok sorát alkotja meg. Ezek, ha szerkezetileg kapcsolódnak, mondattá állnak össze. A mondat tehát a tagmondatok láncszerű, szerkezeti kapcsolódása és minimum egy tagmondatból áll.

3 Implicit nyelvi elemek

A beszélt nyelvben gyakori elemek az ismétlések, a töltelékszavak, a mondatok megszerkesztettsége lazább, szabálytalanabb. Ennek egyik grammatikai következménye az, hogy elmaradhat a főmondat, az utalószó, a kötőszó, a grammatikai, illetve logikai alany, az állítmány, a tárgy, a jelző és az ige. Ezen nem jelölt nyelvtani elemekre bizonyos esetekben következtethetünk akár strukturálisan, akár szemantikailag/kontextuálisan, más esetekben azonban nem (pl. a megkezdett, de befejezetlen tagmondatok esetében). A grammatikailag jólformált és nem jólformált tagmondatokat egyazon szempontrendszer alapján elemezzük.

4 Minimális mondat

A beszélt nyelv lazább szerkesztettségének fentebb bemutatott grammatikai következménye az implicit nyelvi elemek gyakori előfordulása mellett egy másik fontos grammatikai következménye az, hogy egy-egy forduló [3] állhat különálló szavak olyan egymásutániságából, amelyek között semmilyen grammatikai összerendeződés nincs. A tagmondat fentebbi meghatározása alapján ilyen esetekben ezen szavak külön-külön egyetlen tagmondatból álló mondatokat képeznek. Ezek a minimális mondat esetei. Külön figyelmet kell fordítanunk a lexikális tartalom nélküli hangzó megnyilvánulásokra. Ezek a lexikális tartalom nélküli minimális mondat esetei. Csak azokat az eseteket vesszük figyelembe, amelyek a fordulók elején vagy végén jelennek meg. (A tagmondat szavai között megfigyelhető, gyakran bizonytalanságot vagy a kifejezendő gondolat módosítását jelző hangzó megnyilvánulásokat, mint amik nem befolyásolják a mondatszerkezetet, nem jelöljük.) A minimális tagmondatra a példák a következő alfejezetek:

4.1 A befejezetlen tagmondatok

például: *De*… És… *Hm* … *Úúú*…

4.2 A mondatszók

például: köszönések, megszólítások, indulatszavak, töltelékszavak, stb.

4. 3 Egyszavas válaszok

például: *Igen. Nem. Talán.*

4.4. Egymondatos visszakérdezések

például: *ugye*, (akár visszakérdezés, akár töltelékszó), *Legjobb főnök? Legszebb élmény?*
De az *így/úgy* töltelékszavakat nem soroljuk ide, mivel grammatikailag (határozóként) kapcsolódnak a tagmondathoz.

4.5. A mint-tel kezdődő hasonlító határozói alárendelt tagmondatok

Olyan lettem, *mint te*. Ez már nem olyan, *mint az volt.*

4.6. Töltelékszavak

ugye, így/úgy tehát, stb.

5 Teljes tagmondat

Fentebb a beszélt nyelvre különösen jellemző, valamilyen szempontból hiányos szerkezetekről szóltunk. Természetesen a beszélt nyelvben is találkozunk az ettől különböző szerkesztéssel, azaz a strukturális szempontból teljes mondatokkal. Ezek funkcionálisan tartalmazzák a régenst (állítmányt) annak kötelező vonzataival és az alanyt. Ezt a leíró nyelvtan egyszerű mondatának nevezhetjük, illetve összetett mondat esetén az összetételeket.

6 A tagmondatok kódolása

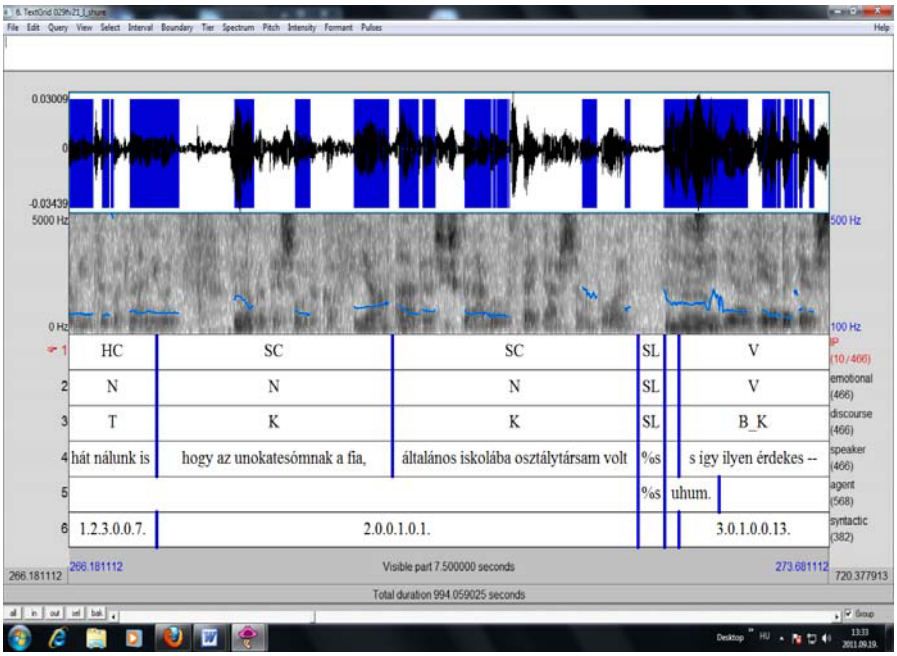
A tagmondatok láncolata lineárisan és hierarchikusan is szervezi a beszédet. Ennek feltárása alapvető célunk. Ennek megfelelően meghatározunk az alá- és mellérendelő tagmondatokat, illetve a tagmondatok közötti grammatikai kapcsolat hiányát (beágyazást, beékelést).

6.1 Szegmentációs szakaszok

Alárendelő tagmondatok esetén egyértelmű a szegmentációs határhelyzet, azaz a tagmondat határa. Mellérendelő tagmondatok esetén vagy új mondat indul kötőszóval, illetve anélkül kezdve, vagy az előtte lévő tagmondathoz kapcsolódik, s így még ugyanannak a tagmondatfüzérnek a tagja, amihez az előző kapcsolódik.

6.2 A számozás

A számozás a tagmondatok közötti sorrendiséget és a tagmondatok közötti viszonyt fejezi ki. A számozás kezdete a hagyományos mondat kezdetét jelöli. A számozás ott fejeződik be, ahol a hagyományos mondat végét lehet érzékelni. A hagyományos mondat végét nem az intonáció és elsődlegesen nem a szemantika, illetve interpretáció határozza meg, hanem a szintaktika.



1. ábra: A szintaktikai annotációs szint kódolása.

Az 1. ábrán láthatjuk a kódolási rendszert, az annotáció 6. szintjén. A kódrendszerben az első szám tehát a tagmondatok sorszámát jelenti. A második szám azt jelöli, hogy az adott tagmondathoz tartozik-e alárendelés, és ha igen, akkor hányas számú tagmondat. Ha nincs, akkor az 0 értékkel van jelölve. A harmadik szám a tagmondathoz tartozó mellérendelő tagmondat(ok) sorszámát jelöli. Ha nincs ilyen, akkor a 0 érték látható. A negyedik számjegy azt mutatja meg, hogy az adott tagmondat hányas számú tagmondatnak az alárendeltje. Itt is megjelenhet a 0 érték. Az ötödik számjegy a grammatikai kapcsolat hiányát jelöli, azt mutatja meg, hogy melyik tagmondat kapcsolódik hozzá úgy, hogy grammatikai elem nem jelenik meg. A számok között pont van. Ha egy elemzési szemponthoz több szám is tartozik, akkor azok vesszővel vannak elválasztva.

7 A hiány kategóriái

7.1. Nem hiányzik semmi

Nem hiányzik semmi abban az esetben, ha érvényesül a teljes tagmondat fent leírt definíciója.

7.2 Hiányzik a főmondat

Mert szeretnék munkát.
Ha így lesz.
Mikor még kicsi voltam.

7.3 Hiányzik az előtte álló mellékmondat

Abban az esetben használjuk ezt a kategóriát, amikor a tagmondat előtt nincs tagmondat, (az előző mondathoz tartozó tagmondat).

És ő nem vette fel a telefont.
Meg el sem jött
De én mindenképp el akartam menni.

7.4 Hiányzik a kötőszó

Attól függ, mit nézünk.
Éreztem, hogy pályakezdőként itt sokat tanulhatok.
Emlékszem, amikor ezt tavaly átéltem.

7.5 Hiányzik az utalószó

Angolt tanultam úgy eddig is, mert nekem az egyetemen kellett.
Sokszor dolgoztam már, hogy minél tapasztaltabb legyek.
Ne mondjátok meg, hogy hová kell menni.

7.6 Hiányzik a grammatikai alany

Csak úgy nem ilyenre számítottam.
Megyek dolgozni.
Nagyon fontos dolgokat mondott nekünk.

7.7 Hiányzik a logikai alany

Hát, általában így szokott lenni.
Nincs szükségem erre egyáltalán.
Nem volt még előző munkahelyem.

7.8 Hiányzik az állítmány

Például: *A főnök kabátban.* Abban az esetben hiányzik az állítmány, ha az ige és annak vonzatköre nem jelenik meg a tagmondatban, van(nak) viszont egyéb szabad határozó(k).

7.9 Hiányzik a tárgy

Sokszor iszik valóban.
Akkor így nem vették észre.
Ő is látta.

7.10 Hiányzik a határozó

Például: a *megy* ige vonzatai: vki, vhová. Ha ezek közül hiányzik a határozó, akkor az hiánynak van feltüntetve.

Nem hitt.
Részt vett.
Pista jártas.

7.11 Hiányzik a jelző

Liter tejet hozott.
Köbméter víz fogyott.
Kiló kenyérrel tért vissza a munkahelyére.

7.12 Hiányzik az ige

Például: *János spagettit*. Ha a tagmondatban megvan(nak) a kötelező vonzat(ok), de a régens hiányzik (eszik/evett/ fog enni).

Péter a kávét.
János könyvet.
A lisztet.

7.13 Befejezetlen tagmondat

Az élőbeszédre jellemző sajátosság, hogy a nyelvtani korrekciók a beszéd folyamatában történnek meg. Ennek grammatikai következménye az, hogy a szerkesztés befejezetlen marad. A tagmondat meghatározás alapján azonban az ilyen befejezetlen szerkezeteket tagmondat értékűnek tekintjük. A befejezetlenséget azonban külön kódoljuk, ugyanis feltesszük, hogy a befejezetlenség által keltett információhiányt egy másik, nem nyelvi modalitás pótolja és így az azonosítható pl. egy arckifejezésben vagy egy mozdulatban stb. Így a szintaktikai annotálás mint a multimodális annotálás része hozzájárulhat ahhoz, hogy az egyik modalitásból hiányzó elemet egy másik modalitás ugyanazon időpillanatában kutathassuk, tehát egy befejezetlen mondat kézmozdulatokkal, mimikával való lezárását nyomon követhessük a szintaktikai szinten is.

7.14 A hiány nem releváns

A hiány nem releváns akkor, ha nem tudunk érvényes hiány kategóriát megállapítani, de a mondat mégsem tekinthető teljesnek.

7.14.1 Mondatszók

7.14.1.1 Indulatszavak:

Hú! Nahát! Ó! stb.

7.14.1.2 Igenlő egyszavas válaszok:

De! Igen! Rendben! Jó! stb.

7.14.1.3 Tagadó egyszavas válaszok:

Nem. Módosítószóval együtt: Még nem. Szerencsére nem. Nem nagyon. Én nem. *Innen még nem.*

De ha a tagadószó mondatrészben van,akkor az alany és az állítmány hiányzik: *Azt mondom, hogy nem. Utazáshoz tudnám kötni, de igazából még nem.* stb.

7.14.1.4. Bizonytalan egyszavas válaszok:

Talán. Lehet. Bizonyára. stb.

7.14.1.5 Köszönések:

Viszlát! Viszontlátásra! Jó napot! De a Jó napot kívánok! köszönésforma nem tartozik ehhez a kategóriához, mert egyértelműen meg tudjuk határozni a mondatban az alanyt, az állítmányt és a vonzatot.

7.14.1.6 Udvariassági formulák:

Szívesen! Nagyon szívesen! stb.

7.14.1.7 Töltelékszavak

Hát, ugye, így, úgy, stb.

7.14.1.8 Megszólítások

András! Kinga! stb.

7.14.2 Egymondatos visszakérdezések

például: *ugye?* (akár visszakérdezés, akár töltelékszó), *Legjobb főnök? Legszebb élmény?* De az *így/úgy* töltelékszavakat nem soroljuk ide, mivel grammatikailag (határozóként) grammatikailag kapcsolódnak a tagmondathoz.

7.14.3 *Mint*-tel kezdődő hasonlító határozói alárendelt tagmondat esetén

Szebb, mint az.
Sokkal jobb lesz így, mint úgy. stb.

7.14.4 Valamilyen okból (például a pongyola nyelvhasználat mértéke miatt) kikövetkeztethetetlen tagmondatok esetén

Ha véletlenül találkozunk egy sziát, de több nem.

8 Összegzés

A Praat szoftver felhasználásával olyan annotációs szabályrendszert dolgoztunk ki, amely lehetővé teszi a spontán beszéd szintaxisának kutatását. Különös hangsúlyt fektettünk arra, hogy a spontán beszéd jellegzetességeit kezelhetővé tegyük a magyar

nyelv mondattana keretei között, mint ami rendszerében nem, csak megvalósulásában különbözik attól. Nem tettünk említést számos problémakörről, amelyek az adott kategóriák átgondolását segítették. Például az egyedi szó- és nyelvhasználatból adódó jelenségekről, sajátosságokról, vagy a töltelékszavak, indulatszavak spontán beszédbe illeszkedő rendszeréről, illetve a pongyola nyelvhasználat következményeként létrejövő szintaktikai problémákról. (Mint például az abszolút és relatív főnév elhelyezkedése a mondat hierarchiájában, a kötőszóval kezdődő mondatok kérdéséről, a főnevesült jelző mondattani szerepköréről, a függő beszédben jelen lévő implicit elemekről, az ellipszis számos kérdésköréről, illetőleg a dialógus másik szereplőjének a vizsgált személy grammatikájára tett hatásáról.) Itt ismertetett kódrendszerünk lehetővé teszi azt, hogy az adatbázist vizsgáló kutatók további szintaktikai elemzéseket folytassanak, kiegészítve, részletezve az általunk létrejött rendszert.

Bibliográfia

1. Keszler B.: Szintagmatan. In: Keszler B. (szerk.): Magyar Grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 355
2. Komlósy A.: Régensek és vonzatok. In: Kiefer F. (szerk.): Strukturális Magyar Nyelvtan I. Akadémiai Kiadó, Budapest (1992) 308—316
3. Iványi Zs.: A nyelvészeti konverzációelemzés. Magyar Nyelvőr Vol. 125 (2001) 74-93 [http://www.c3.hu/~nyelvor/period/1251/125106.htm]

V. Pszichológia, pragmatika, kognitív nyelvészet

A csoportközi értékelés mint a csoporttrauma érzelmi feldolgozásának indikátora a nemzeti történelem elbeszéléseiben

Csertő István[1], László János[2, 3]

[1] Pécsi Tudományegyetem, Pszichológiai Intézet
H-7624 Pécs, Ifjúság útja 6.
csertopi@gmail.com
[2] Magyar Tudományos Akadémia, Pszichológiai Kutatóintézet
H-1132 Budapest, Victor Hugo utca 18-22.
[3] Pécsi Tudományegyetem, Pszichológiai Intézet
H-7624 Pécs, Ifjúság útja 6.
laszlo@mtapi.hu

Kivonat: Egy hosszmetszeti tartalomelemzéses vizsgálatban a csoportközi értékelés mintázatait tártuk fel 1920 és 2000 között kiadott magyar középiskolai történelemtankönyvek trianoni békeszerződésről szóló narratíváiban. A történelmi idő előrehaladtával változó narratív konstrukciókban a külső és a saját csoportra vonatkozó értékelések olyan eloszlási mintázatait tártuk fel három szemantikai dimenzióban, amelyek a pozitív csoportidentitást fenyegető traumatikus esemény érzelmi feldolgozására jellemzőek. A szövegelemzést a NARRCAT (Narrative Psychological Content Analytical Tool) számítógépes tartalomelemző eszköz csoportközi értékelés moduljával végeztük, melyet a PTE Pszichológiai Intézet és az MTA Pszichológiai Kutatóintézet közös narratív pszichológiai kutatócsoportja fejlesztett ki. A komplex elemzőeszköz a NooJ nyelvtechnológiai rendszerben működik, amely lehetővé teszi meghatározott, szószint feletti nyelvi alakzatok azonosítását nagy terjedelmű szövegbázisokban.

1 A nemzeti történelem mint a csoportidentitás narratív konstrukciója

Ahogyan az egyén élettörténeti beszámolói az egyéni identitás tükrét nyújtják, úgy a csoporttörténeti elbeszélések a csoportidentitás állapotairól és folyamatairól tájékoztatnak [2, 3]. A csoport múltjára, jelenére és jövőjére vonatkozó elbeszélések a csoportot érintő események és a csoportközi viszonyok terén interpretációs módokat implikál. A nemzeti történelem narratívái a társadalmi kommunikációban mint természetes közegben létrejövő csoporttörténeti elbeszélések, amelyek gazdag terepet nyújtanak a csoportidentitás és a csoportközi viszonyok dinamikájának vizsgálatára. A narratívák nyelvi-kompozíciós tulajdonságai révén olyan, a csoportközi viszonyokra és csoportfolyamatokra vonatkozó elméletek ellenőrizhetők, illetve árnyalhatók, ame-

lyek esetében a nemzeti csoportok történeti dimenziója jelentős tényező a jelenségek megértése szempontjából [3].

2 Nemzeti trauma, nemzeti identitás és kollektív feldolgozás

A nemzeti identitás a közös múlt narratív konstrukciója, melyet a társadalmi megosztás révén minden csoporttag birtokol. Jelen kutatás értelmezési keretében a nemzeti traumák olyan csoportközi konfliktusok eredményei, melyek a nemzeti identitás alkalmazkodókészségének határait meghaladó mértékű sérülését okozzák, s így újraszervezése válik szükségessé. A nemzeti trauma kollektív elaborációja az identitásnak azt az újraszervezését jelenti, amely a nemzeti történelem hosszú távú rekonstrukciós folyamatában valósul meg. E rekonstrukció célja a traumatikus esemény integrációja egy koherens és fenntartható csoportnarratívába.

A feldolgozott trauma narratívájának a következő feltételeket kell teljesítenie: (1) A traumatikus eseményt a múlt részeként reprezentálja, vagyis oly módon, hogy az eseménynek nincs közvetlen relevanciája az érintett csoportokkal fenntartott viszonyok jelenbeli alakulására. (2) A narratíva koherens, azaz következetesen illeszkedik a történelem eseményeinek láncolatába, valamint a csoporton belül általánosan elfogadott (kanonizált) konstrukció. (3) A narratíva egy fenntartható identitás része, ami azt jelenti, hogy hozzájárul egy pozitívan értékelt nemzeti azonosságtudat fenntartásához, ugyanakkor harmonikus viszonyban áll az érintett csoportokkal fenntartott jelenbeli viszonyokkal.

3 Csoportközi értékelés és traumafeldolgozás

3.2 Csoportközi értékelés és csoportidentitás

A csoportközi értékelés a narratív identitáskonstrukció lényeges nyelvi eszköze, amely az elbeszélt történelmi eseményeket és azok szereplőit jelentésteli és koherens reprezentációvá szervezi. A csoportközi értékelések explicit szociális ítéletek, melyek az eseményben érintett csoportokat, illetve azok képviselőit értékelik. Ezek lehetnek (1) nekik tulajdonított, illetve tetteiket jellemző pozitív és negatív tulajdonságok (pl. *bölcs, jogtalan*), (2) a rájuk irányuló érzelmi reakciók és viszonyulások (*csodál, megvet*), (3) a cselekvéseikre vonatkozó, értékelő jellegű interpretációk (a tényszerű leírás helyett vagy mellett; *vitézkedik, kizsákmányol*), és (4) a jutalmazás és büntetés, illetve elismerés és kritika aktusai (*éljenez, tiltakozik*).

A csoportközi értékelés alapvető szerepet játszik a pozitív szociális identitás fenntartásában. A szociális identitás elmélete [14, 15] azon a tézisen alapul, hogy az egyének önazonosságukat jelentős mértékben azoktól a csoportoktól nyerik, melyeknek tartósan tagjai, és amelyek életükben meghatározó szerepet töltenek be. Egy pozitívan értékelt tagsági csoport pozitív önértékelést és a valahová tartozás biztonságát nyújtja az egyén számára. A szociális identitás azonban nem abszolút, hanem relációs kategória: a saját csoport értékét más, vele azonos típusú külső csoportoktól való pozitív

megkülönböztetettsége adja. A pozitív szociális identitás igénye csoportközi összehasonlításhoz és elfogultsághoz vezet, azaz a saját csoport fel- és a külső csoport leértékeléséhez, amely megjelenhet sztereotipizálásban, diszkriminatív viselkedésben vagy agresszív versengésben [9, 8, 7]. Az értékelésbeli elfogultság a csoport jólétét fenyegető, kiélezett konfliktushelyzetekben felerősödik, megerősítve a csoportkohéziót és a kollektív azonosságtudatot. Kísérletek demonstrálták, hogy az elfogultság a verbális viselkedést is befolyásolja [11].

3.2 A csoportközi értékelés traumafeldolgozásra vonatkozó mutatói

Narratív pszichológiai megközelítésben a csoportidentitást ért trauma kollektív feldolgozása olyan narratív rekonstrukciós folyamat, amely az elfogadhatatlan veszteségélmény narratív leképezésével indul, majd a lezárt múlthoz tartozó, a csoporttörténet egészéhez koherens módon illeszkedő és a fenntartható, pozitív azonosságtudathoz hozzájáruló narratívához vezet. A jelen tanulmány tárgyát képező feltevés szerint a narratív csoportközi értékelés legalább három olyan jelentésdimenzióval bír, amelyek feltételezhetően a traumafeldolgozás folyamatának lényeges eszközévé teszik. Az alábbiakban e három, a narratívákban mennyiségileg mérhető dimenziót és a feldolgozási folyamatra vonatkozó implikációikat határozzuk meg.

Az egyes dimenziókat mindhárom esetben több különböző tartalmi kategória gyakorisági eloszlása, az ezekből létrejövő mintázat jelenti, nem pusztán egyetlen kategória előfordulási gyakorisága. A feldolgozási folyamattal való összefüggésüket oly módon határozzuk meg, hogy a feldolgozatlan és a feldolgozottság felé tartó trauma konstrukciójára jellemző mintázatok közötti különbségeket definiáljuk.

1) Csoportközi elfogultság: pozitív és negatív valencia

A feldolgozatlan trauma konstrukciójában szignifikáns aszimmetria jelenik meg a saját csoport és a külső csoportok értékelésében, a csoportközi elfogultság tendenciájának megfelelően: a saját csoport értékelését pozitív, a külső csoportét negatív túlsúly jellemzi. Ez a mintázat azt implikálja, hogy a saját csoportot nem terheli felelősség a traumatikus esemény bekövetkeztéért, nem vállalja annak következményeit, valamint jóvátételre tart igényt, hiszen a negatívan értékelt esemény felelőssége és jóvátétele a negatívan értékelt szereplőt terheli. Ebben a dimenzióban a feldolgozási folyamat előrehaladását az jelzi, hogy a csoportközi értékelés aszimmetriája csökken, a saját csoport összességében kevésbé pozitívan, a külső csoport pedig kevésbé negatívan értékelődik. E mintázat a negatív eseményért és következményeiért viselendő felelősség megoszlását implikálja. Egy önreflektív, a veszteségre külső, objektívebb nézőpontból tekintő perspektívát alkalmaz az elbeszélés, amely a trauma feldolgozásában fontos tényezőt jelent [6].

2) A jelenre vonatkozó relevancia hangsúlya: narrátori vs. szereplői értékelői perspektíva

A narrátor és a saját csoportot képviselő szereplők értékelései képviselik a csoport értékelő perspektíváját a történelmi narratívákban. Lényeges, hogy míg a narrátor a saját csoport jelenbeli perspektíváját képviseli az esemény vonatkozásában, addig a szereplők értékelései a múlthoz tartoznak, mivel maguk a szereplők is a múltbeli esemény részeként jelennek meg az elbeszélésben. A következő példák illusztrálják a narrátori és szereplői értékelés közti különbséget, lényegében azonos értékelő tartalom mellett. Narrátori értékelés: *A békefeltételek felháborítóan igazságtalanok voltak*. Szereplői értékelés: *A békefeltételeket az ország felháborodott tiltakozással fogadta.*

Feltevésünk szerint a csoporttrauma kezdeti konstrukciójában az értékelések viszonylag nagy hányadát (az időben később keletkezett narratívákhoz képest) a narrátor teszi. Ha ebben a perspektívában hangsúlyos a csoportközi értékelés, az az esemény jelenre vonatkozó relevanciáját, vagyis lezáratlanságát tükrözi. A feldolgozási folyamat során a narrátori értékelések aránya csökken a megelőző konstrukciókhoz képest, ami az esemény jelenre vonatkozó jelentőségének csökkenését implikálja, a jelen és múlt közti pszichológiai távolság növekedését, a rekonstrukciós folyamat az esemény lezárása felé tart.

3) Érzelmi fókusz: érzelmi vs. kognitív értékelés

A narrátor érzelmi és kognitív jellegű értékeléseinek relatív aránya az eseményhez való viszonyulás érzelemtelítettségének mutatója. Az érzelmi-kognitív megkülönböztetés alapja hasonló Pennebaker [6, 16] osztályozásához, amelyet traumatikus életeseményekről szóló egyéni beszámolók tartalomelemzésében használt. Ugyanakkor a csoportközi értékelés szűkebb metszetére vonatkozó vizsgálatunk olyan kategóriarendszert használt, amelyben az érzelmi kifejezések mellett a kódoló nyelvi intuíciója alapján érzelmi reakciókat implikáló morális ítéletek is az érzelmi értékelések közé tartoznak (pl. *kegyetlen, hősies*), míg a kognitív értékelések közt a kognitív mechanizmusokra utaló értékeléseken túlmenően (pl. *átgondolatlan, megfontolt*) helyet kapnak a racionális szempontú illetve általános, érzelmeket nem vagy nem jellemzően implikáló értékelések (pl. *hibás, jó*).

Kollektív traumáról szóló csoporttörténeti szövegekben ahhoz hasonló tendencia várható, amit Pennebaker talált egyéni beszámolókban: A narrátori értékelések körében kezdetben viszonylag nagy arányban szerepelnek érzelmileg telített értékelések, szemben a kognitív értékelésekkel, a kiértékelés érzelmi fókuszának megfelelően. A feldolgozási folyamat során az érzelmileg telített értékelések aránya csökken, szemben a kognitív értékelésekével, amely az érzelmi kontroll és a racionális belátás erősödését implikálja, így az eseményt tárgyként kezelő (s nem élményként megélő) külső, objektívebb perspektíva nagyobb mértékben érvényesül.

4 Vizsgálat: A csoportközi értékelés mint a csoporttrauma érzelmi feldolgozásának mutatója a trianoni béke tankönyvi narratíváiban

4.1 A trianoni béke mint nemzeti trauma

A narratív csoportközi értékelés és a traumafeldolgozási folyamat közti összefüggések vizsgálatához a trianoni békét választottuk releváns eseménynek. Az 1920-ban hatályba lépő trianoni békeszerződés a magyar történelem egyik fő traumája, melyet a nemzet a mai napig nem dolgozott fel maradéktalanul, ugyanakkor a szerződéskötés óta eltelt kilenc évtized elegendő idő arra, hogy a traumatizáció állapotából számottevő elmozdulás történjen a traumatikus esemény integrációja felé. A feldolgozás befejezetlenségére utal az, hogy a határon túli magyarok ügye sem Magyarországon, sem az érintett szomszédos országokban nem ért nyugvópontra, hogy hazánkban mind a mai napig vannak a béke revízióját szorgalmazó csoportosulások, és hogy Trianon története máig nem nyerte el kanonikus formáját. (A közelmúltban állami beavatkozás révén kísérelték meg egységesíteni a Trianonról szóló tananyagot a közoktatásban [5].)

A traumatikus eseményről szóló, 1920 után kiadott magyar történelem tankönyvek fejezetei kiváló szöveges adatbázist nyújtanak a csoportközi értékelés és kollektív traumafeldolgozás közti összefüggések ellenőrzésére. A békekötést közvetlenül követő időszaktól a jelenkorig kiadott tankönyvek Trianon-fejezeteinek hosszmetszeti elemzésével nyomon követhetővé válik a traumafeldolgozással összefüggésbe hozott nyelvi-szemantikai dimenziók változása, s e változások a feldolgozási folyamat keretében értelmezhetők.

4.2 Hipotézisek

A csoportközi értékelés és a feldolgozási folyamat összefüggésére vonatkozóan egyfajta nullhipotézist állítottunk fel alapfeltevésként. Azt feltételeztük, hogy a traumatikus veszteség elfogadásának folyamatát az időtényezőn kívül semmi egyéb nem befolyásolja, mintha légüres térben, társadalmi vákuumban zajlana. Ennek előnye, hogy az értékelés mutatóira vonatkozóan egyértelmű predikciókat lehet tenni, s minden, ettől való jelentős eltérés az eredményekben olyan mozzanat hatásaként értelmezhető, amely a feldolgozás akadályaként jelenik meg.

A csoportközi értékelés három tartalmi dimenziójára vonatkozóan a 3.3 fejezetben leírt általános feltevések alapján a következő predikciókat tettük. Az (1) *értékelés csoportközi aszimmetriájával* mértéke az idő múlásával párhuzamosan csökken, azaz a saját csoport pozitív értékelése és a konfliktusos külső csoportok negatív értékelése egyaránt csökkenő tendenciát mutat. Az (2) *értékelői perspektívára* a narrátori értékelések aránya időben csökkenni fog, így a pszichológiai távolság jelen és múlt között fokozódó hangsúlyt kap a szövegekben. A (3) *narrátori értékelések tartalmára* az érzelmi értékelések aránya fokozatosan csökkenő tendenciát követ, az érzelmi fókusz dominanciája így csökken, míg a racionális belátásé nő.

4.3 Minta

Az Országos Széchényi Könyvtárban elérhető középiskolai tankönyvek adták a mintavétel bázisát. A hosszmetszeti elemzést szolgáló korpuszt 1920 és 2000 között kiadott középiskolai történelem tankönyvek Trianonról szóló fejezetei alkották. A jelölt időszakon belül 10 éves felbontású mintavételt végeztünk: mindazon Trianon-fejezetek bekerültek a mintába, amelyek kerek esztendőkben (1920, 1930 stb.) kiadott tankönyvekben szerepeltek. Ily módon 1920 és 2000 között 10 alkorpuszt kaptunk, melyek számszerű értékelésmutatóiból kíséreltünk meg következtetéseket levonni a feldolgozási folyamatra vonatkozóan.

4.4 Eljárás

A szövegek elsődleges elemzése a NARRCAT számítógépes nyelvi elemzőeszköz értékelés moduljával történt. A NARRCAT moduljai a NooJ nyelvtechnológiai rendszerben működnek [10], amely több nyelvben lehetővé teszi nagy terjedelmű digitalizált szövegkorpuszok morfológiai és szintaktikai elemzését, és erre épülő algoritmusok révén meghatározott nyelvi alakzatok azonosítását. Az értékelés modul az elemzést szolgáló, szófaj és valencia szerinti annotációs jegyekkel jelöli meg az értékelést hordozó kulcsszavakat, amelyek e szempontok szerint külön szótárakba kerültek. Az 1. táblázat rendszerezi a modul szótárait, az egyes szótárakra vonatkozó példákkal és elemszámokkal. Az értékelő kulcsszavak szófaj szerint lehetnek melléknevek, igék, főnevek és határozók. A melléknév- és igeszótárakat az MTA Nyelvtudományi Intézetének használati gyakoriság szerint összeállított digitális szótáraiból állítottuk össze, két független bíráló választásai alapján. A valencia szerint pozitív és negatív értékelések külön szótárakba kerültek. Mivel az értékelések elsősorban tulajdonságokban, valamint cselekvésekben realizálódnak, melyeket melléknevekkel, illetve igékkel fejez ki a nyelv, így a főnév- és határozószótárakat az értékelő melléknevekből és igékből képzett főnevekből, illetve határozókból hoztuk létre. Ez az oka annak, hogy a szótárak elemszámai ismétlődést mutatnak. Az értékelő jellegű érzelmi, illetve mentális állapotokat a NARRCAT önálló érzelem modulja kezeli.

1. táblázat: Az értékelés modul szófaj és valencia szerint osztályozott szótárai, példákkal és az egyes szótárak elemszámával.

Szófaj		Pozitív	db	Negatív	db
Melléknév		*bölcs*	317	*jogtalan*	582
Ige		*vitézkedik* *éljenez*	122	*kizsákmányol* *tiltakozik*	317
Főnév	Melléknévből	*bölcsesség*	317	*jogtalanság*	582
	Igéből	*éljenzés*	122	*tiltakozás*	317
Határozó	Melléknévből	*bölcsen*	317	*jogtalanul*	582
	Igéből	*éljenezve*	122	*tiltakozva*	317

Az értékelések referenciáinak azonosításához (ki kit értékel) és érzelmi-kognitív tartalom szerinti osztályozásához további, szoftveresen támogatott manuális elemzés-

re van szükség. Jelenleg fejlesztések zajlanak e funkciók automatizálása céljából (a szereplőazonosítás korábbi fejleményeiről lásd [17]).

Az elemzés második fázisában a szövegben annotált értékeléseket az Atlas.ti elemzőszoftverrel [4] kódoltuk az értékelés tárgya (magyarok, Antant, Kisantant) és valenciája (pozitív, negatív), az értékelői perspektíva (narrátor, szereplő), valamint narrátori értékelések esetében az értékelés tartalma (érzelmi, kognitív) szerint. Az értékelés tárgya szerinti kódoláskor a magyarok kategóriájába került a nemzet mint egész, és az azt képviselő csoportok, illetve egyéni szereplők, valamint a narrátor mint értékelő. Az Antant, illetve a Kisantant kategóriáiba került a két hatalmi csoport mint egész, az egyes tagnemzetek és az azokon belüli kisebb csoportok, illetve egyéni szereplők. A valencia szerinti kódolás már az elemzés első, automatizált fázisában megtörtént. A perspektíva szerinti kódolásban narrátori és szereplő perspektívát különítettünk el, aszerint, hogy ki értékel a szövegben. Csak a magyarok perspektíváját képviselő értékeléseket vontuk be az elemzésbe, tehát a narrátor és a magyar szereplők értékeléseit. A tartalom szerinti kódolásban az érzelmi és kognitív kategóriákat különítettük el. E tekintetben a kódolást végző szerző egyéni nyelvi intuíciójára hagyatkozott.

4.5 Eredmények

4.5.1 Az értékelés csoportközi aszimmetriája (tárgy és valencia)

Az adatelemzés első lépéseként az egyes csoportokra (magyarok, Antant, Kisantant) vonatkozó pozitív és negatív értékelések gyakoriságait vizsgáltuk. Mind a 10 alkorpusz esetében külön kimutatást készítettünk, ezek adták az adatértelmezés alapját. A csoportközi értékelés hasonlóságai szerint a 10 alkorpusz négy nagyobb szegmensre osztható: 1920-1940, 1950, 1960-1980, 1990-2000 (2. táblázat). Az adatok részletes elemzésére lentebb kerül sor (4.5.3 fejezet), de annyit szükséges itt megállapítani, hogy a négy szegmens által lefedett időszakok megközelítőleg megfeleltethetők négy egymást követő politikai érának: Horthy-korszak (1920-1940), Rákosi-korszak (1950), Kádár-korszak (1960-1980), Rendszerváltás utáni időszak (1990-2000). Ez azt sugallja, hogy a mindenkori uralkodó politikai ideológia rányomta bélyegét a Trianon-reprezentációkra. Az egyes politikai éráknak az eredmények értelmezése szempontjából releváns jellemzőit szintén lentebb ismertetjük (4.5.3 fejezet).[1]

[1] A számszerű adatok eloszlása alapján megállapított korszakhatárokat természetesen nem úgy tekintjük, mint amelyek éles választóvonalat képeznek a változó történelemfelfogások között, azonban az évtizedes mintavételi felbontás nem engedi e felfogásbeli változások finomabb rekonstrukcióját. Ezzel együtt a különböző korszakok Trianon-reprezentációira vonatkozó megállapításainkat alapvetően érvényesnek fogadjuk el.

2. táblázat: Az évtizedenkénti eloszlások alapján kapott négy alkorpusz adatai: évtizedenkénti szószám, értékelés %-os aránya, értékelések eloszlása tárgy és valencia szerint (szövegterjedelemhez mért arányban, zárójelben a nyers gyakoriságok), perspektíva szerint, narrátori értékelések tartalom szerinti eloszlása.

Időszak	Szó/ évtized	Értékelés %	Tárgy / Valencia				Értékelői perspektíva		Narrátori értékelések tartalma	
			Külső csoportok		Magyarok					
			Poz.	Neg.	Poz.	Neg.	Narrátor	Szereplő	Érzelmi	Kognitív
1920-1940 Horthy	2951	1,5	1 (1)	66 (58)	49 (43)	16 (14)	104 (92)	12 (11)	71 (63)	33 (29)
1950 Rákosi	3138	1,5	3 (1)	29 (9)	25 (8)	83 (26)				
1960-1980 Kádár	464	0,9	0 (0)	7 (1)	14 (2)	50 (7)				
1990-2000 Rdszváltás	5419	0,6	2 (2)	41 (44)	8 (9)	7 (8)	30 (32)	24 (26)	17 (18)	13 (14)

4.5.2 Az átlagos szövegterjedelem és az értékelések aránya korszakonként

A következő lépésben megvizsgáltuk, hogy a négy korszakban hogyan alakul a Trianon-szövegek átlagos terjedelme és az értékelés átlagos, szövegterjedelemhez mért százalékos aránya (2. táblázat). Az egyes korszakokon belüli, évtizedenkénti átlagos szószám (összes szószám / évtizedek száma az adott korszakban) mutatja a legjobban, hogy milyen viszonylagos hangsúllyal jelent meg az egyes korszakokban Trianon a tankönyvekben. A Horthy- és a Rákosi-korszak évtizedenkénti átlagos szövegterjedelme megközelítőleg azonos (2951, 3138), majd a Kádár-korszakban drasztikus esés figyelhető meg (464), végül a rendszerváltás utáni időszakban a szószám az összes többi korszak fölé emelkedik (5419).

Az értékelés korszakonkénti, szövegen belüli százalékos aránya (összes értékelés / összes szószám × 100) szintén az esemény viszonylagos jelentőségét, a nemzettörténet szempontjából vett fontosságának változását mutatja. A Horthy- és a Rákosi-korszakban az értékelés aránya azonos (1,5%), majd ehhez képest a következő két korszakban fokozatosan csökken (0,9%, 0,6%).

4.5.3 A csoportközi megkülönböztetés négy mintázata

Az egyes korszakokban megfigyelt, tárgy és valencia szerinti eloszlási mintázatok *közti* eltérés statisztikailag szignifikáns (Pearson $\chi^2 = 135{,}926$; $p = ,000$), tehát a teljes adathalmaz négy történelmi korszak, illetve politikai éra szerinti felbontása releváns. (A cellánkénti gyakorisági adatokat az egyes alkorpuszok esetében a következő képlettel kaptuk: [értékelések nyers gyakorisága / alkorpusz szószáma × 10.000] – egész számra kerekítve. A külső csoportok két kategóriájára, az Antantra és a Kisantantra vonatkozó adatokat összevontan kezeltük, a rájuk vonatkozó értékelések korszakokon belüli eloszlásainak hasonlósága, illetve az értékelések viszonylag kis száma miatt.)

Az egyes korszakokon *belül* a külső és a saját csoportokra vonatkozó pozitív-negatív értékelések eloszlásai közötti különbségek statisztikai szignifikanciáját ugyanezzel az eljárással vizsgáltuk. Az alábbiakban mutatjuk be az egyes korszakokban megfigyelt tendenciákat (lásd 2. táblázat).

1) 1920-1940 (Horthy-korszak)

A békeszerződést közvetlenül követő időszakban tisztán megmutatkozik a csoportközi elfogultság tendenciája az értékelések eloszlásában. A külső csoportoknál a negatív értékelések dominálnak a pozitívakkal szemben: 1 pozitív, 58 negatív értékelés. Ugyanakkor a magyarokra vonatkozó értékelések ezzel ellentétes tendenciát mutatnak: 43 pozitív, 14 negatív értékelés. A külső csoportokra vonatkozó, összesített értékelések és a magyarokra vonatkozó értékelések valencia szerinti eloszlásai szignifikánsan különböznek egymástól (Pearson $\chi^2 = 76{,}555$; $p = {,}000$).

2) 1950 (Rákosi-korszak)

Az 1950-es szövegekben az előző korszakhoz képest egy teljesen más mintázat jelenik meg. Egyrészt itt lényegesen kevesebb a külső csoportokra, mint a magyarokra vonatkozó értékelés: külső csoportok összesen: 10; magyarok: 34 értékelés. Másrészt nem csak a külső csoportok, hanem a magyarok esetében is lényegesen több a negatív, mint a pozitív értékelés: külső csoportok: 1 pozitív, 9 negatív; magyarok: 8 pozitív, 26 negatív értékelés. A két eloszlás között nincs szignifikáns különbség (Pearson $\chi^2 = 2{,}927$; $p = {,}087$). A mintázat hátterében az áll, hogy e korszak szövegeiben Trianon története bizonyos értelemben átkereteződik, mégpedig az ekkor uralkodó szovjet szocialista ideológiának megfelelően. Az eseményben érintett csoportok már nem Magyarország és a győztes hatalmak, hanem a nyugati imperialisták és a szovjet forradalmárok, továbbá ezen a felosztáson belül a szövegek elsősorban a nyugatbarát és a szovjetbarát magyarok szerepére koncentrálnak, melyet azok a békéhez vezető eseményekben betöltöttek.

3) 1960-1980 (Kádár-korszak)

Az 1960-1980 közötti időszak szövegei hasonló mintát mutatnak az előző korszak szövegeihez, ugyanakkor lényegesen kevesebb az értékelések gyakorisága: külső csoportok: 0 pozitív, 1 negatív; magyarok: 2 pozitív, 7 negatív értékelés. A külső csoportokra és a magyarokra vonatkozó értékelések eloszlásai közötti különbség az előző korszakhoz hasonlóan itt sem szignifikáns (Fisher's Exact Test: $p = {,}331$). Az értékelések kis száma részben annak köszönhető, hogy ebben a korszakban sokkal kevesebb és rövidebb szöveg került kiadásra (0,8 szöveg ill. 464 szó / évtized), mint az előzőben (3 szöveg, ill. 3138 szó / évtized). Másfelől az 1960-1980 alkorpuszban a szövegterjedelemhez mért arányokat tekintve is sokkal kevesebb, feleannyi értékelés van, mint az 1950-es alkorpuszban (71 és 140 a két arányszám).

4) 1990-2000 (rendszerváltás utáni időszak)

A rendszerváltás utáni, egyben a szovjet uralom lezárulása utáni időszakban Trianon újra nemzeti keretben tematizálódik, ahogyan a Horthy-korszakban. Egyrészt visszatér a Magyarország – győztes hatalmak reláció, másrészt újra nagyobb hangsúlyt kap

az esemény, amely a szövegterjedelem előző korszakhoz viszonyított jelentős növekedésében mutatkozik meg (évtizedenként 5419 szó szemben a 464 szóval). Részben visszatér a Horthy-korszakban feltárt értékelési mintázat is. A külső csoportokra vonatkozó értékelések újra erős negatív túlsúlyt mutatnak: Antant: 2 pozitív, 35 negatív; Kisantant: 0 pozitív, 9 negatív értékelés. Ugyanakkor a magyarokra vonatkozó értékelések eloszlásában nem jelenik meg a Horthy-korszak szövegeiben talált, csoportközi elfogultságra jellemző pozitív dominancia, az eloszlás ehelyett kiegyenlített: 9 pozitív, 8 negatív értékelés. (A külső csoportokra és a magyarokra vonatkozó értékelések eloszlásai közti különbség ezzel együtt szignifikáns: Fisher's Exact Test: p = ,000) További fontos különbség a két korszak szövegei között, hogy az értékelések szövegterjedelemhez viszonyított aránya lényegesen kisebb a jelenkorban, mint a Horthy-korszakban (0,6% szemben az 1,5%-kal).

4.5.4 Értékelői perspektíva és narrátori értékelések tartalma

Amint fentebb (4.5.3) kifejtettük, a szocializmus idejére eső két alkorpusz szövegeiben Trianon a nyugatellenes szovjet ideológia értelmezési keretében reprezentálódik, amely a kétpólusú világ harcának részévé teszi a békeszerződés történetét, ezzel háttérbe szorítva a nemzeti identitást ért veszteséget. Ebből fakadóan a trauma érzelmi feldolgozottságának állapotára vonatkozóan csak a Horthy-korszak és a rendszerváltás utáni időszak alkorpuszai informatívak, így a narrátori és szereplői értékelői perspektíva relatív arányát, valamint a narrátori értékeléseken belül az érzelmi és kognitív értékelések arányát e két alkorpuszban vizsgáltuk (lásd 2. táblázat).

A kétféle értékelői perspektíva relatív hangsúlyát tekintve, míg a Horthy-korszak szövegeiben összesítve több mint nyolcszor annyi a narrátori, mint a szereplői értékelés (92 és 11), addig a rendszerváltás utáni alkorpuszban a két gyakoriság csaknem azonos (32 és 26). Az arányszámokban kifejezett eloszlások szignifikánsan különböznek egymástól (Pearson $\chi^2 = 25{,}668$; p = ,000).

A narrátori értékelések tartalmát vizsgálva hasonló irányba mutató változás figyelhető meg. Míg a Horthy-korszak szövegeiben több mint kétszer annyi az érzelmi, mint a kognitív értékelés (63 és 29), addig a rendszerváltás utáni szövegekben a két gyakoriság jóval kiegyenlítettebb eloszlást mutat (18 és 14). A két eloszlás közti különbség azonban nem szignifikáns (Pearson $\chi^2 = 1{,}390$; p = ,238), az érzelmi-kognitív arány változása tehát csak tendenciaként értelmezhető.

5 Megvitatás

A kollektív trauma feldolgozására vonatkozó fő hipotézisünk azt jósolta, hogy mind a csoportközi megkülönböztetés mértéke, mind a narrátori értékelések aránya, mind pedig ezen belül az érzelmi értékelések aránya az idő múlásával párhuzamosan csökkenő tendenciát mutat, az ettől eltérő irányú tendenciák pedig a feldolgozást akadályozó tényező hatásaként értelmezhetők. Láttuk, hogy a mindenkori uralkodó politikai ideológia jelentősen befolyásolja a reprezentációs folyamatot, hiszen az évtizedenkénti adateloszlások alapján négy olyan, egymástól eltérő értékelési mintázatot sikerült azonosítani, amelyek a történelmi időben való elhelyezkedésük alapján négy

politikai éra hatásának feleltethetők meg. A négy eloszlási mintázatot időbeli linearitásban vizsgálva az érzelmi feldolgozás szempontjából, azt látjuk, hogy a traumatizáció és retraumatizáció időszaka után beköszöntő szovjet szocialista diktatúra a nemzeti szuverenitás elnyomása révén közel öt évtizeden keresztül megakadályozta a nemzeti identitást ért trauma tematizációját, ezáltal késleltette az érzelmi feldolgozás folyamatát. A Kádár-korszakban a represszió az alacsony szövegterjedelemben jelenik meg. A rendszerváltás után újra az önálló nemzet összefüggésében tárgyalt trianoni béke narratívái a konfliktusban álló csoportok értékelése szempontjából inkább hasonlítanak a revizionista Horthy-korszak narratíváira, mint a megelőző szocialista éra történeteire. A Horthy-korszak és a rendszerváltás utáni kor konstrukcióinak közös pontja a külső csoportok értékelését jellemző erőteljes negatív túlsúly, ami azt mutatja, hogy a jelenkori Trianon-konstrukció megőrzi az áldozat-elkövető viszonyt: a nemzet továbbra is áldozatként jelenik meg, a világháború győztes hatalmaira pedig olyan felelősséget ruház, amely máig nem évült el. Hatékony érzelmi feldolgozásról tehát nem beszélhetünk a vizsgált nyolcvan évet tekintve.

Más mutatók ugyanakkor azt tükrözik, hogy a feldolgozás a Horthy-korszakhoz mint zéróponthoz képest jelentős elmozdulást mutat. Egyfelől a jelenkor lényegesen kevesebb értékeléssel, alacsonyabb „érzelmi hőfokon" beszéli el a traumát, mint Horthy kora, illetve hiányzik a nemzet glorifikációja is. Mindez arra utal, hogy a veszteség véglegesként jelenik meg, a jelenkori szövegek ennek elfogadását közvetítik, sem explicit, sem implicit módon – az értékelés eszközei révén – nem utalnak a veszteség előtti állapothoz való visszatérés lehetőségére avagy igényére. A múlt tehát ebben az értelemben lezárul a narratívákban. Másfelől a jelenkori narratívák a Horthy-korszakhoz viszonyítva pszichológiai távolságot teremtenek a traumatikus múlt és a jelen között. Egyrészt jelentős mértékben csökken a narrátori értékelések aránya, ami az eseményhez való jelenbeli viszonyulás dimenzióját képviseli a narratív konstrukcióban, s e perspektíva hangsúlyának csökkenésével a jelen és a múlt kapcsolata gyengül, a múlt jelenre vonatkozó relevanciája háttérbe szorul. Másrészt a fennmaradó narrátori értékeléseken belül tendencia mutatkozik az érzelmi értékelések csökkenésére, tehát a jelenkori narratívák egy racionálisabb szempontú viszonyt érvényesítenek a Horthy-korszakhoz képest. Ez a mozzanat szintén távolságot teremt múlt és jelen között, azáltal, hogy a veszteség érzelmi aspektusát távolítja a befogadótól.

A feldolgozottság jelen állapotára vonatkozó következtetéseket összegezve úgy tűnik, hogy bár Trianon narratívái a béke által szentesített gazdasági, társadalmi és politikai veszteség véglegességének elfogadását közvetítik, illetve a veszteség élményét távolítják a jelentől, ugyanakkor nem írják felül az áldozat-elkövető viszonyt, a nemzet áldozat szerepét. Ez a perspektíva kívülre helyezi a felelősséget és az események feletti kontrollt, továbbá állandósítja a jóvá nem tett veszteségből fakadó deprimált és ellenséges érzelmi viszonyulást. Ezek a konstrukciós mozzanatok általános mintaként megjelennek a nemzeti múlt más eseményeinek jelenkori narratíváiban is [1, 12], s feltételezhető, hogy a nemzeti identitást megszólító jelenbeli események és jövőképek kapcsán szintén konstrukciós elvekként működnek, amelyek azonban maladaptív megküzdési módokat facilitálnak.

Hivatkozások

1. Fülöp É.: A történelmi pálya és a nemzeti identitás érzelmi szerveződése. PhD értekezés. (2010) Letöltve: http://pszichologia.pte.hu/files/tiny_mce/D-2010-Fulop%20Eva.pdf
2. László J.: A történetek tudománya. Bevezetés a narratív pszichológiába. ÚMK, Bp. (2005)
3. László J.: Narratív Pszichológia. Pszichológia, Vol. 28., No. 4 (2008) 301–317
4. Muhr, T.: User's Manual for ATLAS.ti 5.0 (2004) Letöltve: http://www.atlasti.com/up loads/media/atlman_01.pdf
5. Oktatáskutató és Fejlesztő Intézet: A nemzeti összetartozás napja. Pedagógiai háttéranyag. (2011) Letöltve: http://www.kormany.hu/download/0/cd/30000/A%20nemzeti%20%C3%B6sszetartoz%C3%A1s%20napja.pdf#!DocumentBrowse
6. Pennebaker, J. W.: Putting stress into words: Health, linguistic, and therapeutic implications. Behaviour Research and Therapy, Vol. 31(6). (1993) 539-548
7. Pettigrew, F. T.: The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. Personality and Social Psychology Bulletin Vol. 5, No. 4 (1979) 461–476
8. Sherif, M.: In Common Predicament: Social Psychology of Intergroup Conflict and Cooperation. Boston: Houghton Mifflin (1966)
9. Sherif, M., Harvey, O. J., White, J., Hood, W., Sherif, C.: Intergroup Conflict and Cooperation: The Robber's Cave Experiment. Norman: University of Oklahoma, Institute of Social Relations (1961)
10. Silberztein, M.: NooJ manual. (2003) Letöltve: http://www.nooj4nlp.net/NooJManual. pdf
11. Szabó Zs. P., Banga Cs., Ferenczhalmy R., Fülöp É., Szalai K., László J.: A nyelvbe kódolt társas viszonyok. Az implicit szemantika szociálpszichológiai kutatása. Pszichológia Vol. 30, No. 1 (2010) 1–16
12. Szalai K.: Az ágencia nyelvi jegyei. Az aktív és passzív igék szerepe a narratívumokban. PhD értekezés. (2011) Letöltve: http://pszichologia.pte.hu/files/tiny_mce/doktori/D-2011-Szalai%20Katalin.pdf
13. Tajfel, H.: Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, New York, NY (1978)
14. Tajfel, H.: Human groups and social categories: Studies in social psychology. Cambridge University Press, Cambridge (1981)
15. Tajfel, H., & Turner, J. C.: The social identity theory of intergroup behavior. In: Worchel, S., Austin, W. (Eds.) The Psychology of Intergroup Relations (2nd ed.). Chicago Nelson-Hall. (1986)
16. Tausczik, Y., Pennebaker, J. W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, Vol. 29 (2010) 24–54
17. Vincze O., Gábor K., Ehmann B., László J.: Technológiai fejlesztések a Nooj pszichológiai alkalmazásában. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. JATE Press, Szeged (2009) 285–294

Szemantikus szerepek vizsgálata magyar nyelvű szövegek narratív pszichológiai elemzésében[1]

Ehmann Bea[1], Lendvai Piroska[2], Fritz Adorján[3], Miháltz Márton[2], Tihanyi László[2]

[1] MTA Pszichológiai Kutatóintézet
1132 Budapest, Victor Hugó u. 18-22.
{ehmannb}@mtapi.hu
[2] Nyelvtudományi Intézet,
1068 Budapest, Benczúr u. 33.
{piroska,tihanyil}@nytud.hu; {mmihaltz}@gmail.com
[3] Pécsi Tudományegyetem, Pszichológiai Intézet
7624 Pécs, Ifjúság útja 6.
{kifino}@gmail.com

Kivonat: A narratív pszichológiai tartalomelemzés és a korpusznyelvészet több éve folytatott közös projektje a szemantikus szerepek és a narratív pszichológiai modulok összekapcsolása egyének és csoportok énelbeszéléseinek elemzéséhez. A két munkacsoport korábbi együttes fejlesztései a szemantikus szerepek felismerését úgy oldották meg, hogy a MetaMorpho nyelvi elemzés morfoszintaktikai és szemantikai kimenetét összekapcsolták a NooJ eszköz procedúráival. A jelen munka célja, hogy a korábbi törekvések továbbfejlesztésével magyar nyelvű szövegekben felismerhető váljon az ágencia, és ennek nyelvi kifejező elemeihez automatikusan hozzárendelhetőek legyenek az Ingroup/Outgroup pszichoszemantikai kategóriák. Ekképp a tudományos narratív pszichológia a *semantic role labeling* nyelvészeti terület új alkalmazójaként jelenik meg.

1 Miért fontos a tudományos narratív pszichológia számára a szemantikus szerepek vizsgálata?

A Tudományos Narratív Pszichológia (TNP) a szelf- és csoportnarratívákban azonosítható pszichológiai jelenségek longitudinális, kvantitatív vizsgálatára szolgáló, Magyarországon kifejlesztett elmélet, melynek számos empirikus alkalmazása létezik a szociálpszichológia, a személyiség- és a klinikai pszichológia területén [9]. Az elmélet módszere, a Narratív Pszichológiai Tartalomelemzés (NPTA) fejlődésének alapja a magyar korpusznyelvészekkel és nyelvtechnológusokkal történő együttműködése, melynek során a Narratív Pszichológiai Munkacsoport számos pszichoszemantikai taxonómiát és algoritmust fejlesztett ki [18,10,9] a NooJ nyelvészeti fejlesztési környezet keretében [16].

[1] A kutatást az OTKA 81633K pályázat támogatta.

Az eddig kifejlesztett NPTA-algoritmusok, TNP-modulok a következők: *Aktivitás-Passzivitás* [17], *Érzelem* [7], *Kognitív folyamatok* [21,20], *Értékelés* [1,2], *Intencionalitás* [6], *Tagadás*; *Én- és Mi Referencia* [8], *Perspektíva* [13], valamint a *Szubjektív Időélmény* [5].

E fejlesztésekről és a velük kapott empirikus eredményekről az elmúlt évek során a Munkacsoport a Számítógépes Nyelvészeti Konferenciákon és nemzetközi közleményekben is széleskörűen beszámolt[2].

A Narratív Pszichológiai Tartalomelemző NooJ algoritmusok (modulok) a TNP két fő területén használatosak. A strukturális megközelítés azt vizsgálja, hogy a vizsgált kategóriák – elsősorban az elbeszélői perspektíva, az időélmény és az értékelés – miképpen változnak az énelbeszélések és a csoportelbeszélések egészének belső szerkezetében [4,14,13].

A másik vizsgálódási kör a mintázatelemzés, ami az egyes szógyakoriságok együttjárásából von le pszichológiai következtetéseket: ennek egyik példája, hogy kiscsoportok beszámolóiban a negatív érzelemmarkerek és a szelfreferencia magas együttes aránya csoporton belüli konfliktust jelez; a negatív érzelemmarkerek és a mi-referencia magas aránya viszont együttesen a csoport fenyegetettségére utal [3].

A tudományos narratív pszichológia annyiban lép túl a hagyományos pszichológiai tartalomelemzési koncepción, hogy nem elégszik meg a pszichológiai tartalmak puszta számlálásával és strukturális vagy mintázatelemzésével, hanem azt is vizsgálja, hogy az adott érzelem, kogníció vagy cselekvés milyen cselekvőhöz, illetve milyen elszenvedőhöz tartozik. Minthogy a Narratív Pszichológiai Munkacsoport kiemelt kutatási területe a nemzeti és európai identitás vizsgálata, sarkalatos kérdés, hogy valamely történelmi esemény vagy korszak beszámolóiban a TNP által vizsgált kategóriák a saját csoporthoz vagy a külső csoporthoz tartoznak.

A cselekvő és az elszenvedő kérdésköre a pszichológiában hagyományosan az ágenciakutatás területéhez kapcsolódik. A személyiség- és a klinikai pszichológiában ez főként az énhatékonyság megítélésében fontos, a szociálpszichológiában pedig a humán ágens és a humán elszenvedő egyén vagy csoport nyelvi megjelenítése vagy ennek hiánya a társas-társadalmi-hatalmi felelősség felvállalását, hárítását vagy elkendőzését teszi vizsgálhatóvá.

Ezért fontos a TNP számára a szemantikus szerepek (Semantic Role Labeling) vizsgálatára szolgáló elemzőeszköz kifejlesztése.

2 A magyar és európai történelem narratív pszichológiai korpuszai

Az MTA Pszichológiai Intézetének Oral History Archívumában a következő elektronikus korpuszok állnak rendelkezésre:

- *Történelemkönyv korpusz:* a magyar történelemről szóló könyvek részletei a 10 legfontosabb eseményről, 1900-tól 2000-ig, 10 éves bontásban (kb. 200000 szó);

[2] Cf. http://narrativpszichologia.pte.hu

- *Történelmi regény korpusz:* nemzetek közötti konfliktusokról szóló 6 történelmi regény teljes szövege (kb 700000 szó);
- *Történelem tankönyv korpusz:* általános és középiskolai tankönyvek részletei a 10 legfontosabb eseményről (kb. 210000 szó);
- *Néphistóriai korpusz:* Félig strukturált interjúk a legpozitívabb/legnegatívabb magyar és európai történelmi eseményekről 500 fős rétegzett mintán (kb. 120000 szó).

A két utolsó korpuszt az MTA Nyelvtudományi Intézetének korábban már átadtuk; ezek annotálása számos vonatkozásban már megtörtént. Ezek szolgálnak alapul a szemantikus szerepek vizsgálatára szolgáló fejlesztésekhez.

3 A pszicho-szemantikai szerepek vizsgálatának problematikája

Adott tehát egy elméleti paradigma (a tudományos narratív pszichológia), egy kutatási módszer (a narratív pszichológiai tartalomelemzés), egy magyar nyelvű szövegkorpusz (a történelemszövegek). Amit első körben keresünk, az az, hogy terjedelmes szövegkorpuszokból automatikusan olyan konkordanciákat hozzunk létre, melyek kilistázzák, hogy ki cselekszik, ki érez, ki gondol és ki értékel.

A történelem szövegek vizsgálatakor a 'ki' nem csupán személy lehet – például 'Mátyás király' – hanem csoport is – például 'a tatár hadak'. A narratív szociálpszichológia számára nem az fő kritérium, hogy személy vagy csoport cselekszik-e/érez-e, stb., hanem az, hogy saját csoport (ingroup) vagy külső csoport (outgroup) teszi-e ezt. Ezért az egyes személyek is ingroup, illetve outgroup címkét kapnak.

Kutatásunk egyik konceptuális nehézsége, ami majd a finomabb vizsgálatoknál fog szerepet játszani, hogy az 'ingroup' és az 'outgroup' kategória egyaránt több alhalmazból tevődik össze, például attól függően, mennyire aprólékos, illetve milyen kognitív doménra fókuszál a szövegíró megközelítése. Többnyire egységes ingroupként jelennek meg 'a magyarok' vagy 'az Árpád-házi királyok' olyan szövegrészletekben, amikor „külügyekről", vagyis a szervezetileg azonos nagyságrendű outgroup szereplőkkel való interakcióról olvasunk (pl. 'besenyők', 'jászok', 'keleti lovas népek', 'kun törzsek', 'Európa hatalmasságai', 'német császár', 'orosz fejedelemség' 'úzok', stb). Természetesen a „belügyekről" szóló tudósításkor a magyar színtér szereplői finomodnak ('a főurak', 'Béla király', 'a trónörökös István herceg' stb.), és ezek kétpólusú csoportba sorolásának automatizálása komoly szakmai kihívást jelent úgy a pszichológus, mint a nyelvtechnológus szakember számára.

Részben ez az oka annak, hogy egy egyetemes vagy kutatási igény szerint különböző történelmi korokra lebontott, robusztus 'Ingroup-Outgroup' szólista elkészítése korántsem triviális, hiába tűnik úgy, hogy pl. a 'Tatár Outgroup' szótár nagyjából véges számú elemből és azok variálódásából áll össze ('Dzsingisz kán', 'Batu kán', 'nagykán', 'mongolok', 'mongol törzs', 'nomád sereg', 'tatár hadak', tatár hordák', 'tatárok', stb.). A kézzel összeállított szólisták a variabilitás miatt nagy időbefektetés árán tudják csak a releváns entitásokat lefedni a korpuszokban (pláne az

egyelőre feldolgozatlanokban) előfordulókból. Továbbá, leggyakrabban csakis szövegkörnyezet vagy egyéb, nem objektív kritérium/megegyezés alapján lehet eldönteni, hogy egy entitás melyik csoportba tartozik. Ezért fejlesztésünk során a digitális bölcsészeti kutatásokra jellemző félautomatikus módszerrel dolgozunk, ami a kutatási és implementációs folyamat fontos részeként az automatikus feldolgozás részeredményei után, meghatározott fázisokban, lokális vizsgálattal elvégzett kézi egyértelműsítést és javítást foglal magába.

A továbbiakban a probléma nyelvtechnológiai megközelítéséről és modellezéséről számolunk be. Módszerünk a MetaMorpho nyelvi elemzés morfoszintaktikai és szemantikai kimenetét kapcsolja össze a NooJ eszköz procedúráival. Nyelvészeti szempontból az egyik legnehezebb feladat a szövegekben a koreferencia (illetve az anafora) feloldása, mivel az egyes szereplőket több kifejezés is jelezheti (például: 'IV. Béla leánya' = 'a király leánya' = 'Árpádházi Szent Margit'.) További aspektusa az entitások felismerésének a metonimikus használat, vagyis, hogy önmagukban élettelen dolgok is ágensként, aktív szereplőként említődhetnek – például 'az egri vár hősiesen ellenállt', stb. Ezekkel a jelenségekkel jelenleg csak marginálisan foglalkozunk, mivel többszörösen összetett technológiai megoldást igényelnek.

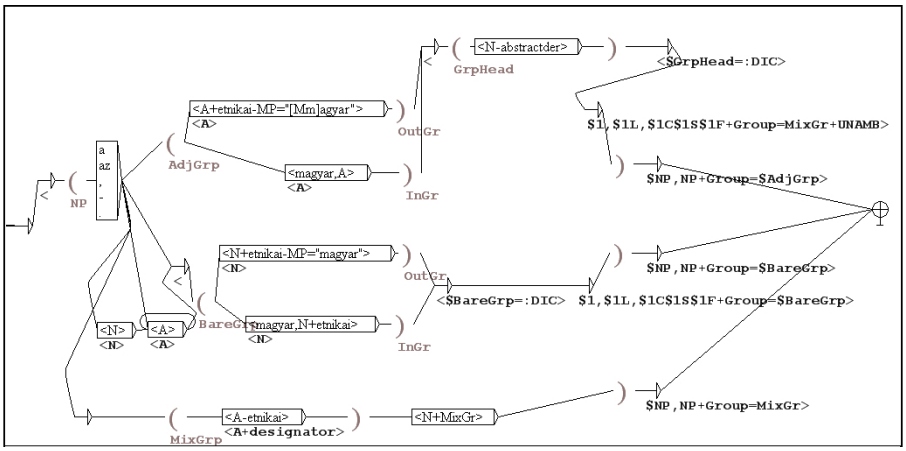
4 Pszichoszemantikai szerepek annotációja

Az automatizált annotációs folyamat kiindulópontja a 'Mixgroup' kategória, vagyis az olyan lexikális elemek, amik az Ingroup vagy Outgroup kategóriához egyaránt tartozhatnak, ilyen például a 'sereg', 'lakosság', 'ország', 'lovasság', stb. A szövegekben ezek megjelenhetnek önmagukban, illetve pl. névelővel és/vagy jelzőkkel együtt, egy NP fejeként. Korábbi munka során elkészült az etnikai főnevek/melléknevek szótára, amit a NooJ fejlesztői környezetben a Mixgroup entitások modellezésében felhasználunk. Létrehoztunk egy NooJ egyértelműsítő prototípus-gráfot (l. 1. ábra), ami

- eldönti, hogy mikor áll főnévi és mikor jelzői szerepben egy etnikai entitás (pl. 'a törökök', ill. 'a török szultán'), és ezeket az NP-ket InGr, illetve OutGr szemantikai címkével látja el;
- begyűjti az összes olyan fejet, ami etnikai jelzővel áll és N+MixGr-ként címkézi őket;
- a MixGr címkéjű főneveket módosító, de eddig az etnikai szólistában nem szereplő jelzőket egy speciális osztályba sorolja, ami azt fogja jelezni, hogy utána egy potenciálisan InGr vagy OutGr elem következhet.

Így egy kb. 600 szóból álló NP halmazhoz jutunk csak a "tatár korpuszon", amit a NooJ-ban félautomatikusan szótárrá alakítunk: az NP-k lemmájukkal és morfológiai jegyeikkel együtt egy külön szótárban eltároljuk. A szótárat a következő elemzési fázisokban, illetve új korpuszok elemzésekor használjuk fel. A fent leírt eljárás azért fontos, mert a szintaktikai szövegelemzők gyakran csak az NP-fejet írják ki; mi olyan gyakorlatias megközelítést választottunk, ami feltételezi, hogy a magasszintű (szemantikai) elemzés során egy már meglévő előfeldolgozó eszköz kimenetére

támaszkodunk, amibe nincs lehetőségünk belenyúlni (vagyis ‘black box’-ként érhető el).



1. ábra: NooJ szintaktikai-szemantikai egyértelműsítő prototípusgráf.

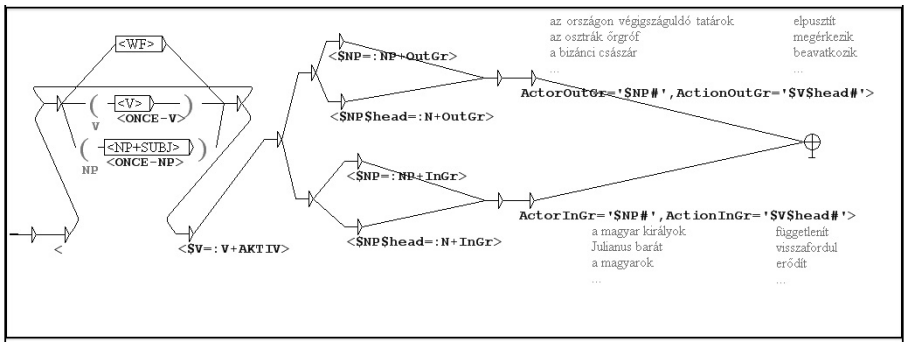
Az entitásdetektáló kör további lépései:

- az egyértelműsítő/címkéző gráfnak a nagyobb lefedés érdekében történő kiterjesztése;
- egyéb mechanizmusok beépítése, pl. hogy megtaláljunk olyan képzett szavakat, amelyeket a NooJban morfológiai lekérdezés alapján az aktivitás stb. szótárba felvehetünk: ‘a mongolok előretöréséről’ > ’előretör’;
- a szemantikailag is anaforikus MixGr kifejezések kézi egyértelműsítése NooJ konkordancia alapján, ahol szövegösszefüggésből lehet eldönteni, hogy az adott csoport saját csoportot vagy külső csoportot jelöl: pl. ’Batu kán visszavonta a *katonáit*’. Sokszor mondathatáron túl átnyúló anaforáról van szó (pl. ’Az *uralkodó* nehéz helyzetbe került.’), ekkor a NooJ kontextusablakának a méretét nagyobb szószámra lehet állítani. A NooJ-ban az illesztett kategóriákhoz tartozó lexikai elemekhez egyben az annotációk is hozzárendelhetők.

5 Thematikus szerep hozzárendelése

Nyelvtechnológiai megközelítésben a thematikusszerep-felismerés (‘semantic role labeling’) kínálkozik alkalmas megoldásként az ágencia detektálására, viszont magyar nyelvre egyelőre nincs létrehozva thematikusszerep-felismerésben felhasználható strukturált szemantikai erőforrás vagy annotált korpusz [lásd pl. 11]. A MetaMorpho [15] képes bizonyos igékhez tartozó thematikus szerepeket felismerni, amelyet szabályalapú koreferenciafeloldással is támogat [12], habár csak viszonylag kevés számú ige mellett. A Vincze Orsolya és Gábor Kata által megalkotott NooJ protézisgráf a MetaMorpho által thematikus szereppel felcímkézett igei bővítményeket találja meg

[19]. Ezen túlmutatva, jelen munkánk célja, hogy a már meglévő erőforrásokhoz igazodva úgy ismerjük fel az ágenciát, hogy ehhez felhasználjuk a történelmi szövegek lexikális elemeihez gráfokkal automatikusan hozzárendelt pszichoszemantikai kategóriákat. A MetaMorpho által felismert főnévi csoportokat és azok mondatban betöltött szerepét kódoló XML-fájlt importáljuk a NooJ-ba, mely után a szótárakkal, illetve szintaktikai mintaillesztésével az In/Outgroup entitások Ágens szerepét igyekszünk meghatározni (l. 3. ábra).



2. ábra: Az ágenciát pszichoszemantikus csoportok alapján szűrő NooJ gráf.

1. Az Aktivitás NPTA gráfot a NooJ elemzőfolyamatban kiemelt szintaktikai elemzőként beállítva a "+AKTIV" címkét kapott igékre szorítjuk a keresést. Így automatikusan kiszűrjük a találatok közül az olyan tartalmú mondatokat, ahol az alany nem cselekvő, pl. '*László király a kunok között érezte jól magát.', 'A főurak közül sokan örültek a király bajának'*. Bár megjegyezzük, hogy a történelmi szövegekben előforduló entitások főképp cselekvőként vagy szenvedőként szerepelnek, és a nem aktív igék viszonylag ritkán, illetve nem az általunk vizsgált "etnikai" entitásokkal fordulnak elő, pl. '*A páncélos katonaság mellett nőtt a könnyűlovasság száma is.', 'A mongol sereget nem egészen helyesen, általánosítva nevezzük - előcsapataikról - tatár seregnek'*.
2. A fent elkészített Group szótár alapján lehetővé tesszük a keresést In/Outgroup entitásokra lebontva, anélkül, hogy ezek lexikai alakjait a gráfba kódolnánk, illetve a szótár új korpuszokon történő iteratív bővítése alapján egy növekvő főnévi lemma- és NP lista, és a szövegek kézzel egyértelműsített annotálása alapján.

6 A fejlesztés további lépései és alkalmazási lehetőségei

Soron következő lépésünk egyrészt a szereplői szótárak és a narratív pszichológiai tartalomelemzési modulok (jelen esetben főként az érzelem, a kogníció és az értékelés) összeillesztése és továbbfejlesztése lesz, másrészt az igei argumentumok thematikus szerepének meghatározása a NooJ-ban írt lokális grammatikák alapján, például:

ha V+AKTIV és Group(SUBJ) == OutGr,
akkor Th_role(OBJ) = Undergoer és Group(OBJ) = InGr.

Ezzel a terjedelmes szövegkorpuszokból automatikusan olyan konkordanciák hozhatók létre, melyek nem csupán azt listázzák, hogy ki cselekszik, ki érez, ki gondol és ki értékel, hanem azt is, hogy ki mindezeknek a tárgya vagy elszenvedője/kedvezményezettje (agent vs. patient/undergoer). Ezek statisztikai feldolgozása révén vonhatók le az egyéni és a csoportidentitással kapcsolatos narratív pszichológiai következtetések.

Hivatkozások

1. Bigazzi S., Csertő I., Nencini, A.: A személy- és csoportközi értékelés pszicholingvisztikája. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2006) 267–276
2. Csertő I.: A személy- és csoportközi értékelés pszichológiai szempontú elemzése elbeszélő szövegekben. In: Alexin Z., Csendes D. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2009) 272–284
3. Ehmann, B., Balázs, L., Fülöp, É., Hargitai, R., Kabai, P., Péley, B., Pólya, T., Vargha, A., Vincze, O., László, J.: Narrative Psychological Content Analysis as a Tool for Psychological Status Monitoring of Crews in Isolated, Confined and Extreme Settings. Acta Astronautica, Vol. 68, No. 9-1) (2011) 1560–1566
4. Ehmann, B., Garami, V.: Narrative Psychological Content Analysis with NooJ: Linguistic markers of time experience in Self reports. In: Proceedings of the 2008 International NooJ Conference. Cambridge Scholar Publishing (2010) 180–190
5. Ehmann, B., Garami, V., Naszódi, M., Kis, B., László, J.: Subjective Time Experience: Identifying Psychological Correlates by Narrative Psychological Content Analysis. Empirical Text and Cultural Research Vol. 3 (2007) 14–25
6. Ferenczhalmy R., László J.: Az intencionalitás modul kidolgozása NooJ tartalomelemző programmal. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2006) 285–295
7. 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 Informatikai Tanszékcsoport (2006) 296–304
8. 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 Cultural Research Vol. 3 (2007) 26–38
9. László, J.: The Science of Stories: An introduction to Narrative Psychology. Routledge, London, New York (2008)
10. László, J., Ehmann, B., Péley, B., Pólya, T.: Narrative psychology and narrative psychological content analysis. In: László, J., Stainton Rogers, W. (eds.): Narrative Approaches in Social Psychology. New Mandate, Budapest (2002) 9–25
11. Márquez, L., Carreras, X., Litkowsky, K. C., Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue. Computational Linguistics Vol. 34, No. 2 (2008) 145-159
12. Miháltz, M.: Knowledge-based Coreference Resolution for Hungarian. In: Proceedings of The Sixth International Conference on Language Resources and Evaluation. Marrakesh, Morocco (2008)

13. Pólya, T., Kis, B., Naszódi, M., László, L.: Narrative perspective and the emotion regulation of a narrating person. Empirical Text and Cultural Research Vol. 3 (2007) 50–61
14. Pólya, T., László, J. and Forgas, J. P.: Making sense of life stories: The role of narrative perspective in communicating hidden information about social identity. European Journal of Social Psychology Vol. 35, No. 6 (2005) 785–796
15. Prószéky, G., Tihanyi, L.: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference. ASLIB, London, United Kingdom (2002) 19–24
16. Silberztein, M.: NooJ Manual (2003) Elérhetőség: www.nooj4nlp.net
17. Szalai, K., László, J.: Activity as a Linguistic Marker of Agency: Measuring in-Groupversus Out-group Activity in Hungarian Historical Narratives. Empirical Text and Culture Research RAM-Verlag: 4 (2010) 50–58
18. 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
19. Vincze O., Gábor K., Ehmann B., László J.: Technológiai fejlesztések a NooJ pszichológiai alkalmazásában. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Egyetemi Kiadó (2009) 285–294
20. Vincze, O., László, J.: Narrative Means of Intergroup Relations: Cognitive States and their role in reducing or increasing intergroup conflict. In: General Meeting of the European Association of Social Psychology (EASP). Stockholm (2011) 30
21. Vincze, O., Tóth, J., László, J.: Representations of the Austro-Hungarian Monarchy in the history books of the two nations. Empirical Text and Cultural Research Vol. 3 (2007) 62–71

Paralingvisztikai jegyek a narratív pszichológiai tartalomelemzésben: a magabiztosság-krízis skála

Puskás László

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

Tanulmányunkban egy újfajta narratív pszichológiai eljárásmóddal lefolytatott vizsgálat kezdeti eredményeit ismertetjük. Arra teszünk kísérletet, hogy a narratív pszichológiai tartalomelemzést és a vokális mintázatok pszichológiai „tartalomelemzését" összekapcsoljuk, vagyis a történetszerkesztésből és a beszéd jellemzőiből az elbeszélő belső állapotaira vonatkozó következtetéseinket egységes keretbe foglaljuk. A lelkiállapot-változás, illetve a krízishelyzet nyelvi tartalmi és fonetikai jegyeit párhuzamosan vizsgáltuk. Megállapítottuk, hogy a narratív tartalmi jegyek struktúrája és a fonetikai struktúra együttesen jelzik a lelkiállapot-változások intenzitását, minőségét. A magabiztosság-dominancia és a krízishelyzet jelzésére, valamint ezek mértékének meghatározására létrehoztunk egy összetett jelzőszámot, amelynek értéke a nyelvi markereket és a vokális jelzéseket egyaránt figyelembe veszi. Ezt a jelzőszámot magabiztosság-krízis indexnek neveztük el. Az index kiszámításánál a nyelvi markerek relatív előfordulási gyakoriságából képzett arányszámokat összegezzük, melyekhez hozzáadjuk a vokális paraméterekre kiszámolt mérőszámokat. A kapott eredmény tartalmaz egy negatív előjelű korrekciós mérőszámot is. Az index értéke egy olyan skálán mozog, amely alapján következtethetünk a közlő kiegyensúlyozottságára, illetve krízishelyzetére.

1 Bevezetés

A narratív pszichológia szerint az elbeszélésben jut kifejezésre az a mód, ahogy az emberek élményeiket, a társas világhoz való viszonyukat megszervezik, identitásukat megalkotják. Ha elfogadjuk azt, hogy az emberek a történetekben és azok révén konstruálják meg önmagukat, és saját pszichológiai valóságukat, el kell fogadnunk azt is, hogy e történetek élményanyaga információval szolgálhat a történetmondó ember alkalmazkodására és megküzdési stratégiáira vonatkozóan is. A narratológia az elbeszélések véges számú alkotóelemét és a véges számú alkotóelemek véges számú variációit írta le, miközben a szöveg végtelenül változatos lehet. A narratív pszichológiai tartalomelemzés ezeket a narratív alkotóelemeket alakítja olyan tartalmi kategóriákká, amelyekhez élményszintű pszichológiai jelentések társíthatók, tartalmakat keres a szövegben, amelyek valamilyen pszichológiai folyamatnak megfeleltethetők. Az alkotóelemek, illetve ezek változatai a szövegben megbízhatóan azonosíthatók, és az elbeszélés így meghatározott elemeihez élményszintű pszichológiai jelentések tár-

síthatók [2,4]. A narratív pszichológiai kutatások eddig figyelmen kívül hagyták az elhangzott közlés fonetikai paramétereit, mint a vizsgálatok eredményeit befolyásoló tartalmi elemeket [3].

Egy elhangzott szövegben nemcsak a nyelvi alkotóelemek, hanem a vokális jellemzők is összefüggenek a közlő lelkiállapotával. Ezek a vokális elemek viszonylag jól körülhatárolhatók, és azonosításuk révén többletinformációhoz juthatunk. Az elhangzott szöveg fonetikai struktúrájában olyan törvényszerűségeket találhatunk, melyek a közlő lelkiállapotával összefüggésben jól megragadhatók. Scherer [8] azzal magyarázza ezeknek az állapotoknak, illetve állapotváltozásoknak a vokális mintázatra gyakorolt hatását, hogy a szervezetben lezajló változások olyan fiziológiai állapotváltozást eredményeznek, amelyek hatással vannak a hangképzési és artikulációs izmokra is. Ezek a változások befolyásolják a hangképzést, melynek révén eltérő akusztikai karakterisztikumok jelenhetnek meg. A vokális mintázatokat elsősorban az érzelemkifejezéssel összefüggésben vizsgálták.

2 Lear két monológjának tartalomelemzése korábbi vizsgálatokban

Pennebaker és Ireland [6] elemezték Shakespeare Lear királyának nyelvhasználatát. Tanulmányukban az egyes szám első személyű névmások (szelf-referencia), a többes szám első személyű névmások (mi-referencia), a pozitív és a negatív érzelmekre utaló kifejezések, valamint a nagy szavak előfordulási gyakoriságát vizsgálták. Ezek krízishelyzetbeli nyelvi mintázatba rendeződésével a Narrcat programmal lefolytatott komplex vizsgálat részeként a 3.2 alfejezetben foglalkozunk.

3 A vizsgálat

3.1 A vizsgálati anyag

Tanulmányunk nem pusztán a leírt szöveget, hanem annak színészi megfogalmazásának tulajdonságait igyekszik vizsgálni az elhangzott szöveg fonetikai sajátosságai és a szöveg tartalma alapján. Lear első és utolsó monológjának szövegét és színészi megjelenítését kíséreljük meg összehasonlítani a Pennebaker és Ireland [6] által elemzett szövegrészletek alapján. Azért szükséges hangsúlyozni, hogy ezen szöveg alapján dolgozunk, mert a tanulmány mindkét monológot rövidített formában közli, és az összehasonlíthatóság miatt szükséges a lehető legteljesebb egyezés. A hanganyagot a Magyar Televízió 1978-ban készült Lear király című tévéjátékának felhasználásával vizsgáltuk meg. A vizsgált monológok szövege magyarul a következő:

„…Tudnotok kell, hogy országunkat három részre osztjuk, erős szándékunk minden gondot és bajt lerázni agg korunkról, átadván ifjabb erőknek, míg magunk tehertől menten mászunk a sír felé. Fiúnk Cornwall, s nem kevésbé szeretett fiúnk Alban, ez órában szilárd akaratunk lányaink hozományát külön kiszabni, hogy jövő viszálynak már most elejét vegyük. (…) Szóljatok leányok (minthogy mi le akarunk mondani az

ország gondjairól s jövedelmeiről), halljuk hát, melyiktek szeret leginkább? Hogy legfőbb kegyünket érdem szerint oszthassuk…"

„Ti mind kőemberek vagytok. Ha nyelvetek, szemetek enyéim volnának, olyan zivatart zúdítanék, hogy meghasadna a Mennynek boltozatja. Ó, vége, örökre. Én tudom, ki holt meg, és ki él. Ő holt, akár a Föld. Dögvész irtson ki gyilkos árulók! Én megmenthettem vol', s vége, vége! Cordelia, Cordelia! Várj egy kicsit! Mit mondasz? Mindig nyájas volt szava, szelíd és halk, nőben nemes vonás. (…) Ki vagy te? Szemem nem jó, de megmondom, meg én. (…) Gomboljatok ki, kérlek. – Köszönöm. [Ez utóbbi két mondat a filmbeli átiratból hiányzik.] Nézzétek! Látjátok ezt? (…)"

3.2 Módszer és eredmények

3.2.1 A szöveg strukturális-tartalmi elemeinek vizsgálata

A Narrcat programmal lefolytatott vizsgálat eredményeit az 1. táblázat mutatja.

1. táblázat: Lear első és utolsó monológjának tartalmi elemei.

	Tagadás	**Szelf-referencia**	**Mi-referencia**	**Értékelés**	**Aktív/ passzív**	**Kognitív**	**Intenció**	**Idő: befejezés**	**Érzelem**	**Idő: örök idő**
Első monológ	1	0	16	2 pozitív	2	3	3	0	2 pozitív	0
Utolsó monológ	1	9	0	4 pozitív 3 negatív	3	3	0	2	0	2

Fentiek alapján, a Lear monológokat felhasználva, felállíthatjuk a lélektani krízis nyelvi jegyeinek profilját (lásd 2. táblázat). Ehhez a szövegszintű mintázathoz szorosan kapcsolódnak a fonetikai paraméterekben bekövetkező változások.

2. táblázat. A krízis nyelvi markereinek mintázata.

	Tagadás	**Szelf-referecia**	**Mi-referencia**	**Értékelés**	**Aktív/ passzív**	**Kognitív**	**Intenció**	**Idő**	**Érzelem**
Változás iránya	Nő/ stagnál	Nő	Csökken	Negatív, nő	Passzív nő/ stagnál	Stagnál/ csökken	Csökken	Idői távolodás jegyek nőnek	Pozitív csökken, negatív nő

A szöveg nyelvi tartalmi elemei mellett megvizsgáltuk a fonetikai struktúrát is. A két, egymással nem megfeleltethető, de egymás hatását erősítő struktúra együttes mérésére pedig bevezettük a magabiztosság-krízis indexet.

3.3 A fonetikai paraméterek alakulása a beszédben

3.3.1 A fonetikai paraméterek vizsgálata

A kiválasztott két monológ vizsgálatához a Praat [7] fonetikai programot használtuk fel, melyet az Amszterdami Egyetem munkatársai fejlesztettek ki. Az érzelmi állapotok fonetikai paraméterekre gyakorolt feltételezett hatásával részben Scherer [8] tanulmánya alapján foglalkoztunk, amely harminckilenc korábbi tanulmány adatait összegezte. Az előfeltevéseinket a 3. táblázat tartalmazza.

3. 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.

Összefoglalóan azt mondhatjuk, hogy lelkiállapot-változás esetén az artikulációs tempó várhatóan csökken az élvezet/boldogság és az őrjöngés/forró düh esetén, míg nő a jókedv/öröm, a szomorúság/levertség, a bánat/kétségbeesés, valamint a félelem rettegés során (ennél fokozottan), a többi lelkiállapot-változás, az elvárások szerint, nem gyakorol hatást rá, illetve ezek hatása előre nem kiszámítható. A hangerő változása, várakozásaink szerint, mind a tizenkét felsorolt lelkiállapot-változásra hatással

van. Nyolc esetben egyértelműen nő a hangerő. Ezek közül háromban fokozottan nő. Egy esetben csökken a hangerő, míg három esetben nem a változás iránya, hanem maga a változás a meghatározó. A hangerő-intervallumoknál négy esetben növekedést, egy esetben csökkenést, egy esetben csökkenést vagy változatlan hangerőt várunk. A beszédszakaszok és a szünetek hosszánál rövidülést várunk négy lelkiállapot-változásnál. Mindkét változó esetén a jókedv/öröm, a bánat/kétségbeesés, a félelem/rettegés, valamint az őrjöngés/forró düh esetén áll be a csökkenés. A beszédszakaszok hosszát, az artikulációs tempó, a hangerő és a hangerő-intervallumok mellett, a magabiztosság-krízis skálázására létrehozott index kialakításánál is felhasználtuk.

3.3.2 A magabiztosság-dominancia és a krízishelyzet pszichológiai skálázásának lehetőségei, a magabiztosság-krízis index

A magabiztosság-dominancia jegyeinek mintázatba rendeződését Lear krízishelyzet előtti megnyilatkozásában a tartalmi elemek vizsgálatánál és a szöveg fonetikai elemzésénél egyaránt megtaláltuk. Ugyanez igaz a krízishelyzetet követő megnyilatkozás mintázatba rendeződésére is. A tartalomelemzés és a vokális jegyek mintázatának vizsgálata nem feleltethető meg egymásnak közvetlenül, még ha kétségkívül egymás hatásait erősítik is, és a megnyilatkozó lelkiállapotának intenzitásáról tudósítanak. A vizsgált jegyek mintázatba rendeződését vizsgálva, igyekeztünk olyan összetett skálázási módszert kialakítani, mellyel a krízishelyzet jellemezhető.

Úgy gondoljuk, hogy nemcsak arról van szó, hogy a vokális jelzések mérésével is leírhatjuk ugyanazt a lelkiállapotot, sokkal inkább arról, hogy a vokális paraméterek és a nyelvi markerek együttesen jelzik a megnyilatkozó lelkiállapotát, és ennek a lelkiállapotnak az intenzitását, amit az is valószínűsít, hogy a vokális paraméterek és a nyelvi markerek nem feleltethetők meg közvetlenül egymásnak. Ebből adódik, hogy eljárásunk két összetevőre oszlott: egyrészt a vizsgált szöveg tartalomelemzésére, másrészt az elhangzott szöveg akusztikai paramétereinek vizsgálatára. Két egymástól teljesen különböző eljárást folytattunk le párhuzamosan, melyekben a vizsgálati egységeink is eltértek egymástól. A szöveg tartalomelemzésénél az elemzési egységünk a szó volt, és a keresett szavak relatív előfordulási gyakoriságát vizsgáltuk. A fonetikai vizsgálatnál a beszédszakaszokat tekintettük elemzési egységnek, amelyek nem feltétlenül feleltethetők meg minden esetben teljes értékű mellékmondatoknak, nyelvtani értelemben. Az akusztikai vizsgálatnál a kiugró értékek gyakoriságát és intenzitását vizsgáltuk.

A magabiztosság-dominancia és a krízishelyzet jelzésére, valamint ezek mértékének meghatározására létrehoztunk egy összetett jelzőszámot, amelynek értéke a nyelvi markereket és a vokális jelzéseket egyaránt figyelembe veszi. Ezt a jelzőszámot magabiztosság-krízis indexnek neveztük el. Az index kiszámításánál a nyelvi markerek relatív előfordulási gyakoriságából képzett arányszámokat összegezzük, melyekhez hozzáadjuk a vokális paraméterekre kiszámolt mérőszámokat. A kapott eredmény tartalmaz egy negatív előjelű korrekciós mérőszámot is. Minél alacsonyabb az index értéke, annál kiegyensúlyozottabb, magabiztosabb a kísérleti személy (a nullához közeli, illetve a negatív érték egyértelműen a dominancia és a magabiztosság jele). Minél magasabb értéket kapunk az indexre, annál erőteljesebb krízishelyzetre utal a

megnyilatkozás. Az index kiszámításához hat arányszámot használtunk fel, melyek értékét egymással összeadtuk:

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.

A magabiztosság-krízis indexbe be kívántuk foglalni az intencióra, az aktivitásra, a kognitív folyamatokra, az értékelésre és az érzelmekre vonatkozó eredményeket is, azonban a két monológban ezek olyan kis gyakorisággal fordultak elő, hogy statisztikailag nem voltak kezelhetők.

3.3.3 A magabiztosság-krízis index segítségével nyert eredmények

Eredményeink egyértelműen azt mutatják, hogy Lear első monológját a kiegyensúlyozottság, a magabiztosság és a dominancia uralja, míg utolsó monológját a súlyos krízishelyzet jellemzi (4. táblázat).

4. táblázat: A magabiztosság-krízis index kiszámítása a hat felhasznált mérőszám alapján.

Mérőszámok	1	2	3	4	5	6	Összesen
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 táblázatból az is kitűnik, hogy az indexhez használt vokális és az írott szövegben mért paraméterek külön-külön eltérő összesített mérőszámokat adnának, és együttesen határozzák meg a krízishelyzet és a kiegyensúlyozottság mértékét.

Szükséges magyarázatot fűznünk a magabiztosság-krízis indexhez felhasznált paraméterekhez és azok kiszámítási módjához. A fonetikai paraméterek vizsgálatánál azt az elvet követtük, hogy a kiválasztott beszédszakaszok számát a vizsgált szöveg szószámával osztottuk el. Erre azért volt szükség, mert a beszédszakaszok több szó-

ból is állhatnak, és ha az egész beszédszakaszt kiválasztanánk, akkor ezzel valamennyi szót kiemelnénk, ami aránytalanságokhoz vezetne, ezért úgy tekintettük, mintha a beszédszakasznak egyetlen szava kerülne megjelölésre, és így a megjelölt szavak számát osztanánk el a teljes szószámmal. A másik kérdés az volt, hogy ha egy beszédszakasz több általunk vizsgált fonetikai paraméternek is megfelel, akkor hányszor vegyük figyelembe. Amellett döntöttünk, hogy valamennyi fonetikai paraméternél külön számítjuk be, mintha annyi megjelölt szó lenne az adott beszédszakaszban, ahány az általunk vizsgált fonetikai paraméternek megfelel, ha úgy tetszik, ezzel súlyoztuk az index fonetikai mérőszámainak összetevőit. Ezt azért tartottuk fontosnak, mert úgy gondoljuk, minél több kiugró értéket tartalmaz egy beszédszakasz, annál intenzívebb a megnyilatkozó lelkiállapota.

A kettő másodperc alatti beszédszakaszok relatív gyakoriságát azért használtuk fel az index kialakításánál, mert úgy véljük, hogy a beszédszakaszok hosszából következtethetünk a beszélő gondolatainak összeszedettségére, az illető fájdalmára, és arra, hogy az adott helyzetre milyen korábban konstruált sémával rendelkezik. Természetesen a kiegyensúlyozott megnyilatkozásban is lehetnek és vannak rövidebb beszédszakaszok, megszólítások, csodálkozások, de a krízishelyzetben, feltételezésünk szerint, jóval nagyobb lehet a relatív előfordulási gyakoriságuk, mivel a válaszreakció, a helyzet újdonságértékéből adódóan, kevésbé automatikus.

A hangerőcsúcsokat tartalmazó beszédszakaszok fontos szerepet töltöttek be az index kialakításánál, hiszen, ahogy azt a 3. táblázatban már korábban ismertettük, bánat/kétségbeesés, szorongás/aggodalom és szégyen/bűntudat esetén növekszik a hangerő, félelem/rettegés, ingerültség/hideg düh és őrjöngés/forró düh esetén pedig fokozottan növekszik a hangerő.

Az alacsony hangerő-intervallumok gyakorisága, feltételezésünk szerint, egyfajta olyan monotonitást kölcsönöz a megnyilatkozásnak, amely az erő és a magabiztosság hiányára utal, rossz lelkiállapotra.

A szelf-referencia és a tagadás előfordulási gyakoriságát vizsgálta Pennebaker és Ireland [6], valamint László és munkatársai [4] is, akik ezek relatív gyakoriságát nézték meg a szövegben. Az énre való túlzott utalás a befelé fordulás jele, míg a 'mi'-re történő utalás a mások irányába való nyitást fejezi ki. Patológiás esetben a magas énreferencia összefüggést mutat a depresszióval, a szuicid tendenciákkal. A tagadást pszichodinamikai szempontból az egészséges emberi környezethez és morális mércékhez való alkalmazkodásra, illetve a világ értéktelenítésére, a destrukcióra és öndestrukcióra való hajlamra vonatkozóan vizsgálták [1]. Krízishelyzetben a megváltozott környezethez való alkalmazkodás problémás, fokozottan fordulhat elő tagadás az elbeszélésben.

A mi-referenciát a magabiztosság-krízis indexnél negatív korrekciós mérőszámként használtunk fel. Erre egyrészt azért volt szükség, mert az indexet alkotó összetevők úgy állnak össze egésszé, hogy minél nagyobb az index értéke, annál erőteljesebb a krízis, és a mi-referencia értéke pont a kiegyensúlyozott megnyilatkozásoknál a legmagasabb, így ott ellentétes hatást érne el. Másrészt a magabiztos megnyilatkozásnál ennek a változónak a negatív értéke jelentősen csökkenti a „véletlenszerűen", a megnyilatkozásba került, általunk vizsgált paraméterek relatív előfordulási gyakoriságának értékét, viszont az erőteljes krízishelyzeteknél kapott indexet kevésbé vagy egyáltalán nem befolyásolja.

Összességében elmondhatjuk, hogy ha csak a megnyilatkozáshoz tartozó magabiztosság-krízis indexet ismerjük, jó eséllyel következtethetünk a beszélő lelkiállapotára is.

4 Megvitatás

Összefoglalóan azt mondhatjuk, hogy az élőszóban is elhangzó megnyilatkozásoknál, a szöveg tartalmi elemein túl, célszerű a fonetikai szerkezet vizsgálata, amely sok esetben árnyalhatja, kiegészítheti, illetve pontosíthatja a hagyományos tartalomelemzés módszereit. Lear két monológjában azt a krízis okozta lelkiállapot-változást vizsgáltuk, amelyet veszteségtörténetként jellemezhetünk.

A tudományos narratív pszichológiai megközelítés az elbeszélések pszichológiai jelentéseit már nemcsak a szavak és témák szintjén vizsgálja, hanem a narratívum szintjén is. Az olyan narratív minőségek mentén törekszik a pszichológiai jelentések vizsgálatára, mint amilyen a struktúra, a szervezettség, a perspektíva, az időviszonyok és a koherencia [5]. Ezzel a vizsgált történetek nyelv feletti tartalmait is vizsgálják.

Tanulmányunkban egy új narratív pszichológiai eljárás meghonosítására teszünk kísérletet, mely összekapcsolja a tudományos narratív pszichológiai tartalomelemzésnek a narratív tartalmakra irányuló megközelítését az elhangzott szöveg fonetikai struktúrájának elemzésével. Vizsgálatunk alapján megállapíthatjuk, hogy a szöveg tartalmi elemei és a fonetikai paraméterek egymással nem megfeleltethető, még ha össze is függő, párhuzamos struktúrát alkotnak, így azok együttes vizsgálatát indokolják. E két párhuzamos struktúra együttes vizsgálata az eredmények minőségi javulását, árnyalását és pontosítását is lehetővé teszi. A verbális és non-verbális kód elemzését a magabiztosság-krízis indexszel kapcsoltuk össze, mely mindkét struktúra elemeit felhasználja.

Vizsgálatunk arról tesz tanúbizonyságot, hogy az akusztikai paraméterek összekapcsolása a lelkiállapot-változásokkal eredményesen alkalmazható technika. Megállapíthatjuk, hogy krízis hatására a megnyilatkozó lelkiállapot-változása mind a megnyilatkozás tartalmi elemeiben, mind pedig annak fonetikai struktúrájában kimutatható, és adatokkal alátámasztható. Meggyőződésünk, hogy a színészi játék modellálta helyzet vizsgálata a spontán megnyilatkozásoknál is alkalmazható, és, a szöveg tartalmi elemeinek vizsgálatával párhuzamosan, alapja lehet egy, a fonetikai struktúrát is vizsgáló, összetett tudományos narratív pszichológiai eljárás alkalmazásának.

Irodalom

1. 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 No. 2 (2005) 181–199
2. László J.: Előszó. In: László J., Thomka B. (szerk.): Narratív pszichológia. Narratívák 5. Kijárat Kiadó, Budapest (2001) 7–15
3. László, J.: Narratív pszichológia. Pszichológia Vol. 28, No. 4 (2008) 301–317

4. László, J.: The science of stories.: An introduction to narrative psychology. Routledge, London; New York (2008)
5. László, J., Ehmann, B., Péley, B., Pólya, T.: A narratív pszichológiai tartalomelemzés: elméleti alapvetés és első eredmények. Pszichológia Vol. 20, No. 4 (2000) 367–390
6. 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
7. Praat: http://www.fon.hum.uva.nl/praat/
8. Scherer, K. R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin Vol. 99 (1986) 143–165. Magyarul: Vokális érzelemkifejezés. Áttekintés és egy modell az eljövendő kutatásokhoz. Fordította: Bodor Péter. In: Barkóczi Ilona – Séra László (szerk.): Érzelmek és érzelemelméletek. Tankönyvkiadó, Budapest (1989)

A multimodális pragmatikai annotáció jelentősége a számítógépes nyelvészetben[1]

Bódog Alexa[1], Abuczki Ágnes[1], Németh T. Enikő[2]

[1]Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
Egyetem tér 1.
4032 Debrecen
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
Egyetem utca 2.
6722 Szeged
{alexa.weirdling, abuczki.agnes}@gmail.com,
nemethen@hung.u-szeged.hu

Kivonat: Jelen tanulmány egy olyan pragmatikai annotációs eljárást mutat be annak szintjeivel, technikai eszközeivel és kezdeti eredményeivel együtt, amely segítségével lehetővé válik a társalgás bizonyos mozzanatainak automatikus fölismerése és kinyerése, valamint a társalgás szerkezetével és menetével kapcsolatos predikciók megtétele. Az annotációs eljárást a multimodális, spontán hétköznapi társalgásokat tartalmazó magyar HuComTech-korpuszon fejlesztettük ki. Az annotációs rendszer nyelvfüggetlen, univerzális kategóriákkal dolgozik, típusos szerkezetű, az egyes szintek egymásra épülnek benne. Az annotációs szintek az alábbiak: a kommunikatív aktusok szintje, a támogató aktusok szintje, a tematikus kontroll szintje, valamint az adott-új információ szintje. Az eljárás megfelel a jelenleg is kidolgozás alatt álló nemzetközi standardizációs elvárásoknak, követelményeknek.

1 Bevezetés: pragmatika és számítógépes nyelvészet

A számítógépes nyelvészet területén a pragmatikai kutatások és fejlesztések évről-évre nagyobb teret nyernek. E tendencia mögött elsődlegesen az ember–gép interakció (a továbbiakban HCI – human–computer interaction) sikeresebbé tételének motivációja húzódik meg, másodsorban pedig a nyelvtudomány azon komputációs igénye, melynek célja a grammatikai és a pragmatikai kompetencia formális modelljeinek létrehozása [1], [2]. A HCI-alkalmazások sikerének egyik kulcsa azért keresendő a pragmatikában, mert e terület az emberek között zajló kommunikatív nyelvhasználat mibenlétével foglalkozik [3]. A nyelvet különböző kontextusokban, különböző célok

[1] A jelen tanulmány alapjául szol-gáló kutatásban Bódog Alexát és Abuczki Ágnest *A felsőoktatás minőségének javítása a kutatás-fejlesztés-innováció-oktatás fejlesztésén keresztül a Debreceni Egyetemen* című, TÁMOP-4.2.1/B-09/1/KONV-2010-0007 projektazonosítójú program, Németh T. Enikőt pedig az MTA-DE-PTE-SZTE. Elméleti Nyelvészeti Kutatócsoportja támogatta.

elérésének érdekében használjuk, s ennek a bázisnak tükröződnie kell az ember és az általa használt gép kommunikációjában is. E mozzanat fontosságát jelzi az is, hogy a hétköznapi felhasználóknak a gépekhez fűződő viszonya sajátosan későmodern természetű: egyrészről igényeljük és talán el is várjuk azt, hogy a gépek megkönnyítsék mindennapi életünket (így ebből a szempontból általában pragmatisták vagyunk[2] és a gépeket értéksemleges eszközöknek tekintjük), másrészről viszont úgy gondoljuk, hogy életünket és céljainkat a gépek igenis befolyásolják, sőt, bizonyos esetekben meg is változtatják – például a gépek használatát a legtöbb esetben explicit módon tanulni kell (ebből a szempontból eszközeinket értékterheltnek tekintjük).[3]

A föntiek alapján elmondható tehát, hogy a komputációs pragmatika fő kutatási problémái azon jellemzők föltárása és gépi kezelése köré csoportosulnak, melyek a hétköznapi, valós nyelvhasználatot alapvetően meghatározzák. Ennek megfelelően központi helyet foglalnak el a különböző referenciális elemek visszakeresésével kapcsolatos problémák, a nyelvészeti pragmatikából ismert beszédaktusok automatikus generálásának és interpretálásának nehézségei, a beszédaktusokon túlmenően a teljes diskurzusok szerkezete automatikus generálásának és interpretálásának kérdései, valamint az abdukció [2]. E problémák komputációs pragmatikai megoldásai a HCI több kutatási és alkalmazási területén alkalmazhatók, így például dialógusrendszerekben, racionális döntési rendszerekben, vagy akár spontán beszédfelismerő rendszerekben is [1], [4].

A jelenleginél természetesebb HCI megteremtését célzó projektek között megkülönböztetett fontossággal bírnak a kommunikáció multimodalitását alapul vevő kutatások, melyek során nemcsak a beszélt nyelvi kommunikatív információmanipuláció jellemzőit tárjuk föl, hanem figyelembe vesszük a nem verbális akusztikus, valamint a vizuális tartományból érkező információkat is.

A fönt említett problémák megoldásához többféle adatgyűjtési és elméletalkotási modellt hívhatunk segítségül, melyek közül a legelterjedtebb módszer a korpusz- és adatbázis-építésen alapuló adatkinyerés, illetve elméleti általánosítások megtétele.

Számítógépes nyelvészek és informatikusok számos sémát fejlesztettek ki azon törekvés során, hogy standardizált kódnyelvet és terminológiát hozzanak létre különböző korpuszannotációk számára. Mivel a korpusz- és adatbázis-építés fő kritériuma a standardok követése és az interoperabilitás, ezért ezeket a sémákat általában XML-ben kódolják, amely lehetővé teszi a gépi feldolgozást. A nyelv verbális aspektusainak kódolása mellett szintén standardizált rendszerré fejlődött a nem verbális jegyek kódolása is, melyekre sajátos kódnyelvek születtek, mint például a nemzetközi élvonalbeli, arcizommozdításokat figyelembe vevő Ekman-féle FACS-kódrendszer (Facial Action Coding System).[4] A multimodális kódolósémák közül úttörőként emelkedett ki a MUMIN[5] multimodális kódrendszer a gesztusok és arckifejezések személyközi kommunikációban betöltött szerepének tanulmányozására. A fenti sémákhoz hasonlóan a HuComTech kutatócsoport is egy többszintű, multimodális an-

2 Vigyázat, nem pragmatikusok, csak pragmatisták!

3 A gépekhez fűződő attitűdjeinkről jó áttekintést ad Ropolyi László [5].

4 A FACS manuáljának részlete elérhető az alábbi weboldalon: http://face-and-emotion.com/dataface/facs/manual/TitlePage.html

5 MUMIN: http://www.ling.helsinki.fi/kit/2006k/clt310mmod/MUMIN-coding-scheme-V3.3.pdf

notációs rendszert épített ki, amely figyelembe veszi a kommunikáció verbális akusztikus, nem verbális akusztikus és vizuális jellemzőit is, így különféle multimodális természetű lekérdezésekre és modellépítésre is alkalmas.

Ugyanakkor nem szabad elfelejtkeznünk arról, hogy a korpusz és adatbázis használata egy elméleti döntés, ahogyan az is, hogy milyen annotációt készítünk, szintaktikait, morfológiait vagy pragmatikait, továbbá, hogy a választott típusú annotációs rendszerünk milyen alapegységekkel és szintekkel dolgozik. A HuComTech-korpuszon alkalmazott multimodális pragmatikai annotáció mögött az az elméleti megfontolás húzódik, hogy a kommunikáció során a kommunikációs partnerek egyszerre, szimultán módon veszik figyelembe a különböző elérhető modalitásokból származó stimulusokat. Ezen elméleti döntés értelmében válhatott a multimodális pragmatikai annotáció alapegységévé a kommunikatív aktus.

Jelen tanulmány a kommunikatív aktusok generálására és interpretációjára összpontosít a HuComTech-korpusz vizsgálata és multimodális pragmatikai annotációja alapján. Célunk kettős: egyrészről szeretnénk bemutatni egy olyan, saját fejlesztésű multimodális pragmatikai annotációs rendszert, mely segítségével oly módon tudjuk leírni és értelmezni a személyközi kommunikatív viselkedéseket, hogy az tevékenyen hozzájárulhasson a beszélt ember-gép interakciót lehetővé tévő dialógusrendszerek modellálásához és kivitelezéséhez. Közvetett célunk pedig az, hogy rávilágítsunk arra, hogy a hagyományosan nem formális természetű nyelvészeti pragmatika aktívan képes hozzájárulni a számítógépes nyelvészethez (és viszont), valamint hogy ez a hozzájárulás nem öncélú. Fontos kiemelni azt, hogy kutatásunk e tanulmány elkészítésekor még nem zárult le – az annotáció jelenleg is folyik, így végleges elméleti általánosítások levonására, valamint eredményeink dialógusrendszerbe történő integrálására egyelőre még nem volt módunk. Ennek ellenére annotációs rendszerünk előnyei már most kézzelfoghatók.

Céljainknak megfelelően elsőként röviden bemutatjuk a HuComTech-csoport által épített korpuszt, annotálásunk terepét, majd pedig a QANNOT-annotációs eszközt. Előadásunk legfontosabb részében saját multimodális pragmatikai annotációs rendszerünk szintjeit mutatjuk be példák segítségével, valamint az annotálás eszközét, az annotációs folyamatot és további kutatási terveinket. Zárásként kísérletet teszünk tanulmányunk metaelméleti reflexiójára is, hogy kimutassuk a nyelvészeti pragmatika és a számítógépes nyelvészet egymásra gyakorolt hatását.

2 A HuComTech-korpusz multimodális pragmatikai annotálásának elméleti alapjai

Multimodális pragmatikai annotációs rendszerünk alapjait egy korábbi tanulmányunkban részletesen kifejtettük [6]. Jelen tanulmányban céljainknak megfelelően arra összpontosítunk, hogy rámutassunk a hagyományos nyelvészeti pragmatika és a számítógépes nyelvészet közös metszéspontjaira, így annotációs rendszerünk elméleti alapjait is e nézőpontból mutatjuk be.

A pragmatikai annotáció a társalgás szegmentálását és címkézését jelenti, melynek során nyelvi információt adunk hozzá a nyelvi szegmensekhez, valamint a nem verbá-

lis kommunikatív viselkedést is szegmentáljuk és címkézzük. A pragmatikai annotáció elsősorban a beszélő szándékának megfelelő, és nem csupán a formában (a felszíni szerkezetben) tükröződő kommunikatív funkciókat jelöli meg, hiszen a sikeres kommunikáció feltétele az, hogy a hallgató/címzett ugyanúgy értelmezze a beszélő/feladó megnyilatkozását és szándékait, ahogyan ő (a beszélő) is kívánta [6].

Multimodális pragmatikai annotációs rendszerünk alapját a *kommunikatív aktusok* képezik. A kommunikatív nyelvhasználat e minimális alapegységei nyelvi szempontból megnyilatkozások [7], amelyek társalgási fordulókba, a fordulók szomszédsági párokba, a párok pedig koherens diskurzusokba szerveződnek. A beszélt nyelvi dialógusokat a társalgáselemzésben általában fordulókra szokás szegmentálni, ám mivel ezek a szegmensek nagyon hosszúak is lehetnek, ezért előnyösebb őket további funkcionális egységekre, kommunikatív aktusokra tagolni. A kommunikáció során minden szint sajátos elvek és megszorítások alapján szerveződik. A nyelvészeti pragmatika oldaláról nézve a kommunikatív aktusok multimodális illokúciós aktusok. Illokúciós aktusok, mivel a bennük kifejezett beszélői és szándékolt hallgatói attitűdök alapján szerveződnek, így előtérbe kerülnek a kommunikációban jelen levő intenciók, s multimodálisak, mivel a verbális közlés mellett figyelembe vesszük a vizuális (a gesztusokkal, valamint a különböző arckifejezésekkel támogatott) és a nem verbális akusztikus (prozódiai) információkat is. Az illokúciós aktusok nyelvészeti pragmatikai kutatásai rámutatnak arra, hogy a partikuláris illokúciós aktusok száma igen magas, így ezek vizsgálata parttalanná válhat mind a kategorizáció, mind a csoportosítás tekintetében. Például a kérésnek mint illokúciós aktus fajtának rengeteg „alfaja" különböztethető meg (kérés, parancs, könyörgés, utasítás, kívánalom stb.), s ezek az aktusok ráadásul még nyelvfüggő természetűek is (az egyik nyelvben megvannak, a másikból pedig hiányoznak). Amennyiben magas szinten általánosító modellt kívánunk létrehozni, úgy ki kell küszöbölnünk a nyelvfüggő, partikuláris kategóriákat – túl kell lépni az „egy jelenség = egy szabály" típusú leírásokon. Multimodális pragmatikai annotációs rendszerünkben ezt a problémát úgy oldottuk meg, hogy nem partikuláris aktusokat, hanem aktustípusokat különböztettünk meg egymástól a Bach és Harnish által kidolgozott illokúciós aktustipológia alapján [8]. A típusos megközelítés egyik pozitívuma az tehát, hogy valamilyen szempont alapján (jelen esetben az aktusban kifejezett beszélői és a szándékolt hallgatói attitűdök alapján) osztályokba, típusokba sorolja a példányszintű (token) jelenségeket, így a rendszer alkalmas lesz általános szabályszerűségek megállapítására, s ebből eredően predikciók megtételére.

Rendszerünkben például a kérések, parancsok, kívánalmak stb. egységesen a direktív aktusok típusába tartoznak. A direktív aktusok olyan aktusokat tartalmaznak, melyek propozicionális tartalma a hallgató egy elvárt/preferált jövőbeli cselekedetére vonatkozik, s amelyek kifejezik a beszélő azon szándékát, hogy a hallgató a szóban forgó aktus hatására hajtsa végre a jövőbeli cselekedetet [6]. A direktívek mellett megkülönböztetünk konstatívokat (melyek a beszélőnek egy propozicionális tartalomhoz fűződő hiedelmét fejezik ki úgy, hogy a beszélő mindeközben szándékozza azt is, hogy az aktus propozicionális tartalmát feldolgozza és higgye a hallgató is), komisszívokat (amelyek a beszélő azon szándékát fejezik ki, amellyel elkötelezi magát egy jövőbeli aktus megtételére) és ún. viselkedő aktusokat is (*acknowledgement*, a beszélő valamilyen affektív, érzelmi, attitűdbeli viszonyulását fejezik ki a hallgató felé). A társalgásban előfordulnak olyan esetek is, amikor a megnyilatkozás semmifé-

le propozicionális tartalommal nem rendelkezik, s a megnyilatkozás konkrét illokúciós ereje nem azonosítható.[6] Ebben az esetben a none (nem azonosítható) címkét alkalmazzuk az annotáció során.

A típusos megközelítés másik előnye az univerzalitás: míg a partikuláris aktusok nyelvfüggőek, addig az aktusok típusai nagy valószínűséggel nyelvfüggetlenek [9]. Az univerzális jelenségek mögött meghúzódó szabályszerűségek föltárása a nyelvészeti pragmatikában és a számítógépes nyelvészetben egyaránt fontos: a pragmatika számára azért, mert absztrakt, általános érvényű megállapításokat tudunk tenni a nyelvhasználatra vonatkozóan, a számítógépes nyelvészet számára pedig azért, mert e megállapításokat fölhasználva túl tud lépni a statisztikai alapú alkalmazásokon.

A kommunikatív aktusok mellett az úgynevezett támogató aktusokat is annotáljuk a multimodális pragmatikai annotáció során. Ezek az aktusok nem bírnak önálló illokúciós értékkel, ehelyett kiegészítik, támogatják a velük egy fordulóban szereplő kommunikatív aktust. Ezen aktusok annotálása azért fontos a nyelvészeti pragmatika számára, mert segítségükkel számot tudunk adni egyrészről az "interakcióban levés" mozzanatairól, másrészről a társalgás formai jegyek alapján történő szegmentálásáról. E két mozzanat a számítógépes nyelvészet számára is fontos: az interakcióban való részvételnek pragmatikai funkciójú multimodális jelölői vannak, például a visszajelzés (backchannel), mely történhet bólogatással, hümmögéssel, ühümözéssel stb. Emellett a társalgásban olyan formai jelölők, például diskurzusjelölők és udvariassági markerek is részt vesznek, melyek segítségével könnyen azonosíthatóvá válnak a megnyilatkozásokat alkotó kommunikatív aktusok típusai. Például hiába hangzik el egy kérdő intonációjú megnyilatkozás, ha a végén szerepel a *légy szíves* kifejezés vagy a *kérlek* szócska: tudjuk, hogy a megnyilatkozás ebben az esetben kérés lesz.[7] Multimodális pragmatikai annotációs rendszerünkben a támogató aktusok közül a visszajelzéseket, az udvariassági markereket, valamint a javításokat (melyek során a beszélő a saját partikuláris kommunikatív aktusához fűződő attitűdjét változtatja meg) jelöljük. Távlati terveink között szerepel a diskurzusjelölők annotálása is.

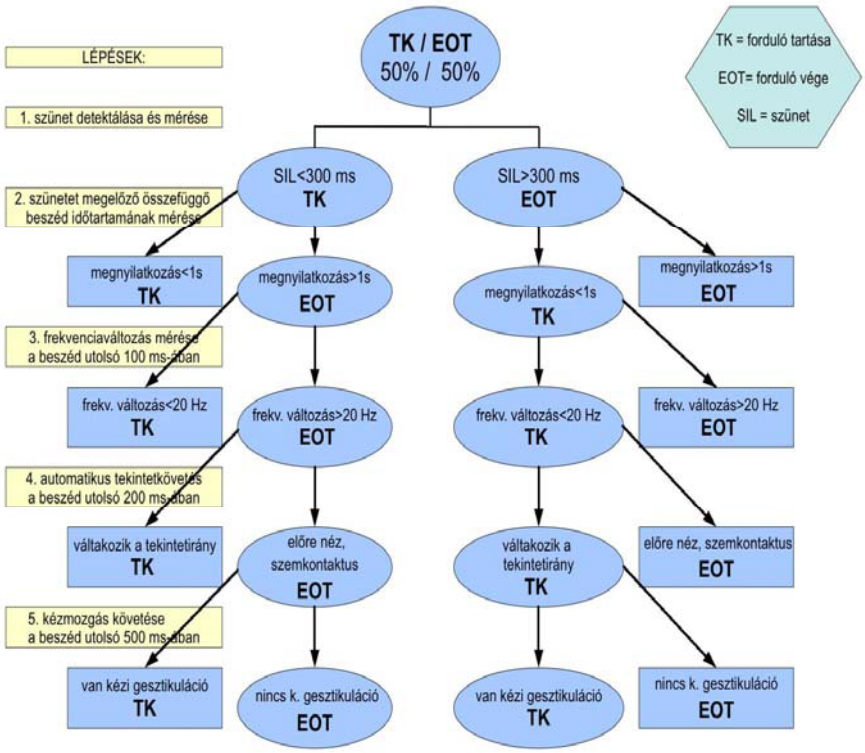
Ahogy korábban említettük, a kommunikatív aktusok és a támogató aktusok egységesen megnyilatkozásokba, a megnyilatkozások pedig társalgási fordulókba szerveződnek. A fordulók tehát több megnyilatkozást, s azon belül több kommunikatív aktust is tartalmazhatnak, határaikat a beszélőváltás jelöli ki. A beszélőváltás azonban nem véletlenszerűen történik a társalgásban: például egy kérdés elvárt következménye a válasz, egy javaslat elvárt következménye az elfogadás, a nyugtázás. Az egymásra következő fordulókból olyan szomszédsági párok bontakoznak ki, ahol a pár második tagja rendszerint az első párra adott elvárható, preferált válasz. Tehát az interakció elméleti modellezése során szintén érdemes párszekvenciákba összekapcsolni a kommunikatív aktusokat. Dialógus-modellekben általában két kommunikatív aktus alkot egy párszekvenciát [10]: az első kommunikatív aktust a gép nyújtja, a második aktus pedig az (emberi) felhasználó fordulója. Mivel a kommunikatív aktusok jellemzően előre megjósolható sorrendben követik egymást (pl. kérdés-válasz és kérés-teljesítés szekvenciákban) [11], [12], így az egyes aktusok jellemző jegyeinek

6 Ilyen eset például a *Jaj!* fölkiáltás.

7 Nyelvészeti pragmatikai terminussal élve a konvencionálisan indirekt illokúciós aktusok automatikus felismerésének lehetőségeit kívánjuk föltárni.

az annotációból való kinyerése megkönnyítheti a dialógusrendszer betanítását azok felismerésére és megfelelő válaszok automatikus generálására is. Ha adott az egyik rész, előre jelezhető a másik [1].

Annotációs rendszerünk lehetővé teszi azt, hogy a többi HuComTech-annotációval egybefűzve megvizsgáljuk a társalgási fordulók szomszédsági párokba való szerveződésének mozzanatait is. E vizsgálat pedig elvezethet minket kommunikatív aktusok közötti döntéshozást segítő, következő aktust jósló döntési fák létrehozásához is. Mivel a kommunikatív aktusok automatikus felismerésének, predikciójának és generálásának első lépése és egyben alapfeltétele a beszélőváltás predikciója is, ezért Abuczki Ágnes [13] kvantitatív vizsgálatokkal, adatbázis-lekérdezésekkel a fordulólezárás és a lehetséges váltási pont tipikus jegyhalmazát gyűjtötte össze Troung és munkatársai [14] modelljéből kiindulva, majd ezeket a jellemzőket vizuális jegyekkel kiegészítve egy döntésfába rendezte (l. 1. ábra).



1. ábra: Döntésfa a forduló lezárásának ('end-of-turn', rövidítése: EOT) és a forduló tartásának ('turn-keep', rövidítése: TK) megkülönböztetésére multimodális jegyek alapján [13].

Az 1. ábrán látható döntésfa a fordulózárás ('end-of-turn', rövidítése: EOT) és a forduló tartásának ('turn-keep', rövidítése: TK) megkülönböztetésére vállalkozik. A döntésfán látható öt lépés közül az első három akusztikai tényezőket, az utolsó két lépés pedig vizuális tényezőket tartalmaz. A beszélőváltás predikciójával egyidejűleg a szomszédsági párok tipikus mintázatának megfelelően, az egyes kommunikatív aktustípusok lekérdezések után kapott megkülönböztető jegyeire támaszkodva, a jegyeket a fenti példához hasonlóan döntési fába rendezve a következő kommunikatív aktust megjósló modellt hozhatunk létre. A pragmatikai annotáció mellett az audio- és videoszinten is annotált HuComTech-korpusz megbízható kiinduló bázisa lehet az egyes kommunikatív aktusok együtt járó jellemzői összegyűjtésének, ami hozzájárulhat az emberi beszélő által végrehajtott aktusok automatikus felismeréséhez. Az egyelőre még csak vázlatosan modellált dialógusrendszer feladata elsősorban „csupán" a fordulók végének detektálása lesz a tipikus fordulóvégi jellemzők (audio- és vizuális markerek) együttes előfordulása és bizonyos időtartamú események egymást követő sorrendje alapján. A megnyilatkozás végének detektálása után pedig a gépi ágens felteheti a beépített forgatókönyvének (*scenario*) megfelelő következő kérdést. Így a kérdések és válaszok láncából felépül a dialógus. A szomszédsági párok sorozatából épül ki a társalgás egésze, melynek során akár több témát is egymásba fűzhetünk. Ezért annotációnkba a tematikus kontroll szintjét is bevezettük, mellyel célunk az volt, hogy korrelációkat tudjunk megállapítani az egyes kommunikatív aktusok szekvenciális szerveződése, a fordulókezelés, valamint a globális diskurzusszerveződés mozzanatai között. Annotációs rendszerünkben megkülönböztetjük a témakidolgozás, az egyes társalgási témák motivált egymásba fűzése, illetve a motiválatlan témaváltás mozzanatait.

A pragmatikai annotáció utolsó szintjén a társalgás univerzumába kerülő új lexikai információkat jelöltük. Erre azért volt szükség, hogy a későbbiekben megvizsgálhassuk azon hipotézisünket, amely szerint az új információ bevezetése élénkebb, erőteljesebb gesztikulációval és nagyobb intenzitással jár együtt. [13] kvalitatív elővizsgálata a szemantikailag új lexikális információ kézi bejelölése után azt az eredményt hozta, hogy a gesztus csúcspontja (ún. *stroke*) és a szemantikailag legfontosabb verbális egység gyakran egybeesik. Ezt a feltételezést kvantitatív módszerekkel, vagyis a tervezett lekérdezések statisztikai elemzésével is kívánjuk igazolni a HuComTech-korpuszban.

3 A multimodális pragmatikai annotációs séma

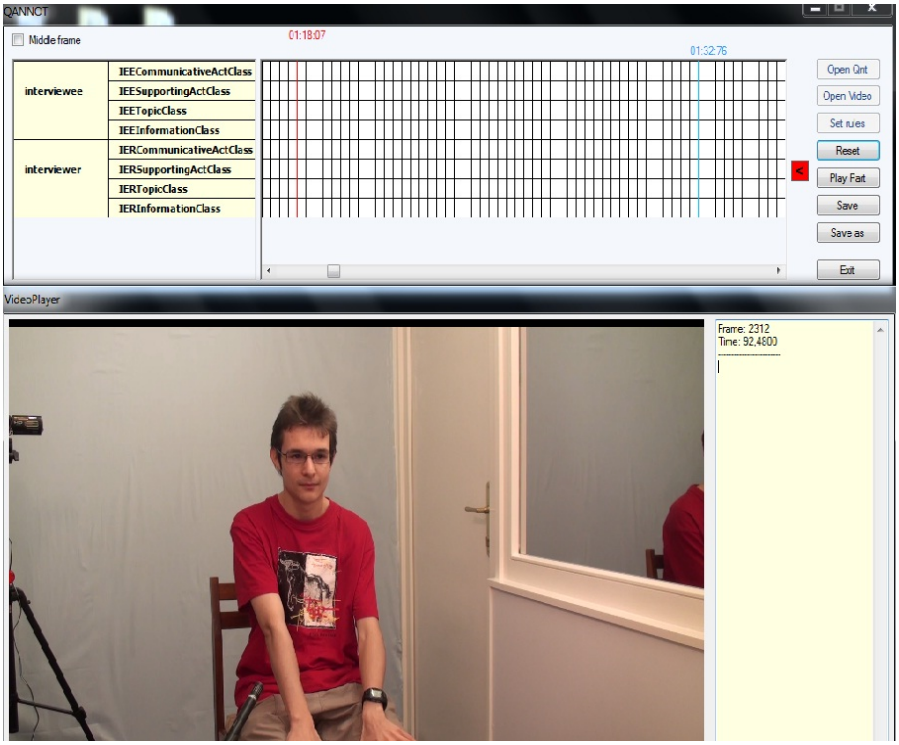
A HuComTech-projekt multimodális pragmatikai annotációjának sémáját az alábbi táblázatban összegezzük:

1. táblázat: A HuComTech-projekt multimodális pragmatikai annotációs sémája.

kommunikatív aktusok típusai (the level of communicative act types):
konstatívok (constatives) = ítélkezők: válaszadás, megerősítés, informálás, predikció, visszaemlékezés
direktívák (directives) = végrehajtók: kérés, parancs, javaslattétel
komisszívok (commissives)= elkötelezők: beleegyezés (pl. egy fogadásba), fölajánlás, ígéret
viselkedők (acknowledgements): üdvözlés, búcsúzás, elfogadás (pl. meghívásé)
indirekt (indirect)
nem azonosítható (none)
támogató aktusok szintje (the level of supporting acts):
visszajelzés (backchannel)
udvariassági marker (politeness marker)
javítás (repair)
nincs aktus (none)
tematikus kontroll szintje (the level of thematic control):
témakezdeményezés (topic initiation)
témakidolgozás (topic elaboration)
témaváltás (topic change)
információ szintje (the level of information type):
adott (given)
új (new)

A multimodális pragmatikai annotáció eszközét, a QANNOT-programot Szeghalmy Szilvia (Debreceni Egyetem) hozta létre 2010-ben a HuComTech-csoport számára. A QANNOT-ban az annotáció egysége – amely egységekhez *timestamp*eket (kezdő- és végpontokat) lehet rendelni – a *frame*. A pragmatikai annotáció jelenleg beállított szegmentálási egysége 8 frame per second, vagyis nyolc frame reprezentál egy másodpercet. Az annotáció során az annotátorok elsőként a kommunikatív aktusok típusainak címkéit helyezik el az annotálni kívánt videó idővonalán. Ezután a támogató aktusok címkézése következik, majd a témaváltás. Végezetül az adott-új információ címkézése történik. Mivel a QANNOT egyszerre jeleníti meg az összes annotációs szintet, így az annotátorok szimultán módon össze tudják hasonlítani és szinkronizálni egymással a különböző szinteken elhelyezett címkéket.[8]

8 Ez nemcsak a multimodális pragmatikai annotáción *belül* fontos, hanem akkor is, amikor a különböző annotációkat egybe kívánjuk vetni, s korrelációkat megállapítani például a Praatban zajló prozódiai és a QANNOT-ban zajló multimodális pragmatikai annotáció címkéi között.



2. ábra: A multimodális annotáció felhasználói felülete a QANNOT-programban.

4. Tervezett lekérdezések a HuComTech-korpuszon

A kutatás következő szakaszában (a 2011-es MSzNy konferencia időpontjáig) kvantitatív elemzést kívánunk végezni adatbázis alapú címkelekérdezések segítségével, melyekről előadásunkban részletesen be fogunk számolni. Multidimenziós vizsgálatot fogunk végezni, vagyis a dialógusok horizontális és vertikális szerkezetét egyaránt elemezni fogjuk a különböző típusú (audio, video, szintaktikai és pragmatikai) annotációk bizonyos szintjeinek (a diskurzus, a tekintetirány, a kommunikatív és támogató aktusok, valamint a tematikus kontroll szintjének) szimultán többszintű megjelenítése és együttes előfordulásukra vonatkozó címkelekérdezések segítségével.

A horizontális (szekvenciális) elemzés részeként az annotáció horizontális vetületét fogjuk vizsgálni, vagyis ennek segítségével az időben egymást követő jelenségek (elsősorban kommunikatív aktusok) mintázatát próbáljuk feltárni.

A vertikális címkeelemzés keretében pedig audio-, video- és pragmatikai címkék együttjárását keressük: első lekérdezéseink során azt vizsgáljuk meg, hogy az egyes kommunikatívaktus-típusok (konstatív, direktív, komisszív, viselkedő, indirekt) jellemzően milyen embléma típusú gesztusokkal (figyelem, egyetértés, nem egyetértés,

visszautasítás, kételkedés, számok és alak, valamint méret mutatása[9]) és milyen arckifejezésekkel (semleges, boldog, meglepett, szomorú, elgondolkodó, feszült[10]) (a kategóriák részletes bemutatásáért l. [15]) járnak vagy kezdődnek együtt (vagyis melyik kommunikatív aktusba esik bele egy gesztus vagy arckifejezés kezdőpontja). Ezeket az eredményeket olyan formában szeretnénk megkapni, hogy hány-hány darab emblématípus jelenik meg az egyes kommunikatív aktus-típusok végrehajtása közben. Vagyis a fenti vertikális természetű lekérdezések fő célja az egyes aktus-típusokat kísérő nem verbális-vizuális, nem verbális-akusztikus és verbális jegyek felfedése, amelyek szisztematikus rendszerbe foglalása és explicitté tétele elvezethet minket a kommunikatív aktusok automatikus felismeréséhez.

Következő lépésként, a szekvenciális (horizontális) elemzés során a kommunikatív aktusok egymást követő tipikus sorrendjeit szeretnénk megállapítani. Ezzel validálni szeretnénk a szomszédsági párok [11] által felállított tipikus láncolat alkalmazhatóságát magyar spontánbeszédkorpuszon is. Ezt a lekérdezést úgy fogjuk elvégezni, hogy diskurzusszinten[11] a záró (turn give közben végrehajtott) és a nyitó (turn take közben végrehajtott) kommunikatív aktusokat párosítjuk, majd a kapott aktuspárokat csoportosítjuk és megszámoljuk. Mivel a párszekvencia első fele előrejelzi a második felét - különösen formális, kanonikus szituációkban -, így ez a megközelítés grafikus és multimodális felhasználói felületek működtetéséhez egyaránt megfelelő feltételeket biztosít. Eredményeinkkel ezáltal nemcsak a kommunikatívaktus-típusok felismeréséhez, hanem azok automatikus generálásához és összefonásához, diskurzusba kapcsolásához is célunk hozzájárulni.

Következő lekérdezésünk arra a kérdésre keresi a választ, hogy az egyik beszélő által végrehajtott visszajelzés (*backchannel*) a másik beszélő által végrehajtott mely kommunikatívaktus-típusba és hány alkalommal esik bele.[12] Ezzel azt kívánjuk feltárni, hogy leggyakrabban milyen aktustípusra következik reakcióként a visszajelzés, vagyis mi a visszajelzés leggyakoribb funkciója.

A kommunikatív aktusok akusztikai markereinek feltárásához a Praat-program [16] áll rendelkezésünkre. A Praat-programban – melyben a HuComTech-korpusz audioannotációja zajlik - a spektogram horizontális irányban mutatja az időtartamot, vertikális irányban pedig a frekvencia (hangmagasság) skálázását (Hz mértékegységben). A fenti adatok millisecundumonkénti értékeinek feltöltése lehetővé teszi a felvételek fonetikai elemzését és fonetikai jellegű (például intenzitásra és alapfrekvenci-

9 Sémánkban a címkék angolul szerepelnek: attention, agree, disagree, refusal, doubt, numbers, size.

10 Sémánkban a címkék angolul szerepelnek: natural, happy, surprised, sad, recalling, tensed.

11 A HuComTech-korpusz audioannotációja tartalmaz egy diskurzusszintet, ahol a társalgás fordulókra van bontva [13]. A fordulókat a következő címkék jelölik: T (turn taking: a forduló átvétele/kezdete), K (turn keeping: 'a forduló megtartása'), G (turn giving: 'forduló átadása') és BC (backchannel: 'a hallgató fél rövid, figyelmet jelző visszajelzése'). Egy beszélő fordulóján belül akár több kommunikatív aktus is előfordulhat, tehát az audioannotáció további információkkal bővül a pragmatikai szinten.

12 Olyan visszajelzéseket (BC) vizsgálunk, amelyek kezdőpontja beleesik a másik beszélő által végrehajtott kommunikatív aktus időtartamába. Aktustípusonként egyesével szükséges lekérdezni a kommunikatív aktusok darabszámát és időtartamát, illetve a bennük végrehajtott visszajelzések darabszámát.

ára vonatkozó) lekérdezések végrehajtását. Ezek után elsődleges célunk az egyes kommunikatívaktus-típusok átlagos intenzitásminimumának, -maximumának és -átlagának lekérdezése lesz, annak érdekében, hogy ezekkel az eredményekkel is hozzájáruljunk az egyes aktustípusok megragadásához és formalizált leírásához, amely a későbbiekben elvezethet minket a beszélő kommunikatív szándékának automatikus felismeréséhez, illetve előrejelzéséhez.

5 Összegzés

A jelen tanulmányban bemutatott pragmatikai annotációs rendszer fő előnye abban rejlik, hogy univerzális kategóriákkal dolgozik, vagyis a felvételek nyelvétől függetlenül univerzálisan alkalmazható, hiszen a kommunikatív és a támogató aktusok típusai, valamint a tematikus kontroll tulajdonságai egyaránt univerzális jellemzői a társalgásnak. A rendszer interoperábilis XML-sémája lehetővé teszi az annotációs szempontok, annotálandó kommunikatív jelenségek bővítését újabb szintek és címkék bevezetésével. Ugyanakkor a fölöslegessé vált szintek és címkék is törölhetők (például a *none* címkét bevezetését követően hamarosan töröltük). A fordulók mint strukturális elemek és a kommunikatív aktusok típusai mint funkcionális elemek együttes szerepeltetése lehetővé teszi, hogy a fordulókból kibontakozó szomszédsági párokhoz megfelelő kommunikatívaktus-típusokat tudjunk rendelni. Mivel a QANNOT-program képes egyszerre megjeleníteni az összes annotációs szintet, így lehetővé válik az egyes szintek címkéinek szimultán összehasonlítása (például a kommunikatív aktusok összevetése az audio- és videoannotáció címkéivel), illetve a címkestatisztikai adatbázisba való feltöltés után bizonyos kommunikatív jelenségekre jellemző multimodális jegyhalmazok explicit formában történő felfedése. Ez közelebb vihet minket olyan multimodális jegyhalmazok meghatározásához és finomításához, amelyek segítségével nagy biztonsággal meg tudjuk jósolni a következő forduló kommunikatív aktusát/aktusait a társalgásban.

Mindezen megfontolásokat figyelembe véve a HuComTech-korpusz pragmatikai annotációja tevékenyen hozzájárul az ember-gép kommunikációs technológiák nyelvészeti aspektusainak modellezési lehetőségeihez. Ha a számítógépes nyelvészet adatbázisokra alapozva kívánja a kommunikációt modellálni, akkor annak szüksége van a pragmatikára, hiszen jól megalapozott pragmatikaelméleti döntéseket igényel annak meghatározása, hogy milyen legyen a társalgások pragmatikai annotációja. Ugyanakkor a pragmatika számára is nyereséggel jár a komputációs nézőpont, mert rákényszeríti a pragmatikusokat, hogy a kommunikatív nyelvhasználatra vonatkozó megállapításaikat explicit formában fogalmazzák meg, úgy, hogy azok formalizálásra alkalmasak legyenek és ezáltal algoritmizálhatóakká váljanak.

Bibliográfia

1. Bunt, H., Black, W.: The ABC of computational pragmatics. In: Bunt, H., Black, W. (eds.): Abduction, belief and context dialogue: Studies in computational pragmatics. John Benjamins, Amsterdam (2000) 1–46
2. Jurafsky, D.: Pragmatics and computational linguistics. In: Horn, L. R., Ward, G. (eds.): The handbook of pragmatics. Blackwell, Oxford (2002) 578–604
3. Németh T. E: Pragmatika. In: Kiefer F. (szerk.): A magyar nyelv. Akadémiai Kiadó, Budapest (2006) 222–261
4. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara (1997) 88–95
5. Ropolyi L.: Technika és etika. In: Fekete L. (szerk.): Kortárs etika. Nemzeti Tankönyvkiadó, Budapest (2004) 245–292
6. Abuczki Á., Bódog A., Németh T. E.: A multimodális pragmatikai annotáció elméleti alapjai az ember–gép kommunikáció modellálásában. In: Németh T. E. (szerk.) Ember-gép kapcsolat. A multimodális ember–gép kommunikáció modellezésének alapjai. Tinta Könyvkiadó, Budapest (2011, megjelenés alatt)
7. Németh T. E.: Megnyilatkozás: típus - példány. Néprajz és Nyelvtudomány Vol. 35 (1994) 69–101
8. Bach, K., Harnish, R. M.: Linguistic communication and speech acts. MIT Press, Cambridge (1979)
9. Verschueren, J.: Understanding pragmatics. Arnold, London (1999)
10. Bogdan, C., Kaindl, H., Falb, J., Popp, R.: Modeling of interaction design by end users through discourse modeling. In: Proceedings of the 13th international conference on Intelligent user interfaces. ACM, New York (2008)
11. Levinson, S. C.: Pragmatics. Cambridge University Press, Cambridge (1983)
12. Schlegoff, E. A.: Sequence organization in interaction: A primer in conversation analysis. Cambridge University Press, Cambridge (2006)
13. Abuczki, Á.: A multimodális interakció szekvenciális elemzése. In: Németh T. E. (szerk.) Ember-gép kapcsolat. A multimodális ember–gép kommunikáció modellezésének alapjai. Tinta Könyvkiadó, Budapest (2011, megjelenés alatt)
14. Troung, K. P., Poppe, R., Heylen, D.: A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech (2010) 3058–3061
15. Földesi, A.: Unimodális funkcionális annotáció a HuComTech-korpuszban. In: Bódog, A. (szerk.): Az ember–gép kommunikáció technológiájának elméleti alapjai. IKUT zárókötet. (előkészületben)
16. Boersma, P., Weenink, D.: Praat: doing phonetics by computer 5.0.02. Institute of Phonetic Sciences, University of Amsterdam (2007) http://www.praat.org

Metaforikus kifejezések szerkezeti jellemzői

Babarczy Anna

BME Kognitív Tudományi Tanszék, Budapest 1111, Egry József u. 1.
babarczy@cogsci.bme.hu

Kivonat: A tanulmány a tág értelemben vett metaforikus kifejezések előfordulási jellemzőit vizsgálja magyar írott és kvázi beszélt nyelvi korpuszban. Az elemzés célja olyan lexikális kifejezések vagy morfoszintaktikai konstrukciók kézi azonosítása, melyek a korpuszban előforduló szavak alapjelentésétől eltérő jelentésére utalnak. A fő kérdések, melyekre választ keresünk a következők: (a) Javítható-e számottevően a gépi metaforaazonosítás teljesítménye, ha a metaforikus jelentést jelző kifejezést nem csak egymondatos ablakon belül, hanem annál távolabb is keressük? (b) Található-e olyan nyelvtani szerkezet vagy konstrukció, amely jellemző a metaforikus kifejezésekre, és amely figyelembe vétele megkönnyítheti a metaforák gépi azonosítását? és (c) Megfigyelhetők-e tipikus eltérések a fenti két tekintetben különböző szövegfajták között?

1 Bevezetés

1.1 Metaforák a kognitív nyelvészetben és a nyelvtechnológiában

Az elvont nyelv kérdése egy klasszikus nyelvfilozófiai problémára vezethető vissza, ami magyarázatot keres arra a kérdésre, hogy honnan származhat az a tudás, amiről nem lehet közvetlen tapasztalatunk. Két egymással ellentétes, bár egymást nem teljes mértékben kizáró hipotézis terjedt el a szakirodalomban ennek magyarázatára: a fogalmi metafora elmélet [13], [14] és a nyelvi elvonatkoztatás elmélete [25].

A fogalmi metafora elmélet arra az empirikus megállapításra épül, hogy az emberi nyelvben (többé-kevésbé) szisztematikus kapcsolat létezik adott konkrét tartományok és adott elvont tartományok között: hideget és meleget kifejező szavak például viszonylag konzisztensen írnak le érzelmi állapotokat, mint ahogy téri relációkat meghatározó kifejezéseket szisztematikusan alkalmazunk idői relációk leírására. Az elmélet szerint tehát az elvont fogalmak elsajátítása és mentális reprezentációja a konkrét tudásból származik, ami pedig az embert körülvevő világ testi tapasztalatában gyökeredzik.

Az elvont nyelv kérdésének másik megközelítése a nyelvi elvonatkoztatás elmélete [25], ami pszicholingvisztikai kísérletek eredményein és a gépi nyelvtanulás tapasztalatain alapul. Az elmélet szerint mind a konkrét, mind pedig az elvont fogalmak elsajátítását a nyelvi inputból kivont statisztikai minták segítik. A feladat kivitelezhetőségét a nyelvnek az az empirikusan bizonyított tulajdonsága biztosítaná, hogy egy-egy nyelven belül a hasonló disztribúciójú szavak többnyire azonos fogalmi tartományba

tartoznak – ha a disztribúció fogalmát megfelelő pontossággal definiáljuk [15], [8], [19].

A természetes nyelvi szövegek sekély szemantikai elemzése, azaz az argumentumok és határozók tematikai azonosítása a gépi nyelvfeldolgozás egyik kulcskérdése. A feladat egyik legnehezebb problémája a formailag hasonló, de szemantikailag eltérő argumentumok/határozók megkülönböztetése, azaz a tág értelemben vett metaforikus kifejezések helyes azonosítása. Az alábbi mondatokban, például, a *labdával* argumentum a konkrét értelemben vett játszás eszköze, míg az *ötlettel* és a *játszott* között más típusú kapcsolat van, mivel az ige itt metaforikus értelemben szerepel. Amint a (3) példa mutatja, nem állíthatjuk azonban azt, hogy egy ötlet nem lehet eszköz.

(1) Eljátszott az ötlettel.
(2) Eljátszott a labdával.
(3) Mindenkit feldühített az ötlettel.

A sekély szemantikai elemzés terén két elterjedt statisztikai megközelítés létezik: az emberi erővel annotált korpuszból való gépi tanulás [17] és a teljesen automatikus gépi tanulás [3]. Az előbbi rendszer morfológiailag és/vagy szintaktikailag elemzett, és argumentumcímkékkel (pl. PATIENS, AKTOR, HELY, MÓD) ellátott korpuszból von ki statisztikai mintákat a predikátum-argumentum előfordulásokra vonatkozóan, és ezek alapján azonosítja az argumentumszerkezeteteket új szövegekben. A másik, kevésbé erőforrás-igényes, de kevésbé sikeres módszer csak morfológiai és/vagy szintaktikai annotációval ellátott korpuszból alkot lexikont, melyben a predikátumokhoz argumentum-valószínűségeket rendel. Jelenleg egyik módszer sem képes a metaforikus szerkezetek megbízható azonosítására.

1.2 A gép metaforaazonosítás előző eredményei

A kutatás korábbi eredményeinkre épít, ahol a fogalmi metafora elméletből kiindulva forrás- és céltartományi szavak együttes előfordulása alapján próbáltunk metaforikus kifejezéseket azonosítani egy korpuszban [1]. A metaforajelző szavakat három különböző módon definiáltuk. Az első egy asszociációs kísérletre épült, ahol egyetemi hallgatók a forrás- és céltartományokat képviselő szavakhoz szorosan kapcsolódó szavakat soroltak fel. A második módszer az így kapott szólistákat szótári szinonimákkal egészítette ki, a harmadik módszer pedig a kísérleti korpuszból kézzel kivont forrástartományi szavakat vette alapul a tesztkorpusz metaforáinak azonosításához. Mindhárom kísérlet esetében a forrás- és céltartományi szópárokat egy-egy mondaton belül kerestük. A legjobb eredményeket a harmadik, korpuszalapú módszer adta, de itt is 50 százalék alatt maradt mind a találati arány, mind pedig a pontosság. Az eredmények tehát azt mutatják, hogy egy forrás-cél tartománypároson belül nem bármilyen asszociáció vezet metaforikus értelmezéshez, és a valóban metaforicitásra utaló relációk mibenléte leginkább az adott szöveg nyelvi tulajdonságain múlik. Az is kiderült, hogy nem minden esetben van szükség egy mondaton belül mindkét tartománybeli kifejezésre a metaforikusság értelmezéséhez. Mindez a metaforák koncep-

tuális természete helyett azok disztribúciós tulajdonságainak fontosságára világít rá. A módszer gyenge eredményei azonban arra utalnak, hogy az eddigieknél részletesebb elemzésre van szükség. Erre tesz kísérletet a jelen tanulmány a nemzetközi irodalomból már ismert eredmények felhasználásával.

Deignan főként a metaforikus kifejezésekben szereplő szavak grammatikai és kollokációs természetét vizsgálva arra mutatott rá, hogy a pszicholingvisztikai kísérletekben használt példák problémákhoz vezethetnek [4], [5]. A nyelvi metaforák grammatikai viselkedésének vizsgálata is olyan fontos részletekre világít rá, amelyeket a konceptuális metaforaelméletben figyelmen kívül hagynak. Ugyancsak Deignan elemzéseiből derül ki, hogy a különböző szavak, kifejezések többnyire más-más grammatikai jellemzőkkel, illetve logikai relációkkal rendelkeznek a szó szerinti és a metaforikus használatban. Az „az emberi viselkedés állati viselkedés" konceptuális metafora esetén például azok a szavak, amelyek a forrástartományban szerepelnek, és entitásokat jelölnek, metaforikus használatukban többnyire igeként vagy melléknévként fordulnak elő. A szerző egyéb metaforatípusok vizsgálata alapján számos példával mutatja meg, hogy metaforikus használatban a szavak jóval kevesebb grammatikai szabadsággal rendelkezek, mint amikor szó szerinti jelentésükben jelennek meg. Ez azt jelenti, hogy a forrástartományban lévő entitások közti logikai reláció nem egyszerűen megismétlődik a céltartományban, ahogyan azt a kognitív metaforaelmélet jósolná, hanem át is alakul: a szavak metaforikus jelentésükben önálló életet kezdenek élni.

A British National Corpus egy részének kézi elemzése precízebb megállapításhoz vezet: egy új elemzés szerint az itt előforduló 241 metaforikus kifejezésből 164-et ige vezetett be [22]. Ez a megfigyelés összecseng a gépi metafora azonosítás egyik klaszszikus tanulmányának kitételével, amely szerint az ige által bevezetett metafora operatív definíciójának tekinthetjük azt a tulajdonságát, hogy a metaforikus kifejezésekben valamiféle szelekciós megkötés megszegése fordul elő [26]. Erre a megfigyelésre épül Fass met* elnevezésű félig-meddig gépesített rendszere [6], amely szó szerinti, metaforikus, metonimikus és anomalikus ige alapú kifejezéseket kísérel meg megkülönböztetni egymástól. A rendszer három lépésben működik. Először egy kézi erővel alkotott szelekciópreferencia-szótár és részontológia segítségével különíti el a szó szerinti jelentést (ahol az argumentumok megfelelnek az ige szelekciós preferenciáinak) minden nem szó szerinti jelentéstől (ahol az argumentumok nem felelnek meg a szelekciós preferenciáknak). A második lépésben a rendszer egy forrás- és céltartomány részontológiával veti össze a vonzatszerkezetet, és ha megfelelést talál, metaforikusnak címkézi a kifejezést. A módszer problémája az, hogy a jelentős kézi beavatkozás ellenére vagy erősen alulgenerál (nem találja meg a metaforákat) vagy erősen túlgenerál (mindent metaforának ítél). Az eredmények szinte kizárólag azon múlnak, hogy mi szerepel a kézileg megalkotott ontológiában. Ez a probléma visszavezethető a metaforák kézi azonosításának bizonytalanságára, amit a rendkívül alacsony annotátorok közötti egyetértés is mutat [1].

Shutova és munkatársai új munkájukban az argumentumstruktúra módszert a korpuszból kinyert forrás- és céltartomány-szólista keresési módszerrel kombinálták [22]: olyan kifejezéseket kerestek, ahol az ige jelöli a forrástartományt és az alany vagy a tárgy a céltartományt. A forrás-, illetve céltartományi szavakat klaszteralgoritmusok segítségével korpuszból állították össze. Az igék közül kiszűrték azokat,

amelyek – szintén korpuszelemzések szerint – gyenge szelekciós preferenciákat mutatnak. A szerzők hipotézise szerint az olyan kifejezések, ahol erős szelekciós preferenciájú forrástartományi igék céltartományi vonzatokkal fordulnak elő, metaforikusnak tekinthetők. A módszer eredményeként 79 százalékos pontosságot értek el. Az értékelés azonban nem egy „gold standard" mintához képest történt, hanem a gépi elemzés eredményének utólagos kézi ellenőrzésével. Ebből következően a rendszer fedési arányáról nincs információnk, és az eredmények nem vethetők össze más módszerek ereményeivel.

Végül Baumer és munkatársai egy hasonló klaszteralapú megoldást egészítenek ki szemantikaiszerep-címkézéssel (Semantic Role Labelling, SRL) [2]. Az SRL segítségével a szintaktikai elemzéssel ellátott korpuszban különböző szintaktikai szerkezetekből is ki tudják vonni a tematikai szerepeket (pl. az angol passzív szerkezet alanyáról megállapítható, hogy az ige páciens argumentuma). A rendszer jelenleg kísérleti stádiumban van.

2 A metaforikus kifejezések kézi elemzése

A fenti eredmények tehát korlátozott sikert értek el, ami részben azzal magyarázható, hogy még mindig nincs pontos képünk a metaforák mibenlétéről. A jelen elemzés célja ezért a konceptuális metaforaelmélettől elvonatkoztatva olyan lexikális kifejezések vagy morfoszintaktikai konstrukciók kézi azonosítása és elemzése, melyek a korpuszban előforduló szavak alapjelentésétől eltérő jelentésére utalnak (a továbbiakban ezt metaforikus jelentésnek fogjuk nevezni). Alapjelentés alatt a szó konkrét, fizikai vagy téri jelentését értjük. Egy „metaforajelző" elem lehet egyetlen szó, ahogy a (4) példában a *kétségbeesés* jelzi az *összefűz* ige metaforikus jelentését a predikátum szelekciós megkötéseinek megszegésével. Ezzel szemben az (5) mondat metaforikus jelentése csak a tágabb kontextusból következik, amiből kiderül, hogy a királyi udvarról van szó, és annak a támogatásában való bizalomról.

(4) A halálra rémült pár (amennyiben a házasság valamely ősi formája nem is, a kétségbeesés bizonyára összefűzte őket) egyre nehezebben haladt. (National Geographic)
(5) Ne csak az udvarra építs. (Filmfelirat)

2.1 Korpusz és annotációs rendszer

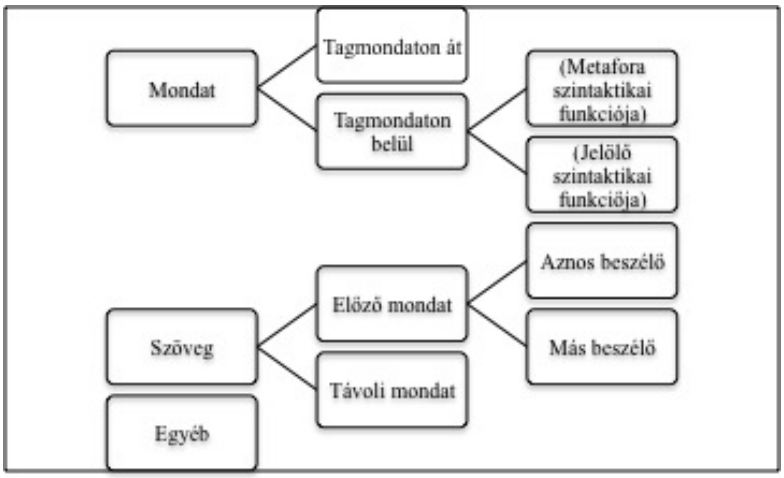
Az elemzés alapjául egy regényből, a National Geographic magyar nyelvű kiadásából és filmfeliratokból álló összesen 36.355 szavas korpusz szolgált. A korpusz összetételét az 1. táblázat mutatja. Az elemzési korpusz egy nagyobb korpusz része, a három szövegtípust arányosan reprezentálja. A szövegkontextus jelentősége miatt a szövegek nem mondathatárokon, hanem epizódushatárokon vannak elválasztva.

1. táblázat: Szövegszavak száma korpuszban.

Regény	National Geographic	Filmfelirat	Összes
19 544	7 252	9 559	36 355

Az elemzés a nemzetközi gyereknyelvikorpusz-kutatásokban ismert CHAT formátumban a CLAN annotációs és statisztikai elemzőprogramok használatával készült. A formátum és az eszközök előnye, hogy lehetővé teszik a szöveg rugalmas tagolását és hosszabb távú, akár mondatokon átívelő függőségek kezelését.

Az annotációs rendszer a metaforikus kifejezés és az azt jelző elem egymástól való távolságát, mindkettő tagmondatbeli, illetve egymáshoz viszonyított grammatikai funkcióját és lexikális azonosságát jelöli. A rendszer felépítését az 1. ábra mutatja némi egyszerűsítéssel. Az első szint az jelöli, hogy a metafora és az azt jelző elem egy mondaton belül, vagy két különböző mondatban jelenik-e meg. Az ábrán „egyéb" címkével jelölt kategóriába olyan esetek tartoznak, ahol nem lehet metaforicitást jelző elemet azonosítani (mint például a fenti (5) mondat esetén), vagy a metaforikus szó morfológiai alakja jelzi a metaforicitást (pl. *mélységesen*). Mind a metafora, mind pedig a metaforicitásra utaló elem szintaktikai szerepét (ige, alany, tárgy, egyéb vonzat vagy határozó) valamint egymáshoz viszonyított nyelvtani funkcióját (fej-dependens) jelöli az annotáció.



1. ábra: Az annotációs rendszer szintjei.

2.2 Eredmények

A fő kérdések, melyekre választ keresünk a következők: (a) Javítható-e számottevően a gépi metaforaazonosítás teljesítménye, ha a metaforikus jelentést jelző kifejezést nem csak egy-mondatos ablakon belül, hanem annál távolabb is keressük? (b) Talál-

ható-e olyan nyelvtani szerkezet vagy konstrukció, amely jellemző a metaforikus kifejezésekre, és amely figyelembevétele megkönnyítheti a metaforák gépi azonosítását? és (c) Megfigyelhetők-e tipikus eltérések a fenti két tekintetben különböző szövegfajták között?

Az elemzés összegzett eredményeit a 2. táblázat mutatja. A mondaton kívüli metaforicitásra utaló elemek (lásd 6. példa) alacsony átlagos valószínűsége (10%) arra utal, hogy nem javítható jelentősen az automatikus gépi azonosítás teljesítménye a keresőablak tágításával. A szövegtípusok között azonban van némi különbség: a beszélt nyelvet reprezentáló filmfelirat korpuszban valamivel gyakoribb, 17%, a metaforikus mondatot megelőzően előforduló metaforicitás jelző elem (az írott és a beszélt szövegek közötti különbség statisztikailag szignifikáns, Chi2 = 20.9, p = 0.002, valószínűleg nem a véletlen műve).

(6) - És mondja csak Bondy úr, hogyan jutott erre a **gondolatra**?
- Hogyan? – válaszolta G. H. Bondy szórakozottan. – Tulajdonképpen hogy az igazat megvalljam, az öreg van Toch **vezetett rá.**

A néhány mondaton átívelő metafora elemzéséből az is kiderül, hogy a metaforicitást jelző elem nem feltétlenül a metaforikus szót tartalmazó mondatot közvetlenül megelőző mondatban jelenik meg, hanem ennél nagyobb is lehet a távolság.

2. táblázat: A metaforikus kifejezések és a metaforicitást jelző nyelvi elemek egymástól való távolsága.

Metaforicitás jelző	Regény	National Geographic	Filmfelirat	Összes (átlag)
Nem azonosítható (%)	1%	2%	9%	2%
Mondaton kívül (%)	6%	8%	17%	10%
Mondaton belül (%)	93%	90%	75%	86%
Összes N (100%)	147	62	60	269

Összesen 237 olyan metaforikus kifejezés fordul elő a korpuszban, ahol a metafora és a metaforicitásra utaló elem egy mondatban jelenik meg. Az ilyen esetek túlnyomó többségében (223 metafora), a két elem egy tagmondaton belül található. A 3. táblázat az egy tagmondaton belül előforduló metaforikus kifejezés és metaforajelző elem egymáshoz való nyelvtani viszonyának valószínűségeit mutatja. A fej-módosító viszony jelzős (7a), névutós (7c), birtokos (7b), stb. szerkezetekre utal, az ige-vonzat viszony pedig olyan tagmondatokra, ahol a metaforikus kifejezés az ige, a metaforicitást jelző szó pedig az ige nyelvtani alanya (8a), tárgya (8b) vagy más esetű vonzata (8c). Az egyéb kategóriába azok a tagmondatok tartoznak, ahol a metafora és a metaforicitásra utaló kifejezés is valamilyen bővítmény.

(7a) ... termékeny vita folyt ...
(7b) ... egy régi vita lángját ...
(7c) ... felügyelete alá helyezték ...
(8a) ... sok történet kering ...
(8b) Ne keverj bele személyes érzelmeket.
(8c) ... kockázatos ugrás volt az ismeretlenbe.

A számokból kiderül, hogy a British National Corpus elemzési eredményeinek megfelelően a metaforikus értelemben használt kifejezések többsége a magyar korpuszban is ige, és a helyes értelmezést segítő kifejezés a bővítménye. Az ilyen esetek egy részében maga a vonzatkeret kínálja a metaforikus értelmezést (pl. *A Róka nem ad a pontosságra*), míg máskor a vonzat lexikális tulajdonságai a meghatározóak (pl. *...ugyanúgy süllyedne el a mi kultúránk*).

3. táblázat: Az egy tagmondatban előforduló metaforikus kifejezések és a metaforicitást jelző nyelvi elemek viszonya.

Metafora -- Jelző	Regény	National Geographic	Filmfelirat	Összes (átlag)
Fej -- Módosító (%)	23%	9%	4%	12%
Módosító -- Fej (%)	15%	9%	4%	10%
Ige – Vonzat (%)	58%	80%	89%	76%
Egyéb (%)	4%	2%	2%	3%
Összes egy tagmondaton belül N	124	54	45	223

Az egy mondaton belül, de két különböző tagmondatban megjelenő metafora és metaforicitás jelző párok túlnyomó többsége beleillik a 3. táblázatban felsorolt grammatikai szerkezetek valamelyikébe, bár egy-egy nehezebben elemezhető konstrukciót is találunk, mint például a (9) mondat.

(9) Úgy látszott, rövidesen leomlanak az utolsó **korlátok**, melyeket a világtengerek mind ez ideig az emberiség **fejlődése elé** emeltek.

Az automatikus gépi metaforaazonosítás szempontjából a fenti megfigyelések annyit jelentenek, hogy a vonzatkeretek és a vonzatszelekciós preferenciák beépítése a rendszerbe elvben jelentősen javíthatja a teljesítményt, amint ezt a nemzetközi tapasztalatok is mutatják. A számokból az is kiderül azonban, hogy más visszatérő grammatikai konstrukciót is találunk a metaforikus kifejezések között: a metaforicitást jelző kifejezés gyakran módosítja a metafora fejet, vagy megfordítva, a metafora módosítja a jelző elemet. Bár a nyelvtani konstrukció más, a metaforicitás elvi meghatározása megmarad: a kollokációs-szelekciós preferenciák megszegése jelzi a nem szó szerinti értelmezést. Továbbra is kulcskérdés marad tehát, hogy milyen módszerrel definiálhatjuk a szelekciós preferenciákat a pontos eredmények eléréséhez.

Bibliográfia

1. Babarczy, A., Simon, E., Bencze, I., Fekete, I.: A metaforikus nyelvhasználat korpuszalapú elemzése. In: Tanács,A., Vincze, V. (szerk): VII Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged. (2010)
2. Baumer, E.P.S., White, J.P., Tomlinson, B.: Comparing Semantic Role Labeling with Typed Dependency Parsing in Computational Metaphor Identification. Workshop on Computational Approaches to Linguistic Creativity (CALC-10) at HLT/NAACL (Los Angeles, CA) (2010)
3. Burgess, C., Lund, K.: Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes Vol. 12 (1997) 177–210
4. Deignan, A.: Metaphor and corpus linguistics. John Benjamins, Amsterdam/Philadelphia (2005)
5. Deignan, A.: Corpus linguistics and metaphor. In: Gibbs Jr., Raymond W. (szerk.): The Cambridge Handbook of Metaphor and Thought. Cambridge University Press, Cambridge (2008) 280–294
6. Fass, D.: met*: A method for discriminating metonymy and metaphor by computer. Computational Linguistics Vol. 17, No. 1 (1991) 49–90
7. Gentner, D., Holyoak, K. J., Kokinov, B. N. (eds): The analogical mind: perspectives from cognitive science. MIT Press, Boston (2001)
8. Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., Trueswell, J.: Hard words. Language Learning and Development Vol. 1 (2005) 23–64
9. Heywood, J., Semino, E., Short, M.: Linguistic metaphor identification in two extracts from novels. Language and Literature Vol. 11 (2002) 35–47
10. Kintch, W.: Predication. University of Colorado Technical Report 99-02 (1999)
11. Kintsch, W.: Metaphor comprehension: a computational theory. Psychonomic Bulletin and Review Vol. 7, No. 4 (2000) 257–266
12. Kövecses, Z.: Metaphor: A Practical Introduction. Oxford University Press, Oxford (2002)
13. Lakoff, G.: The contemporary theory of metaphor. In: Ortony, A. (ed.): Metaphor and Thought (2nd ed.). Cambridge University Press, Cambridge (1992)
14. Lakoff, G., Johnson, M.: Metaphors we live by. University of Chicago Press, Chicago, IL. (1980)
15. Landauer, T. K., Dumais, S. T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review Vol. 104, No.2 (1997) 211–240
16. Martin, J.H.: A corpus-based analysis of context effects on metaphor comprehension. In: Stefanowitsch, A., Gries, S.Th. (eds): Corpus-Based Approaches to Metaphor and Metonymy. de Gruyter, Berlin (2006) 214–236
17. Pradhan, S., Hacioglu, K., Ward, W., Jurafsky, D., Martin, J. H.: Support Vector Learning for Semantic Argument Classification. Machine Learning Journal Vol. 60, No. 1 (2005)
18. Saffran, J. R. , Aslin, R. N., Newport, E. L.: Statistical learning of 8-month-olds. Science Vol. 274 (1996) 1926–1928
19. Schutze, H.: Dimensions of meaning. In: Proceedings of Supercomputing Vol. 92 (1992) 787–796
20. Schwaneflugel, P.J. (ed.): The psychology of word meanings. Lawrence Erlbaum Associates, Hillsdale, NJ (1991)
21. Shutova, E, Sun, L., Korhonen, A.: Metaphor Identification Using Verb and Noun Clustering. In: Coling 2010 (2010)
22. Shutova, E., Teufel, S.: Metaphor corpus annotated for source - target domain mappings. In: Proceedings of LREC 2010. Malta (2010)

23. Simon, E., Szamarasz, V.: Preparations for a multilingual corpus analysis of metaphor. Doktorandusz konferenciaelőadás. Budapest (2008)
24. Steen, G.: Towards a procedure for metaphor identification. Language and Literature Vol. 11 (2002) 17–34
25. Vinson, D. P., Vigliocco, G.: Semantic feature production norms for a large set of objects and events. Behavior Research Methods Vol. 40, No. 1 (2008) 183–190
26. Wilks, Y.: Making preferences more active. Artificial Intelligence Vol. 11, No. 3 (1978) 197–223

VI. Szemantika

Az intenzionalitás számítógépes nyelvészeti kezelése – avagy a ℜeALIS λ szintfüggvénye

Alberti Gábor[1]

PTE BTK Nyelvtudományi Tanszék
ℜeALIS Elméleti és Számítógépes Nyelvészeti Kutatócsoport
7624 Pécs, Ifjúság útja 6.
alberti.gabor@pte.hu

Kivonat: Kutatócsoportunk szeme előtt továbbra is [5], [6] az a hosszú távon kifizetődő cél lebeg, miszerint az intelligens számítógépes nyelvészeti célokat (pl. fordítás, kivonatolás) az egymással kommunikáló humán interpretálói „elmék" ℜeALIS-modelljének [1]-[3] implementálására alapozva kívánjuk megvalósítani. A jelen munkaszakaszban a mondatok (alkotta diskurzusok) intenzionális jelentésrétegének megragadását tűztük ki, ami első lépésben az elmélet kínálta elvek és ötletek [8] specifikálását és célorientált formalizálását jelenti, második lépésben pedig az erre épülő implementációt. Döntően magyar lexikai tételeken mutatom be az intenzionalitás „tetten érését" és formális megragadását, ami a legkisebb toldalékok komplex jelentéstani analízisétől, a legkülönfélébb szófajba eső szavak elemzésén keresztül, nagyobb diskurzusegységek interpretálói információállapotba való beágyazódása intenzionális tényezőinek feltárásáig terjed. Megközelítésünk kiemelkedő erényének tartjuk, hogy nemcsak az „üzenetet" alkotó szavak puszta jelentéséből összeálló információt tárjuk fel és implementáljuk, hanem az üzenet megbízhatóságát is, valamint az üzenet forrását jelentő interpretáló információállapotának releváns tényezőit, a grice-i értelemben vett „ideális beszélői" karaktertől való eltérés elemzése révén.

Kulcsszavak: reprezentacionalista dinamikus diskurzusszemantika, intenzionalitás, információállapot, mód és modalitás, aspektus

1 Bevezetés

Minthogy középtávon kifinomult gépi fordításra és megbízható információkivonatolásra törekszünk, ezúttal egy olyan rövid távú projektet indítottunk, ami a poszt-montagoviánus [11], (S)DRT-re alapozott [15] [9], ℜeALIS nevű [1] [2] reprezentacionalista dinamikus diskurzusszemantika megközelítésében (2. szakasz) a diskurzusreferensek „intenzionális szintjeinek" [8] a gyakorlati kidolgozására irányul, majd a

[1] A szerzőt e cikk alapjait jelentő kutatásaiban az OTKA T60595 sz. projektje támogatta, a konferencia-részvételt pedig a TÁMOP-4.2.1.B-10/2/KONV/2010/ KONV-2010-0002 (A Dél-dunántúli régió egyetemi versenyképességének fejlesztése). Értékes megjegyzéseikért elsősorban a ℜeALIS ESzNy Kutatócsoport következő tagjainak szeretnék köszönetet mondani: Kleiber Juditnak, Károly Mártonnak és Kilián Imrének.

kapott reprezentációk implementálására az egymással kommunikáló interpretálói „elmék" komplex modelljében – ahogyan azt a ℜeALIS formálisan megragadja [4] négy belső függvénye segítségével: a formulaépítő σ-ról [6] [18], a horgonyzó/azonosító α-ról [7], a „dobozszint"-kijelölő λ-ról [8], [16] illetve a kurzor szerepű κ-ról van szó.

A projekt első felében tehát – megalapozandó az implementációt – a ℜeALIS elméleti konstrukcióit bizonyos nyelvi elemek csoportjaira alkalmaztuk, döntően magyar lexikai elemekre (3-4. szakasz). Olyan specifikált formális reprezentációkat dolgoztunk ki, amelyek pontosan megragadják az érintett morfémák és szavak összetett intenzionális karakterét, a mód és modalitás toldalékaira, az aspektusjelölő elemekre, különféle modális (segéd-) igékre, adverbiumokra, melléknevekre és partikulákra (pl. *bevesz*, *fog*, *valószínűleg*, *állítólagos*, *is*). A második projektszakaszban belefogtunk e reprezentációk implementálásába a kommunikáló interpretálói „elmék" ℜeALIS-modelljében [16]. A nyelvi elemek komplex intenzionális karakterizálásának a feladata, a λ szintfüggvénynek köszönhetően, végső soron arra redukálódik, hogy a DRS stílusú „dobozstruktúrában" minden egyes referenshez hozzárendeljünk egy $\gamma=\langle\langle\mu_1,\tau_1,i_1,\pi_1\rangle, \langle\mu_2,\tau_2,i_2,\pi_2\rangle, ..., \langle\mu_k,\tau_k,i_k,\pi_k\rangle\rangle$ „világocska-indexet" – vagy még inkább egy $\Gamma=\{\gamma^1, \gamma^2,..., \gamma^N\}$ indexhalmazt – e „dobozstruktúrában" elfoglalt pozíciójuk (pozícióik) / szintjük (szintjeik) kifejezése végett. Hamarosan kiderül, hogy a rendezettnégyes-sorozatok e Γ halmaza miből is áll össze, és hogy ez a matematikai konstrukció hogyan képes egységesen megragadni a legkülönfélébb nyelvi kifejezésekben rejlő intenzionalitást, illetve a szövegkörnyezet és a kontextus adta intenzionális hatásokat (5. szakasz).

2 A ℜeALIS alapjai

Mindenekelőtt felvázolom a jelen tárgyalásunk szempontjából releváns vonásait annak a háttérelméletnek, amelyen a szemantikai elemzések, a DRS stílusú reprezentációk és a számítógépes implementáció lépései nyugszanak.

A ℜeALIS (*REciprocal And Lifelong Interpretation System*, azaz Kölcsönös és Élethossziglani Interpretációs Rendszer) olyan új poszt-montagoviánus [11] elméletként mutatható be, amely a koherens (kis-)diskurzusokká összeálló mondatok formális jelentéselemzését nyújtja [15] [9], középpontjában az „interpretálók" lexikai, személyközi és kulturális / enciklopédikus tudásának egy *élethossziglani* modelljével, mely az interpretálók egymásról való *kölcsönös* tudását is megragadni hivatott. A teljes (40 oldalas) definíciós rendszer elérhető angolul az interneten ([1] http://lingua.btk.pte.hu/realispapers), magyarul pedig egy idén megjelent könyvben [2]; az elmélet különféle aspektusairól és alkalmazásairól pedig mostanában számos publikáció látott napvilágot [3]-[8] [16] [18].

Ami most igazán releváns, az a Kamp-féle DRS-ek újfajta felhasználása: az interpretálói információállapotok élethossziglani reprezentációi gyanánt lehet őket alkalmazni. Nyilván gigantikus dobozstruktúrák adódnak így, de matematikai tartalmuk alig bonyolultabb, mint az eredeti DRS-eké; a beágyazott „dobozrendszerek" viszont – ezek a logikai műveletekre nézve nem zárt, véges „információtárak" – készen kínálkoznak a Montague-féle formális diskurzus-szemantikában használatos (végtelen) *lehetséges világok* [11] helyettesítésére [8]; melyek megalapozottsága korántsem megfelelő [19]. A korlátlanul egymásba ágyazható „dobozok" segítségével

ugyanis meg tudjuk ragadni az interpretálói hiedelmek, vágyak és szándékok ('BDI') – nem ritkán egymás hiedelmeire, vágyaira és szándékaira vonatkozó – szövevényes rendszerét. Egy interpretáló információállapota tehát „világocskáknak" – az említett véges információtáraknak – egy olyan felcímkézett fastruktúrájaként definiálható, ami gyakorlatilag az ő elméjének – „belső világának" – a formális modelljeként szolgál, amely része a teljes univerzum „külső világot" is tartalmazó modelljének. Ami talán meglepő megközelítés, de semmi intuícióellenes nincs abban, hogy az emberi elméket is a világ(modell) részének tekintsük.

Ezek alapján a *szimultán rekurziós* definíciós technika kínálkozik a ℜeALIS mint episztemikus multiágens rendszer formális megfogalmazására: $\Re = \langle W_o, W, Dyn, Tru \rangle$, ahol az ágensek szerepét a világról – és azon belül (tipikusan!) egymás elméjének tartalmáról – folyamatosan információt gyűjtő interpretálók játsszák. W_o a külvilágot jelöli, ami egy idődimenziót is tartalmazó „teljes történelem", amire alapítva mind (igazságértékelő) statikus interpretációt definiálhatunk (Tru), mind (DRS-építő / a tudásgyarapodást felmérő) dinamikus interpretációt (Dyn), kölcsönhatásaikat is [KGR] megragadva. A W egy függvény, amelynek a W[i,t] értéke egy i interpretáló t időpillanatbeli *információállapotát* adja meg. A fentiek értelmében ez egyfelől a világ egy reprezentációját jelenti, másfelől nézve viszont a világ(modell) egy részletét; amennyiben ez utóbbi aspektust kívánjuk érzékeltetni, akkor *belső világként* utalhatunk a – világocskák felcímkézett fastruktúrájaként szerveződő – W[i,t] konstrukcióra. A modális kifejezések interpretációja a megfelelően felcímkézett világocskák tartalmára épül, a külvilágé (vagy bárminemű „lehetséges világé") helyett.

Ez nem kevesebbet jelent, mint hogy a ℜeALIS megközelítésében a szokásos értelemben vett *intenzionalitás* egyszerűen nem is létezik: a (teljes világmodell részét képező elmék leírásában szereplő) interpretálói világocskák hordozzák mindazt az információt (BDI, feltevések, álmok), ami másutt a lehetséges világokra van bízva. Úgy is fogalmazhatunk tehát, hogy a ℜeALIS rendszerében az interpretáció mindig *extenzionális*, csak a bázist képező modellzóna lehet többféle: a W_o külvilág vagy egy W[i,t] interpretálói belvilág valamely szektora, vagy – látjuk majd, mennyire gyakran! – a külvilág és több interpretáló különféle világocskáinak valamilyen kombinációja. Mindemögött az a hipotézis húzódik meg, hogy minden olyan (nyelvészeti) probléma, amelyről Montague-t követve [11] azt szokás gondolni, hogy megoldása a (végtelen) lehetséges világok kontrukciójáért kiált, megoldható a (véges) világocskákéra alapozva.

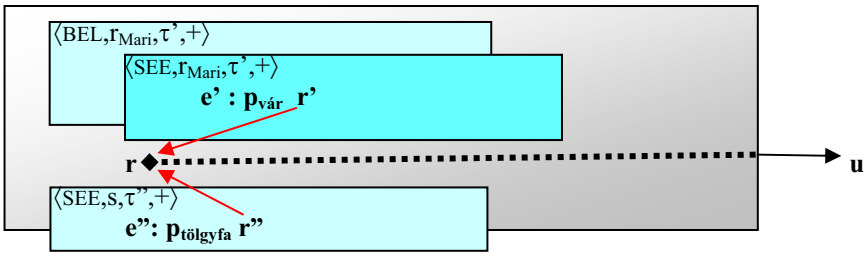
Szemléltetésül e cikkben álljon a *modális horgonyzás* – azaz az eltérő modális kontextusokon átívelő keresztreferencia – makacs problémája [20:243]. Az alábbi (1a) kétmondatos kisdiskurzus második mondatában azt nem tudják megmagyarázni, hogy *a várkastély* határozott kifejezés egyfelől modálisan alá van rendelve egy megelőző mondatban szereplő összetevőnek, másfelől viszont a második mondat a maga egészében nem áll modális alárendeltségben. Ez a jelenség azért jelent súlyos problémát a lehetséges világok *eliminációján* nyugvó szemantikai megközelítésben, mert az érintett mondat különböző részeinek interpretálása különböző eliminációt igényelne: *a várkastély* referenciáját Mari hiedelmei alapján kalkulálhatjuk ki, miközben a mondat állítmánya hamisnak bélyegzi az éppen e kalkuláció alapját jelentő előfeltevést. A ℜeALIS megközelítésében viszont, amik megfelelnek az „eltérő modális kontextusoknak", azok egyazon világmodell részét képezik – minthogy valamennyi interpretálói belvilág egyazon világmodellbe tartozik. Referenseik összehorgonyzásának en-

nél fogva elvi akadálya nincsen, csupán a referensek közötti „elérhetőség” megfelelő feltételrendszerét kell meghatározni.

Az alábbi (1e) reprezentáció például egy „ideális interpretáló” dinamikus interpretációjának a releváns részletét mutatja. Egy mondat (illetve diskurzus) *dinamikus interpretációja* az interpretálói információállapot kiterjesztéseként definiáltatik [1, 2.2.] [2, 4.2.]. Ami tulajdonképpen történik e „kiterjesztés” során, az nem más, mint hogy új szektorok épülnek ki az interpretálói információállapotban, köszönhetően a bemeneti performancia (morfémáról morfémára való) interpretálói feldolgozásának: a felcímkézett világocskák részben rendezett szövevénye új blokkokkal gyarapodik. Egy mondat *statikus interpretációja* (igazságértékelése) a külvilág bázisán vagy / *és* potenciálisan akár több interpretáló bizonyos világocskáinak a bázisán definiálandó. E struktúrák valamiféle egyesítését ($W_o+\Sigma W[\iota,\tau]$) kell a dinamikus interpretáció kimenetével ($W[i,t]$) összevetni, és meghatározni, hogy létesíthető-e közöttük elégséges *mintaillesztés*.

1. példa. MODÁLIS HORGONYZÁS – MINT AZ INTENZIONÁLIS AZONOSÍTÁS EXTRÉM ESETE

a. Mari úgy vélte, hogy a fák mögött egy várkastély van.
   A várkastély egy hatalmas tölgyfának bizonyult.
b. Általános világocskaindex: $\gamma=\langle\langle\mu_1,\tau_1,i_1,\pi_1\rangle, \langle\mu_2,\tau_2,i_2,\pi_2\rangle, ..., \langle\mu_k,\tau_k,i_k,\pi_k\rangle\rangle$
c. Az r’ világocskaindexe: $\gamma'=\langle \langle \text{BEL},r_{\text{Mary}},\tau',+\rangle, \langle \text{SEE},r_{\text{Mary}},\tau',+\rangle \rangle$
d. Az r” világocskaindexe: $\gamma''=\langle \langle \text{SEE},r_{\text{speaker}},\tau'',+\rangle \rangle$
e. A RELEVÁNS VILÁGOCSKÁK VIZUÁLIS MEGJELENÍTÉSE:



Az (1a) pontbeli első mondat egy r’ referens bevezetésével járul hozzá a diskurzusjelentéshez, amelyhez az az információ kapcsolódik, hogy „Mari várkastélynak vélte látni az r’ dolgot (a τ’ pillanatban)”. A második mondat egy állítást tesz valamiről, ami minden bizonnyal a beszélő vizuális megfigyelésén alapul.[2]

A DRT jól ismert „dobozstruktúrájának” [15] a ℜeALIS formalizmusában a világocskák *felcímkézett* részbenrendezése felel meg [1, 1.2.4.] [2, 3.2.4.]. Az (1e) reprezentáción fogom bemutatni e címkéket. Olyan rendezett négyesek, amelyek a következő tényezőket adják meg: a címke *modalitását* (pl. hiedelem / vágy / szándék / feltevés / megfigyelési mód), *közvetlen gazdáját*, *időpillanatát* és *polaritását* (pozitív

[2] Felvetődhet az olvasóban, hogy a pontos formula-feltöltése az olyan “dobozoknak”, mint az (1e) vagy a majdani (3d) pontbeliek, önkényes elemeket is tartalmaz, amelyek nem feltétlenül kompozicionális mondatelemzésből származnak. A tárgyalás jelenlegi szakaszában erre azt válaszolom, hogy az önkényesség a releváns dobozstruktúrát nem érinti. A cikk 5. szakaszában pedig visszatérünk majd a kérdésre egy tágabb perspektívából.

/ semleges / negatív). Az (1e) pontban a felső dobozpár például azt az információt hordozza, hogy egy τ' időpillanatban Mari (r_{Mari}) úgy hiszi (BEL), hogy egy e' eventualitást lát (SEE), melynek információtartalma: egy r' referens várkastély (a $p_{vár}$ predikátum a 'várnak lenni' állítást fejezi ki). Az alsó (egyetlen) doboz pedig azt az információt nyújtja, hogy a beszélő (s) vizuális észleli egy τ" (későbbi) pillanatban, miszerint valami – egy r" diskurzusszereplő – nem más, mint egy tölgyfa. Az (1c-d) formulák – a Bevezetésben előrevetített (1b) általános képletnek megfelelően – a világocskaindexeket közlik az r' és az r" referensek esetében. Az r' indexe azt fejezi ki, hogy egy Mari által τ' pillanatban látni vélt dologról van szó, míg az r" indexe egy, az adott beszélő által τ"-ben látott „valamire" utal.

Ez a formalizmus is megjeleníti tehát, hogy a lehetségesvilág-szemantikák számára problematikus modális horgonyzási jelenség miért is az: *a várkastély* kifejezést tartalmazó második mondat a beszélő perspektíváján nyugszik, és nem Marién; mégis sikeres a szóban forgó szinguláris határozott főnévi szerkezet indukálta antecedenskeresés. Vajon ez hogyan magyarázható a ℜeALIS rendszerében?

Az *unicitás* jelenti az antecedenskeresés sikerének zálogát: lennie kell egy olyan világocskának, amelyben egy referens egyedi az adott világocskában abban a tekintetben, hogy a szinguláris határozott főnévi szerkezet hordozta állítás csakis őrá igaz. Az alábbi (2a) kisdiskurzus második mondata például nem elégíti ki ezt az unicitási kritériumot – nem is jól formált a diskurzus, pedig modáliskontextusváltásról szó sincsen.

Az *elérhetőség* jelenti az antecedenskeresés sikerének másik tényezőjét. Az (1a) probléma precíz megoldása *akkommodációt* is igényel, egy referensnek ugyanis elérhetőnek kell lennie egy másik referens számára, amennyiben össze kívánjuk horgonyozni őket azonos referenciájuk kifejezése végett [15]. A ℜeALIS rendszerében az elérhetőség a lehető legkézenfekvőbb módon definiálható a világocskahierarchiára alapítva: r1 *elérhető* r2 számára, amennyiben r1 lejjebb helyezkedik el r2-höz képest a hierarchiát matematikailag definiáló részbenrendezés szerint [1, 2.2.3.6.] [2, 4.2.3.6.].

Milyen információ akkommodálását váltja ki a szinguláris határozott kifejezés az (1a) második mondatában? Azét, hogy a beszélő elfogadja, hogy „valóban van egy jókora entitás a fák mögött". Ennek ábrázolása úgy fest a diskurzus interpretálójának szemszögéből, hogy a diskurzus dinamikus interpretációjához tartozó relatív gyökérvilágocskába – ami a részbenrendezés szerint a legalsó világocska – bevezettetik egy r referens. Ami tehát mind r' („a Mari féle várkastély"), mind r" („a beszélő tölgyfája") számára elérhető; r' és r" tehát egyaránt odahorgonyozható az r referenshez, megragadva ezáltal koreferenciális viszonyukat, amelyet az ábrán a közös u jelöletük is mutat.

Gyanúsan egyszerűnek tűnhet persze az akkommodációhoz való folyamodás. Gondoljuk azonban meg: a beszélő számára kézenfekvő stratégiát jelent a lehető legkevesebbet „(ki)mondani", és ehelyett annyit rábízni a hallgatói információállapotra, amennyit csak lehetséges(-nek gondol a beszélő). Ahelyett, hogy a formális szemantikai elemzések során a szavak által expliciten ki nem fejezett információt ignoráljuk (mereven elhatárolódva leírásától), inkább arra kéne törekedni, hogy az információnak ezt az implicit rétegét is megragadjuk. A ℜeALIS „élethossziglani" megközelítése lehetővé teszi az implicit információ formális kezelését.

2. példa. UNICITÁS ÉS AKKOMMODÁCIÓ

a. Egy ódon városban megnéztünk két kastélyt. **A kastély* gyönyörű volt.
b. Péter tegnap megnősült. + c. / d.
c. *A pap* roppant harsányan beszélt. / d. ??*A kutya* nagyon hangosan ugatott.

A fenti (2b)+(2c/d) kétmondatos kisdiskurzus-variációk az akkommodáció iskolapéldájaként szolgálnak [14]. A mi kultúránkban egy *pap* „kitüntetett szereplője" *lehet* egy esküvőnek, míg ugyanez nem mondható el egy kutyáról. Mindazonáltal az sem zárható ki, hogy egy interpretáló a (2b)+(2d) diskurzust is kifogástalannak értékeli egy adott kontextusban: annyi szükséges, például, hogy ott legyen az információállapotában egy darabka tudás egy kutyáról, aki megkülönböztetett szerepet játszik Péter életében. Fontos hangsúlyozni, hogy sem a pap az egyik diskurzus-variációban, sem a kutya a másikban nem jelenik meg az esküvőt tartalmazó interpretálói információállapot valamiféle *logikai* következményrelációra való lezárásában; a kohézió tehát a jelen mondatok tartalma és az interpretáló által egykor – akár korlátlanul régen – elsajátított tartalmak között lép fel. Ha tehát számot akarunk adni a (2c) / (2d) folytatások eltérő megítéléséről, akkor aligha fordulhatunk a logikailag zárt lehetséges világokhoz; a ℜeALIS nyújtotta élethossziglani megközelítés ígér megoldást. A (2c/d)-beli szinguláris határozott kifejezés olyan eljárást indít el a dinamikus interpretáció során, ami az interpretálói információállapot kiterjesztését eredményezi a diskurzuskezdő (2b) mondat megértését követően; olyan kiterjesztését, amelyben lennie kell(ene) egy világocskának unicitást élvező pappal / kutyával. Az előbbi esetben a feladat végrehajtható, akkommodálva a mi nyugati kultúránkra jellemző esküvőre vonatkozó enciklopédikus információt; az utóbbi esetben pedig akkor, de csakis akkor hajtható végre, ha Péterre vonatkozó megfelelő személyközi információ akkommodálható.

Az alábbi (3a) pontban egy másik kontextust mutatok be, amelyben egy adott darabka információ („a felesége") úgy használható fel egy személy azonosítására, hogy közben a beszédpartnerek számára eltérő modális kontextusokhoz tartozik; mi több, mindketten tudván tudják, hogy hamis tartalmat hordoz. A ℜeALIS – ismét – olyan megoldást kínál, ami a releváns referensek bizonyos világocskákban való *unicitására* épül. A (3d)-ben a világocskablokkok azt ábrázolják, hogy *a felesége* szinguláris határozott kifejezés a vendég számára éppen ezt jelenti: „az egyetlen x személy a kontextusban, akire az igaz, hogy ő (a vendég) a portástól azt reméli, hogy az elhiszi, hogy az x illető a felesége, annak ellenére, hogy jól tudja az x-ről, hogy nem az"; míg a portás számára a következő meghatározás nyújtja az unicitást: „az egyetlen y személy a kontextusban, akire az igaz, hogy úgy gondolja, hogy a vendég azt reméli tőle, hogy elhiszi, hogy y a felesége neki (mármint a vendégnek), miközben persze tudja, hogy nem a felesége". A (3b-c) az imént meghivatkozott indexek formális leírását közli, hogy világos legyen, mi a mögöttes matematikai tartalma az olyan vizuális megjelenítéseknek, mint a (3d)-beli, amire aztán a kommunikáló interpretálók ℜeALIS-modelljének implementációját is fel lehet építeni.

3. példa. SIKERES REFERÁLÁS HAMIS INFORMÁCIÓ SEGÍTSÉGÉVEL:

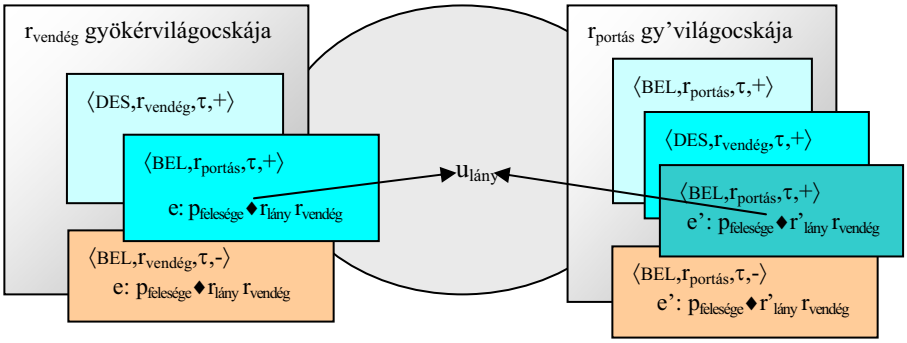
a. *Egy férfi érkezik egy motelbe egy lány társaságában, aki korántsem a felesége, egy olyan országban, ahol a portásnak a jogszabályok értelmében nem lenne szabad egy szobában elszállásolni őket. Az persze nem áll a portás anyagi érdekében, hogy ajtót mutasson nekik. Inkább mindketten úgy*

emlegetik a lányt, mintha a vendég felesége lenne, noha tisztában vannak vele, hogy ez az „előfeltevés" hamis; sőt, még azt is tudják, hogy a másik is tisztában van az igazsággal. A portás például ezt mondja: Remélem, ízleni fog a feleségének ez a pezsgő.

b. $\Gamma_e = \{ \langle\langle BEL,r_{vendég},\tau,-\rangle\rangle, \langle \langle DES,r_{vendég},\tau,+\rangle, \langle BEL,r_{portás},\tau,+\rangle \rangle \}$

c. $\Gamma_{e'}=\{\langle\langle BEL,r_{portás},\tau,-\rangle\rangle,\langle\langle BEL,r_{portás},\tau,+\rangle, \langle DES,r_{vendég},\tau,+\rangle, \langle BEL,r_{portás},\tau,+\rangle\rangle\}$

d. A RELEVÁNS VILÁGOCSKÁK VIZUÁLIS MEGJELENÍTÉSE:



3 Modális melléknevek, adverbiumok, kötőszavak, (segéd-) igék

Az alábbi (4) példában egy a (3)-hoz hasonló elemzéshez vezető jelenséget szemléltetek. Az *állítólagos* melléknévről van szó, amit Kiefer [17:188] *szabálytalanként* sorol be, a (4b-c), (4d-e) tulajdonságai alapján, összevetve a szabályos *öreg* melléknévvel.

Megközelítésünkben kézenfekvően adódik a szabályos és a szabálytalan melléknevek közötti különbség: az előbbiek egy *predikátummal* járulnak hozzá a diskurzusreprezentációhoz (mint a $p_{tölgyfa}$ vagy a $p_{felesége}$ a 2. szakasz elemzéseiben), míg az *állítólagos* a világocskacímke modális összetevőjét szabja meg. A beszélő olyan információval utal egy szereplőre, amelynek igazsága mellett nem kötelezi el magát (4g), miközben ugyanazon mondat állítmányának tartalma mellett igen (4f). Így utal a szereplőre: „egy olyan személy, akiről legjobb tudomása szerint van, aki (r*) azt gondolja, hogy kém" (4h). A (4b) „anomáliáról" – miszerint az „állítólagos P"-ből nem feltétlenül következik a P – a (4h) reprezentáció számot ad, hiszen deklaráltan nincs elkötelezve a beszélő a P igazsága mellett (4g). Az állítmányi szerep visszautasítása pedig (4d) abból adódik, hogy az *állítólagos* hozadéka nem egy $p_{állítólagos}$ predikátum.

4. példa. *ÁLLÍTÓLAGOS*: EGY SZABÁLYTALAN (AVAGY MODÁLIS) MELLÉKNÉV

a. Tegnap Mari találkozott *egy állítólagos kémmel*.

b. Egy *állítólagos kém* az kém. → *nem (feltétlenül) igaz*

c. Egy *öreg kém* az kém. → *feltétlenül igaz*

d. *Pál *állítólagos*. → *rosszul formált.* e. Pál *öreg*. → *jól formált*

f. $\Gamma_{e:találkozik} = \{ \langle\langle BEL,r_{speaker},\tau,+\rangle\rangle \}$

g. $\Gamma_{s:kém}=\{ \langle\langle BEL,r_{speaker},\tau,0\rangle\rangle,$

h. $\langle\langle BEL,r_{speaker},\tau,+\rangle,\langle BEL,r^*,\tau,+\rangle\rangle \}$

A segédigék hasonló modális hatásmechanizmusára német példákat mutatok be. Az (5a-b) mondatokban egyaránt megvan az a jelentésfaktor, hogy a beszélő elhatárolódik magától az s állapotról szóló állítástól, miszerint Péter beteg volt ('$\langle\langle \mathrm{BEL}, r_{speaker}, \tau, 0\rangle\rangle$'; ld. (5c-d)). Az (5c-d)-ben közölt több négyesből álló formulák azt a beszélői vélekedést fejezik ki, hogy a beszélő másnak (r*) tulajdonítja az állítást (5c), illetve úgy gondolja, hogy az alany szándéka elhitetni másokkal (r*) a betegség fennállását (5d).

5. példa. A NÉMET *SOLL* ÉS *WILL*: MODÁLIS SEGÉDIGÉK

a-b. Peter soll / will krank gewesen sein. 'Peter beteg volt.' *(de ld. (5c-d))*
Peter *soll / will* beteg van.PERF van.INF
c. $\Gamma_{s:beteg/a.} = \{\langle\langle \mathrm{BEL}, r_{speaker}, \tau, 0\rangle\rangle, \langle\langle \mathrm{BEL}, r_{speaker}, \tau, +\rangle, \langle \mathrm{BEL}, r^*, \tau, +\rangle\rangle\}$
d. $\Gamma_{s:beteg/b.} = \{\langle\langle \mathrm{BEL}, r_{speaker}, \tau, 0\rangle\rangle, \langle\langle \mathrm{BEL}, r_{speaker}, \tau, +\rangle, \langle \mathrm{INT}, r_{Peter}, \tau, +\rangle, \langle \mathrm{BEL}, r^*, \tau, +\rangle\rangle\}$

A (6) példa képletei egy olyan ideális beszélő információállapotának (egyszerűsített) modelljét állítják fel, aki egy *valószínűleg*-gel módosított tartalmú mondatot dolgozott fel. Az igazságértékelés szempontjából az első érdekesség az, hogy hamis állításról akkor sem beszélhetünk, ha az s állapotról („Mari otthon van") szóló állítás maga a külvilág alapján hamis. A (6a) mondat tehát nem a külvilágról ad információt (6b), hanem az „ideális beszélő" információállapotáról, mondjuk a grice-i értelemben [13], amire az SDRT is alapít [9]. A kérdésre majd az 5. szakaszban visszatérünk. Az elemzés a világocskacímke modális összetevőjének finomabb értékskáláját alkalmazza: a 'BEL_{great}' a hiedelem alacsonyabb fokozatára utal, mint a biztos tudásra utaló 'BEL_{MAX}'. A (6c) formulái tehát ezt közlik: a beszélő (s) *valószínűsíti*, hogy Mari otthon van, és szándékában áll a hallgatóját (i) is erről a valószínűségről meggyőzni. A beszélő azt is sugallja a (6a) közléssel, hogy nincs közvetlen érzéki tapasztalata Mari otthon létével vagy ennek ellentétével kapcsolatban, és hallgatójáról is ezt gondolja (6d), illetve azt, hogy közlésével tudott valami újat mondani a hallgatónak (6e), vagyis az nincs Mari otthon létével kapcsolatos biztos tudás birtokában.

6. példa. *VALÓSZÍNŰLEG* : EGY MODÁLIS ADVERBIUM

a. Mari valószínűleg otthon van.
b. Irreleváns az interpretációnál, hogy s ("M. otthon van") fennáll-e W_o-ban.
c. $\Gamma_{s:otthon\text{-}van} = \{ \langle\langle \mathrm{BEL}_{great}, s, \tau, +\rangle\rangle, \langle\langle \mathrm{INT}, s, \tau, +\rangle, \langle \mathrm{BEL}_{great}, i, \tau', +\rangle\rangle,$
d. $\langle\langle \mathrm{SEE}, s, \tau, 0\rangle\rangle, \langle\langle \mathrm{BEL}_{great}, s, \tau, +\rangle, \langle \mathrm{SEE}, i, \tau, 0\rangle\rangle,$
e. $\langle\langle \mathrm{BEL}_{great}, s, \tau, +\rangle, \langle \mathrm{BEL}_{MAX}, i, \tau, 0\rangle\rangle \}$

A kötőszókban is rejlik intenzionalitás; amit a ℜeALIS eszköztárával meg tudunk ragadni formálisan, és a világocskaindexekre alapozva implementálhatunk is. A (7a) válaszból például az is kiderül, hogy a beszélőnek nincs biztos tudása sem az s' állapotra nézve („M. Delhiben van."), sem az s"-re nézve („M. Bombayben van.") – vagy meg akarja téveszteni a hallgatóját (7b), azaz nem viselkedik „ideális beszélőként". Jobban belegondolva azt is megkérdőjelezhetjük, hogy a klasszikus logika által javasolt s=s'∨s" információról lehet-e *biztos* tudása a beszélőnek ('BEL_{MAX}'), miközben a diszjunkciónak sem az s' tagjáról, sem az s" tagjáról nincsen biztos tudása. Ezért a (7c) formulában olyan tudásmodellt állítottam fel, amelyben a *vagy* hatása egy

’BEL_{amax}’ hiedelemérték választásában mutatkozik meg: ez igen erős, de mégsem teljes és közvetlen bizonyosságra utal.

7. példa. INTENZIONALITÁS A KÖTŐSZÓKBAN

a. (Hol van Mari?) Delhiben vagy Bombayben.
b. $\Gamma_{\text{s':Delhiben}} = \{\langle\langle \text{BEL}_{\text{MAX}}, s, \tau, 0\rangle\rangle\}$; $\Gamma_{\text{s":Bombayben}} = \{\langle\langle \text{BEL}_{\text{MAX}}, s, \tau, 0\rangle\rangle\}$
c. $\Gamma_{\text{s:[s' or s"]}} = \{\langle\langle \text{BEL}_{\text{amax}}, s, \tau, +\rangle\rangle\}$
d. Am Montag wusste ich nicht, *dass/ob* du am Sonntag in der Kneipe gewesen warst.
-On hétfő tud.MÚLT.E1 ÉN nem, $\text{hogy}_{\text{dass/ob}}$ te -On vasárnap -bAn a.DAT kocsma van.PERF van.MÚLT.E2
Hétfőn nem tudtam, hogy vasárnap a kocsmában voltál / voltál-e.
e. $\Gamma_{\text{s[dass]:kocsmában}} = \{\ \langle\langle \text{BEL}_{\text{MAX}}, s, \tau', 0\rangle\rangle, \langle\langle \text{BEL}_{\text{MAX}}, s, \tau, +\rangle\rangle\ \}$
f. $\Gamma_{\text{s[ob]: kocsmában}} = \{\ \langle\langle \text{BEL}_{\text{MAX}}, s, \tau', 0\rangle\rangle\ \}$

A fenti német példapár (7d) a *hogy*-nak megfelelő alárendelő kötőszók közötti választásról szól, illetve ennek egyetlen érdekes mozzanatáról: míg látszólag csupán egy korábbi információállapotról tájékoztat a mondat, amelyben az s állapotról szóló információ egy semleges hiedelemvilágocska-szektorban van (7e-f), az egyik kötőszóval a beszélő elárulja, hogy egy későbbi információállapotában az s már pozitív tudásként van jelen (7e).

A szakasz utolsó példájában (8) egy olyan magyar ige szerepel, amely az interpretálói információállapotban rendkívül gazdag indexhalmazzal címkéz fel egy s eventualitást, ami egyébként (a megítélésem szerint preferált értelmezés szerint) a külvilágra vetítve hamis (8b). Egész kis dráma bontakozik ki az s információ „vándorlását” nyomon követve világocskáról világocskára, az indexhalmazt áttekintve (8c-f). Egy τ’ pillanatban Mari nem gondolta úgy, hogy Pál nős (s), egy későbbi τ pillanatban viszont már így gondolta (8c). A változást egy (nem feltétlenül ismert) r* „intrikus” idézte elő, aki tudja, hogy s hamis, és úgy gondolja, hogy Mari sem hiszi igaznak (8d). Arra vágyik (’DES’) viszont, hogy Mari úgy higgye, hogy s igaz, és ezért tenni is akar (8e); az INT címke a szándékra utal, amellyel a címke közvetlen gazdája (r*) saját belvilágának komplementumát a belvilágában megfogalmazódó vágyaihoz akarja igazítani – mint láttuk (8c), sikeresen. Mi több (8f), Mariról azt sugallja a (8a) mondat, hogy úgy hiszi, az intrikus is nősnek gondolja Pált, és sejtelme sincs arról, hogy tudatosan be akarta csapni őt.

8. példa. *BEVESZ* : EGY GAZDAG INTENZIONÁLIS MINTÁZATÚ IGE

a. Mari bevette, hogy Pál nős.
b. Az s állapot („Pál nős.”) nem áll fenn W_o-ban.
c. $\Gamma_{\text{s:nős}} = \{\ \langle\langle \text{BEL}, r_M, \tau', -\rangle\rangle$ or $\langle\langle \text{BEL}, r_M, \tau', 0\rangle\rangle, \langle\langle \text{BEL}, r_M, \tau, +\rangle\rangle,$
d. $\langle\langle \text{BEL}, r^*, \tau, -\rangle\rangle, \langle\langle \text{BEL}, r^*, \tau', -\rangle, \langle \text{BEL}, r_M, \tau', +\rangle\ \rangle,$
e. $\langle\langle \text{DES}, r^*, \tau', +\rangle, \langle \text{BEL}, r_M, \tau, +\rangle\ \rangle, \langle\langle \text{INT}, r^*, \tau', +\rangle, \langle \text{BEL}, r_M, \tau, +\rangle\ \rangle,$
f. $\langle\langle \text{BEL}, r_M, \tau, +\rangle, \langle \text{BEL}, r^*, \tau, +\rangle\rangle, \langle\langle \text{BEL}, r_M, \tau, 0\rangle, \langle \text{INT}, r^*, \tau, +\rangle, \langle \text{BEL}, r_M, \tau, +\rangle\rangle\}$

4 A mód, a modalitás és az aspektus intenzionalitása a magyarban

Ízelítőt szeretnék adni a magyar mód- és modalitástoldalékok intenzionális modellezésére irányuló alprojektünk eredményeiből. Az alábbi táblázat néhány múlt idejű kombináció (egyszerűsített) elemzését mutatja be.

Minden kombináció (legalább) kétértelmű. A beszélő (s) vagy valaki más (r*) hiedelmeit, vágyait és/vagy szándékait fejezi ki (BEL, DES, INT), egy modalitáson belül finomabb intenzitási skálát is megkívánva (MAX > amax > great > med). A BEL_{MAX} címke például a teljes bizonyosságra utal. Az ⟨INT,r*,π⟩ címketípus az r* személy utasítását (π=+; ld. c., g.), tiltását (π=-) vagy engedélyét (π=0; ld. a., e.) jelzi, a címke polaritási összetevőjétől függően. A BEL-PART modális tényező egy e eventualitás „részleges tudásának" a megragadására hivatott (l. a b., d. episztemikus olvasatokat); amin nem bizonytalan tudást értek, hanem olyan információdarabok ismeretét, amelyek a ℜeALIS élethossziglani interpretálói belső világaiban az e tudásdarabhoz asszociálódnak mintegy „tanúskodva" az e fennállása mellett. A *hazamehetett* alaknál például a táblázatban ez a két intenzionális elemzés szerepel: a. „a beszélő szerint valaki hazament, mert engedélyt kapott erre" (nyilván az is vizsgálandó, hogy r* az engedélyezéshez megfelelő pozícióban van-e); b. „meglehetősen valószínű, hogy valaki hazament, mert vannak emellett tanúskodó jelek (nincs ott az irodájában, sőt a kabátja és az esernyője sincs ott, elmúlt már 18^{10}, stb.)".

↓Mód Modalitás→	*hazamegy + -(Vt)t*		*hazamegy + -(Vt)t + vol- + -nA*	
-hAt	*hazame·het·ett*		*hazame·het·ett vol·na*	
	a. ⟨INT,r*,0⟩ ⟨BEL_{MAX},s,+⟩	b. ⟨BEL_{med},s,+⟩ ⟨$BEL\text{-}PART_{great}$,s,+⟩	e. ⟨INT,r*,0⟩ ⟨BEL_{MAX},s,-⟩	f. ⟨DES_{great},s,+⟩ ⟨BEL_{MAX},s,-⟩
kell	*haza kell·ett ↙men·ni(e) / menni↘*		*haza kell·ett vol·na ↙men·ni(e) / menni↘*	
	c. ⟨INT_{MAX},r*,+⟩ ⟨BEL_{MAX},s,+⟩	d. ⟨BEL_{amax},s,+⟩ ⟨$BEL\text{-}PART_{MAX}$,s,+⟩	g. ⟨INT_{MAX},r*,+⟩ ⟨BEL_{MAX},s,-⟩	h. ⟨DES_{amax},s,+⟩ ⟨BEL_{MAX},s,-⟩

1. ábra. A magyar mód és modalitás múlt idejű alakjainak modális elemzése.

Hasonlóképpen modellezhetjük az aspektusok intenzionális karakterét. Vegyük például górcső alá a (9a)-beli progresszív válaszmondatot! A progresszivitásból adódóan fellép egy Imperfektív Paradoxonként emlegetett jelenség [10:147]: nem dönthető el a mondat igazságértéke pusztán a külvilági tények alapján. Csak a szóban forgó nap 18^{10} előtti időszaka tesztelendő externálisan, vagyis a hazautazási esemény kumulatív szakaszának egy kezdőintervalluma (9c). A teljes esemény lefolyásáról a beszélő nem garantál biztos tudást (9b), csupán erős valószínűséget sugall (9b). A 18^{10} utáni időszakra vonatkozóan tehát „internális" információ áll rendelkezésre: egyrészt az említett beszélői valószínűsítés, ami a „dolgok szokásos rendjének" ismeretéből fakadhat (9b), másrészt (legalábbis preferáltnak hat egy ilyen értelmezés) az alanynak tulajdonított szándék. Úgy látom egyébként, hogy a (9b-d) intenzionális karakter egy az egyben a jövő idő jellemzésére is alkalmas: a (9e) mondatot is úgy értelmezzük (egyik jelentésében), hogy az eseményről biztos tudás persze nincs, de valószínű, hogy lefolyik (9b), mert a beszélő rendelkezésére állnak erről tanúskodó jelek (9c), és

preferáltan az alany szándéka is megvan (9d). A progresszív tehát végső soron nem más, mint „jövő a múltban”.

9. példa. A MAGYAR PROGRESSZÍV ASPEKTUS ÉS A JÖVŐ IDŐ

a. (Mit csinált Péter 2003. május 4-én 18^{10}-kor?) Utazott (éppen) haza.
b. $\Gamma_{e:hazautazik} = \{ \langle\langle BEL_{MAX}, s,\tau,0\rangle\rangle, \langle\langle BEL_{great}, s,\tau,+\rangle\rangle,$
c. $\langle\langle BEL\text{-}PART_{MAX}, s,\tau,+\rangle\rangle$
d. $\langle\langle INT, r_{Peter},\tau,+\rangle\rangle \}$
e. Péter haza *fog* utazni.

5 Az információ beágyazása az interpretálói információállapotba

Az előző két szakaszban különféle lexikai egységek intenzionális karakterének a hatását tárgyaltam a dinamikus interpretáció kimenetére. Vannak azonban pragmatikai hatások is.

Kézenfekvő például, hogy az *irónia* egyszerűen megfordítja bizonyos világocskák polaritási címkéjét (π=-). Más esetben megsejthető, hogy a beszélő blöfföl; ilyenkor a megfelelő világocska polaritása: π=0. Nem nyertünk volna hát semmi információt? Dehogynem! Csak nem a külvilágról, hanem a beszélő sanda szándékáról... – hogy például elhiggyünk valamit, ami talán nem is úgy van; vagy hogy elhitesse velünk, hogy ő tud valamit.

Elméleti hátterünk élethossziglani jellegéből adódóan kézenfekvő lehetőség kínálkozik a *megbízhatóságáról* modellt alkotni akár az információnak, akár az interpretáló ágenseknek. Össze kell vetni egy információdarabra nézve különböző interpretálók intenzionális mintázatait, illetve rögzített interpretálókat tekintve azok intenzionális mintázatait különféle eventualitások vonatkozásában. A legegyszerűbb alkalmazandó elv például az, hogy megbízhatóbb az az információ, ami független forrásokból ugyanabban a formában érkezik, és ez az egybeesés a források megbízhatóságát is növeli. Ilyen elveknek kell irányítaniuk az információ áramlását az ideális interpretáló részbenrendezett világocskahálózatában, illetve annak meghatározását, hogy az információforrásként szolgáló ágensek milyen módon térnek el az „ideális beszélő” default képétől, ami a lexikai intenzionális hatások tárgyalása során (3-4.) mindig a kiindulópontunk volt.

Mivel a ℜeALIS a kommunikációban álló interpretálók „élethossziglani” és „kölcsönös” multiágens rendszere, különböző kérdéstípusok intenzionális modellezésére is készen kínálkozik. Az alábbi (10a-e) pontokban a *kiegészítendő kérdésekre* vonatkozóan vázolok fel egy világocskaindexekre épülő elemzést. Az r* referens Pál (adott időpontbeli) feleségeként határozódik meg a (10b)-ben. A szintén r*-ról szóló e* eventualitás pedig a (10c) pontban a (10d)-ben meghatározott világocskamintázatban jelenik meg, lehorgonyzatlan (azonosítatlan) p* predikátummal. A kérdő formából adódóan az e* olyan, hogy (10d) a beszélő nem tudja eldönteni az igazságértékét, de szándékában áll elérni ezt; valószínűsíti továbbá, hogy a hallgató birtokában van a releváns tudásnak, és reméli, hogy hajlandó is lesz megosztani vele. A (10c)-beli p* „lehorgonyzatlanságának” jelentősége a következő: a formális pragmatikai kezdeményezések [9] sarokkövének tekinthető „Maximalizáld a diskurzuskoherenciát!” elv

arra fogja késztetni a hallgatót, hogy a p* predikátumreferenst a lehető leghatékonyabban horgonyozza le. A válasz hatékonyságát nyilván a kérdező információállapotának növekményére alapozva határozhatjuk meg. A (10e.1) válasz például nyilván a legkedvezőtlenebb, mert aligha nyújt információnövekményt a kérdező meglévő enciklopédikus tudásához képest. A 3. válasz pedig hatékonyabb a 2. válasznál, akkor – és csakis akkor –, ha a kérdező ismeri a megnevezett személyt; egy azonosított entitás referensének a megtalálása ugyanis elérhetővé teszi mindazt a roppant információtömeget, ami e referenshez kapcsolódott „élethossziglan".

10. példa. A KÉRDÉS KÉRDÉSE

a. Ki volt Pál felesége akkoriban?
b. e: $p_{felesége}$ t r* $r_{Pál}$
c. e*: p* t* r*
d. Γ_{e*} = { $\langle\langle BEL_{MAX},s,\tau,0\rangle\rangle$, $\langle\langle INT,s,\tau,+\rangle,\langle BEL_{MAX},s,\tau,+\rangle\rangle$, $\langle\langle BEL_{great},s,\tau,+\rangle,\langle BEL_{MAX},h,\tau,+\rangle\rangle$, $\langle\langle DES,s,\tau,+\rangle,\langle INT,h,\tau',+\rangle,\langle BEL_{MAX},s,\tau',+\rangle\rangle$}
e. 1. „Egy nő."
   2. „Egy pincérnő a kedvenc indiai éttermünkből."
   3. „Az elbűvölő Shabana Singh."
f. Ki *is* volt Pál felesége akkoriban?
g. Γ^{+}_{e*} = { $\langle\langle BEL_{great},s,\tau,+\rangle,\langle BEL_{MAX},s,\tau'',+\rangle\rangle$,
   $\langle\langle BEL_{amax},s,\tau,+\rangle,\langle BEL_{MAX},h,\tau,+\rangle\rangle$,
   $\langle\langle BEL_{great},s,\tau,+\rangle, \langle BEL_{amax},h,\tau'',+\rangle, \langle BEL_{MAX},s,\tau'',+\rangle\rangle$ }
h. Tunteeko Pekka Marja*n* / Marja*a*? 'Péter ismeri Marit?'
   ismer-E3-Q Péter Mari-***ACC*** / Mari-***PART*** (e: p_{ismer} t $r_{Péter}$ r_{Mari})
i. Γ_{e} = $\langle\langle BEL_{MAX},s,\tau,0\rangle\rangle$, $\langle\langle BEL_{great},s,\tau,+/-\rangle\rangle$, $\langle\langle INT,s,\tau,+\rangle,\langle BEL_{MAX},s,\tau,+/-\rangle\rangle$
j. ... És PÉTERT *is* hívtuk meg!
k. $\Gamma_{e:[Péter\ az...]}$ = { $\langle\langle BEL_{MAX},s,\tau,+\rangle\rangle$, $\langle\langle INT_{MAX},s,\tau',+\rangle$,... }

A fenti (10f) példa újabb csodálatos megnyilvánulása egy piciny nyelvi elem sokrétű intenzionális hatásának. Lelkesedésem tárgya ezúttal az *is* szócska – diskurzuspartikulaszerű szerepben. A (10g)-ben foglaltakat teszi hozzá a kérdőszó szemantikai-pragmatikai kontribúciójához (10d): a beszélő biztos benne, hogy egykor birtokában állott az e* tudás ($\tau''<\tau$), és majdnem biztosra veszi, hogy a hallgatója most is tudja; preferáltnak érzem továbbá azt az értelmezést, hogy a kérdező úgy véli, hogy hallgatója tudja róla, hogy egykor birtokában állott neki is az e* információ (az együtt töltött „régi szép időkben"...).

Az eldöntendő kérdés annak jelzése, hogy a beszélő sem abban nem biztos, hogy egy bizonyos e eventualitás igaz, sem abban, hogy hamis, és szeretne biztosat tudni. A (10h) finn példa annyiban különleges, hogy a tárgy esetjelölése (Akkuzatívusz / Partitívusz) arról is információt ad (10i), hogy a kérdező pozitív vagy negatív választ vár-e (el).

Az *is* szócska egy másik sajátos jelentéshozadékával zárom az intenzionális mintázatok elemzését. A fenti (10j) fókuszos mondat csak olyan diskurzusban hangozhat el, ahol előtte ugyanaz a tartalom ugyanolyan fókuszkonstrukcióval mint szándék (10k) fogalmazódott meg.

Hivatkozások

1. Alberti, G.: ℜeALIS: An Interpretation System which is Reciprocal and Lifelong. Workshop 'Focus on Discourse and Context-Dependence' (16.09.2009, 13.30-14.30 UvA, Amsterdam Center for Language and Comm.). http://www.hum.uva.nl/aclc/ events.cfm/C2B8E596-1321-B0BE-6825998CFA642DB2, http://lingua.btk.pte.hu/realispapers (2009)
2. Alberti, G.: ℜeALIS: Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
3. Alberti, G.: ℜeALIS, avagy a szintaxis dekompozíciója. Általános Nyelvészeti Tanulmányok Vol. 23. (szerk. Bartos H.) (2011) 51–98
4. Alberti, G., Károly, M., Kleiber, J.: The ℜeALIS Model of Human Interpreters and Its Application in Computational Linguistics. In: Cordeiro, J., Virvou, M., Shiskov, B. (eds.): Proceedings of ICSOFT 2010, 5th International Conference on Software and Data Technologies, Athens, Greece. Vol. 2. SciTePress Portugal (2010) 468–474.
5. Alberti, G., Károly, M., Kleiber, J.: From Sentences to Scope Relations and Backward. In: Sharp, B., Zock, M. (eds.): Natural Language Processing and Cognitive Science. Proceedings of NLPCS 2010. SciTePress, Funchal, Madeira, Portugália (2010) 100–111
6. Alberti G., Kilián I.: Vonzatkeretlisták helyett polaritásos hatáslánccsaládok – avagy a ℜeALIS σ függvénye. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2010. SzTE Informatikai Tanszékcsoport. http://www.inf.u-szeged.hu/mszny2010 (2010) 113–126
7. Alberti, G.: The Grammar of ReALIS and the Implementation of its Dynamic Interpretation. Informatica Vol. 34, No.1 (2010) 103–110
8. Alberti, G., Kleiber, J.: Where are Possible Worlds? (Arguments for ℜeALIS). SinFonIJa4, Budapest (2011)
9. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge Univ. Press (2003)
10. Dowty, D. R.: Word Meaning and Montague Grammar. D. Reidel Publishing Company, Dordrecht (1979)
11. Dowty, D. R., Wall, R. E., Peters, S.: Introduction to Montague Semantics. D. Reidel Publishing Company, Dordrecht (1981)
12. Farkas, J.: A produktív finn képzések. Alberti, G. (szerk.): Vonzatok vonzásában. PTE BTK Nyelvtudományi Doktori Iskola (2012)
13. Grice, H. P.: Logic and Conversation In: Cole, P., Morgan, J.L. (eds.): Syntax and Semantics Vol. 3: Speech Acts. Academic Press, New York (1975) 41–58
14. Kálmán, L.: Deferred Information: The Semantics of Commitment. Kálmán, L., Pólos, L. (eds.): Papers from the Second Symposium on Logic and Language. Akadémiai, Budapest (1990) 125–157
15. Kamp, H., van Genabith, J., Reyle, U.: Discourse Representation Theory. In: Gabbay, D., Guenthner, F. (eds.): Handbook of Philosophical Logic, Vol. 15.Springer-Verlag, Berlin (2011) 125–394.
16. Károly, M.: Interpretáció és modalitás – avagy a ℜeALIS λ-függvényének implementációja felé. In: Tanács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2011. SzTE Informatikai Tanszékcsoport. http://www.inf.u-szeged.hu/mszny2011 (2011) 284–296
17. Kiefer, F.: Jelentéselmélet. Corvina, Budapest (2000)
18. Kilián, I.: Tárgymodell változatok a ℜeALIS nyelvi elemzéshez. In: Tanács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2011. SzTE Informatikai Tanszékcsoport. http://www.inf.u-szeged.hu/mszny2011 (2011) 276–283
19. Pollard, C.: Hyperintensions. ESSLLI 2007, http://www.cs.tcd.ie/esslli2007 (2007)
20. Roberts, C.: Anaphora in Intensional Contexts. In: Lappin, Sh. (ed.): The Handbook of Contemporary Semantic Theory. Blackwell, Oxford (1996) 215–246

Tárgymodellváltozatok a ReALIS nyelvi elemzéshez

Kilián Imre

ReALIS ESzNyK / PTE TTK Informatika Tanszék
7624 Pécs, Ifjúság útja 6.
kilian@gamma.ttk.pte.hu

Kivonat: Forrásnyelvek célnyelvre átalakítását (pl. fordítóprogramokban) a két metamodell közötti átalakítási szabályrendszerként értelmezhetjük. A ReALIS elmélet (http://lingua.btk.pte.hu/realispapers) esetében ez a ReALLan forrásnyelv, és a választott Prolog nyelvű tárgymodellváltozatok közötti leképezés megadását jelenti. Szövegfeldolgozási célokra Prolog nyelven általában a relációs tárgymodellt alkalmazzák, mert ez a nyelv jellegéből fakadóan a szövegnyelvtani szerkezet relációt nemcsak az adott (felismerési) irányban, hanem fordítva, szöveggenerálási irányban is képes kiszámítani. Hatékonysági okokból azonban még további tárgymodellváltozatokat is érdemes számításba venni. A következtetéses tárgymodell esetében az elemzett szöveg szavai tényállításokká, a ReALIS lexikonban ábrázolt nyelvtani információk szabályokká képződnek le, amelyek egy célállításból meghíva előállítják az elemzett szöveg nyelvtani szerkezetét. A Prolog logikán túli eszközeinek használatával a deduktívan megvalósított elemzési feladat abduktívan megvalósított szöveggenerálássá alakítható. A ReALIS lexikonban tárolt nyelvtanának, és az elemzési folyamat aszinkron jellegének a Prolog visszafelé következtető stratégiája helyett azonban jobban megfelel egy előre haladó modell. A cikkben tárgyalt Contralog modell a Prolog előre haladó kiterjesztése, amellyel magyar mondatok ReALIS elmélet szerinti elemzését mutatjuk be.

1 ReALLan: a ReALIS nyelvleíró nyelve

Természetes nyelvi megvalósítások egyik sarokköve a nyelvi információk leírási módja. Ezt célszerűen valamilyen nyelvleíró formális nyelven tehetjük meg. Ha csupán a szöveges kinézetet megadó valamelyik *nyelvtani formalizmusra* (pl. BNF) szorítkozunk, akkor a kinézet oltárán feláldozzuk az adatszerkezetet és annak az értelmezését. Objektumorientált rendszerekben a formális nyelv *metamodelljét* pl. UML-ben adjuk meg, amely a nyelv elemeit grafikus módon rögzíti, és amelyhez az *érvényességi szabályokat* az OCL megszorítás-leíró nyelvvel adhatjuk meg. A mi esetünkben a Prolog megvalósítás miatt a ReALLan a Prolog egy résznyelve, vagyis az alapnyelvtan egyfajta alkalmazói megszorítása. Mivel a Prolog típustalan, ezért erre a célra egy Prolog *típusleíró nyelvkiegészítést* (ReALType) valósítottunk meg.

A ReALLan nyelvleíró nyelven a rendszer teljes lexikalizmusa miatt a lexikonbéli elemekhez rendelhető nyelvtani információk rögzítésének szabályait lehet megadni. A nyelv alapvetően *jegyszerkezetes*, egy jegyszerkezet mátrix megadása alapvetően

Prolog listában, `JEGY:ÉRTÉK` párokkal lehetséges. Ehhez az általános leíráshoz képest a következő *bővítéseket* és *nyelvtani könnyítéseket* (syntactic sugar) tesszük lehetővé:

- Ha egy jegy értéke szintén összetett, és a jegygeometriában megadott összes jegyet tartalmazza, akkor a jegynevek megadása nem kötelező, és a Prolog listakifejezés helyett kerek zárójelekkel *teljes Prolog kifejezés* is megadható. Pl. `agr:[pers:1,nr:sing]` helyett `agr(1,nr)` is írható.
- Azonos értékek (KIG összefutó élek) jelölésére (fordításidejű egyesítés) *Prolog változókat*, és a `=/2` funktort használjuk. Pl: `PRED=desire(SUBJ,OBJ).`
- A fordításidejű egyesítés mellett a `:=/2` funktorral a *jobboldal kiértékelésére és futásidejű egyesítésre* is lehetőséget adunk. Pl. az

```
RDES1:=[argn(ord(-7,nei),cat(+2,noun),
        case(+2,nom)),
        argd(cat(+7,gqd))]
```

…kifejezés futásidőben egyesíti a Prolog változót, mint referenst a szövegben megfelelő helyzetben talált alanyesetű, főnévvel úgy, hogy a szerkezet általánosított kvantordeterminánsszerepben van.

2 Tárgymodell: Horn-klózok

A tárgymodellek leírásához érdemes rögtön az átalakítási szabályrendszert is hozzákapcsolni. Ha a szigorú objektumorientáltság elvei mellett maradunk, akkor ez úgy történik, hogy a forrás- és a célkörnyezet metamodelljét kapcsolatnyalábbal kapcsoljuk össze, melyet az átalakítások szabályait rögzítő OCL-megszorításokkal látunk el. Bár most nem kívánjuk az UML modelleket bemutatni, a metamodellek és az átalakító relációk fogalma a modellező eszköztől független, és a Prologhoz kötődő környezetben is alkalmazható úgy, hogy a forrás- és célkörnyezet fogalmait, valamint a közöttük megvalósítandó *átalakítási relációt* adjuk meg.

A célkörnyezet a *Horn-klózok osztálya*. Ez az elsőrendű logika azon részosztálya, amelyekben a klózok következményoldalán több literál diszjunkciója helyett legfeljebb *egyetlen literál* állhat.

$p_1; p_2;\ldots p_k :- n_1, n_2,\ldots, n_1.$

A részosztály azért figyelemre méltó, mert a Prolog programozási nyelv is ezt használja úgy, hogy a következtetéseket a háttérben egy rögzített stratégiájú, rezolúciós tételbizonyító végzi. A visszafelé haladó, lineáris-, egység- és alaprezolúciós stratégia tételbizonyításra gyengécskének tűnik, de cserébe a nyelv nem logikai eszközeivel meglehetősen rugalmas és magasszintű működés írható elő.

A ℜeALIS céljaira a Horn-klózokra alapuló relációs és következtetéses tárgymodellt is, ez utóbbira pedig a Prolog eredeti, *visszafelé haladó*, ill. a Horn-klózok újonnan kifejlesztett, *előre haladó* értelmezésére alapuló tárgymodellt is kidolgoztuk.

2.1 Relációs tárgymodell

A Horn-klózok relációs tárgymodell szerinti alkalmazásakor egy program bemenet/kimenet relációját egy adott Prolog szabály számítja ki. Ha egy reláció több részrelációból van összetéve, akkor azokat a szabály feltételében nevezzük meg úgy, hogy a be- és kimenő paraméterek egymáshoz láncszerűen kapcsolódnak. Az ilyen szerepű változókat a Prolog programozók *akkumulátorpárnak* nevezik.

```
reláció(BE,KI):-
  rész1(BE,TMP1),rész2(TMP1,TMP2),…,részN(TMPN-1,KI).
```

A Definite Clause Grammar (DCG) formalizmus relációs tárgymodell szerinti nyelvtani elemzésekor a <bemenet-elemzetlen szöveg> párt használjuk akkumulátorként, a tetszőleges argumentumszerkezethez az akkumulátorpárt pedig a DCG előfordító maga hozza létre.

```
nonterm(…)→nonterm1(…),nonterm2(…),…,nontermN(…).

nonterm(…,BE,KI):-
  nonterm1(…,BE,TMP1),nonterm2(…,TMP1,TMP2),…,
  nontermN(…,TMPN-1,KI).
```

A megoldás egyik hátránya: a *relációk nemdeterminisztikus kiértékelése* miatt az eredményreláció számossága legrosszabb esetben az egyes részrelációk számosságának a szorzata is lehet. Ha viszont a szorzatban az első részreláció számossága nagyobb, akkor a nemdeterminizmus visszalépéses kezelése miatt egészen az első relációig tartó, ún. *mély visszalépés* történik.

A ℜeALIS relációs tárgymodell szerinti megvalósításában a bemenő paraméter az elemzendő szöveg, a kimenő pedig a szövegnek megfelelő logikaikifejezés-szerkezet. Értelmes részrelációk lehetnek: szóalaktani, nyelvtani-szemantikai elemzés, ill. pragmatika. Ilyen értelmezés mellett ugyanazt a szabályt használhatjuk elemzésre, (ha híváskor TEXT adott, LOGEXPR viszont változó), illetve szöveggenerálásra is (ha híváskor TEXT változó, de LOGEXPR adott).

```
text2logic(TEXT,LOGEXPR):-
  morphology(TEXT,MORPHLIST),
  syntaxSemantics(MORPHLIST,PUREEXPR),
  pragmatics(PUREEXPR,LOGEXPR).
```

Sajnos a relációs tárgymodell és az ezzel összefüggő Prolog DCG formalizmus a mi céljainkra nemigen alkalmas. A ℜeALIS környezeti feltételei (pl. vonzatok bizonyos távolságban) csak úgy lennének elemezhetők, ha azokat a bemenő szövegben előre-hátra mozgással ellenőriznénk. Ennek a megvalósítása is körülményes, és komoly hatékonysági aggályokat is felvet.

A ℜeALIS megvalósítás célkitűzése a szöveg és a diskurzusreprezentáció közötti reláció kiszámítása. Ez (Prolog-szerű értelmezésben) mindkét irányú kapcsolatot jelenti. Ha a szöveg adott, akkor a program azt a reprezentációs kifejezést számítja ki,

amely az adott logikai rendszerben és az interpretáló belső tudatállapotát leíró tudásbázisban (ontológiában) kiértékelhető, bizonyítható, vagy hozzávehető a tudásbázishoz. Az ellenkező irányban: ha a tudáskezelő összetevő által (pl. egy kérdésre adott válaszként) egy logikai kifejezést kapunk, akkor a reláció a szöveg képét állítja elő.

A megoldás másik hátránya, hogy a szöveg legalább egy bekezdésnyi, de esetleg akár több oldalnyi hosszú is lehet. Ez egyrészt a feldolgozás időigényét behatárolja, másrészt a hosszú bemenő adatokon az igen mély visszalépések csökkenthetik az elemzés hatékonyságát. Harmadrészt a szélsőségesen összetett adatszerkezetek sok Prolog-megvalósítás fizikai határait is feszegethetik (pl. veremtúlcsordulást okozhatnak).

2.2 Következtetéses tárgymodell Horn-klózokon

A következtetéses tárgymodell esetében a bemenő szöveget nem listaparaméterként, hanem *tényállításokként* ábrázoljuk. A cikkben feltételezzük, hogy a szóalaktani elemzés már megtörtént, és már csak a nyelvtani-szemantikai elemzés van hátra.

```
word(peter,1,1,noun('Péter',proper,nom,sing-3)).
word(peter,1,2,verb('hasonlít',[], decl, pres, sing-3)).
word(peter,1,3,noun('az',pro(point),sub,sing-3)).
word(peter,1,4,art(def,cons)).
word(peter,1,5,adj('vörös')).
word(peter,1,6,adj('ukrán')).
word(peter,1,7,adj('futó')).
word(peter,1,7,noun('bajnok',common,sub,sing-3)).
```

A ℜeALLan szabályok követel-kínál mechanizmusa szinte kínálja magát arra, hogy Horn-klózokká képezzük le őket. Az alábbi klóz pl. a 'hasonlít' ige és kötelező vonzatai közötti kapcsolatot írja le.

```
regArg2(ID,S,XV,verb('hasonlít',[],MODE,VTIME,AGR),
             XS,noun(SUBJ,SKIND,nom,AGR),-7,
             XO,noun(OBJ,OKIND,sub,OAGR),7):-
  verb(ID,S,XV,'hasonlít',[],MODE,VTIME,AGR),
  gqdet(ID,S,XS,SUBJ,SKIND,nom,AGR), order(XV,XS,-7,nei),
  gqdet(ID,S,XO,OBJ,OKIND,sub,OAGR), order(XV,XO,7,nei).
```

Szintén Horn-klózok írják le a ℜeALIS σ (sigma) függvényének megfelelő eventuális kifejezések részkifejezésekből történő felépítését is.

```
sigma3(ID,S,XV,TIME,SUB,OB,CLAUSE):-
  regArg2(ID,S,XV,verb('hasonlít',[],_MODE,VTIME,_AGR),
       XS,SUBJ,_PRS,XO,OBJ,_PRO), {TIME =.. [VTIME,_]},
  sigma3(ID,S,XS,TIME,SUB,CLAUSE,
           (desire(TIME,SUB,OB):-CONS)),
  sigma3(ID,S,XO,TIME,OB,CONS).
```

A fenti állítás eredményeképpen a mondat logikai alakjaként a következőket kapjuk. (A kettős implikáció egy egyszerű normáló program segítségével átalakítható feltételek konjunciójává.)

```
CLAUSE=((similar(pres(T),SUB,OB):-
           run(T, OB),ukrain(T, OB),
           red(T, OB),champion(T, OB)):-
             name(T,SUB,'Peter'))
```

2.3 Visszafelé haladó tárgymodell (Prolog)

A visszafelé haladó tárgymodell magát a Prolog értelmezőt használja következtető motorként úgy, hogy az általános következtetéses tárgymodellt használja. Ebben a megközelítésben az elemzést a logikai alakra változóként hivatkozó *célállítás* hívásával indítjuk. Ha visszavezethető a célállítás a szöveget rögzítő tényállításokra, akkor a mondat elemezhető volt, és a közben elvégzett változóhelyettesítésekből kiadódik a célállításban szereplő logikai alak is.

A megközelítés egyik hátránya, hogy a bizonyításhoz *hipotézist* kell felállítani, ez gyakorlatilag a célállítás. A bizonyítás időpontjában már minden ténynek ismertnek kell lennie – a rendszer nem alkalmas csővezeték- (pipe) -szerű feldolgozásra.

Másrészt viszont a visszafelé bizonyítás logikája szerint még az ismétlődő részbizonyításokat is újra és újra elvégzi, ezzel romlik a hatékonysága.

A fentebb vázolt tárgymodell alapvetően *deduktívan, felismerőként* használható, mégis kicsi módosítással *abduktív, szöveggenerátor* célú használatra is alkalmas. Ha a célállítást a logikai alak megadásával, de hiányzó szövegkép-tényállításokkal indítjuk, akkor a visszafelé bizonyítás során előbb-utóbb a tényállítások szintjéig ér. Ha az üres tényállításokat *visszaléptethető állításfelvétellel* (`assert`) valósítjuk meg, akkor a program végeredményben abduktív bizonyítást fog végezni.

```
word(ID,S,X,WORD):-
  (assert(w(ID,S,X,WORD));
  retract(w(ID,S,X,WORD)), fail).
```

2.4 Contralog: Horn-klózok előre haladó értelmezése Prologban

A Contralog tervezésekor cél volt, hogy az előre- és visszafelé haladó működés integrálható legyen úgy, hogy a logikai forrásnyelv ugyanaz (a Horn-klózok nyelve), amit részben maga a Prolog visszafelé haladóan, részben pedig az előrehaladó motor akként értékelhet ki. A kétféle rezolúciós stratégia pedig a programozó által vezérelhetően legyen váltható: egyrészt a Prologból legyen meghívható az előrehaladó motor, másrészt az előrehaladó végrehajtásból legyen meghívható a Prolog.

A Contralog programnyelv a Horn-klózok nyelvét (a Prolog nyelvet) előrehaladó stratégiát megvalósítva képezi le a Prolog nyelvre magára úgy, hogy egy inkrementális fordítóprogram a beolvasott Contralog-szabályokat Prolog-szabályokká fordítja le, és a szabványos Prolog futtatókörnyezetben működteti. [4]

Az így létrehozott rendszerben tehát minden fordítva működik, mint a Prologban:

- A következtetést nem a célállítások, hanem a *tények* indítják.
- ha van olyan szabály, amelynek feltételrészében egy adott tény szerepel, akkor megvizsgáljuk, hogy a feltétel többi részét már sikerült-e bebizonyítani korábban. Ha igen, akkor a *szabály tüzel*, vagyis a következményrészt sikerült bebizonyítanunk.
- A bebizonyított következmény újabb *egységklóz rezolvenseket* (bebizonyított tényeket) jelent, amelyet a *munkatáblán* (blackboardon) eltárolunk, és ezzel a ténnyel folytatjuk a bizonyítást.
- A következtetési folyamatot a *célállítások* állítják le.
- Célállítás elérésekor, vagy ha bármi okból a bizonyítás az adott láncon tovább nem folytatható, a rendszer visszalép, és egy korábban nyitva hagyott alternatíva mentén próbálkozik újra.

A Prolog-Contralog kapcsolatot kétféleképpen lehet működtetni:

- a Contralog-szabályok feltételrészében a `{}/1` literál közvetlen Prolog cél meghívását eredményezi.
- A Contralog *importok* azok a tények, amelyek egy modul következtetési láncát elindítják. Ez az indító tényeknek megfelelő Prolog tüzelési szabályok exportját jelenti.
- A Contralog *exportok* viszont azok a predikátumok, amelyeket az előre haladó stratégia szerint tényként kikövetkeztettünk, és vagy másik modul importját elégítjük ki vele, vagy a Prolog futtatórendszer egy predikátumát hívjuk meg. A Contralog-exportokból Prolog-importok lesznek (, bár ezt a fogalmat a szabványos Prolog nem ismeri).

A fent ismertetett alapműködésen túl az elburjánzó következménytények törlésére logikán kívüli eszközöket vezettünk be:

- minden tárgymodulban létrehoztunk egy, a *munkatáblát teljesen törlő* Prolog eljárást, amit a `MODULE:clean` hívással indíthatunk.
- egyes tények kikövetkeztetésekor *letilthatjuk a következtetést* az adott szálon (a tényt a munkatáblán tároljuk ugyan, de a megfelelő tüzelő eljárásokat nem hívjuk meg). Ezt a működést a `:- lazy NAME/ARITY.` deklaráció hatására válthatjuk ki.
- egyes tények kikövetkeztetésekor az *azonos névjegyű tényeket mind töröljük* a munkatábláról (`:- var NAME/ARITY.`), vagy egyes argumentumokat – a relációs technológiához hasonlóan – kulcsként tekintve, csak az azonos kulcsú tényt töröljük. Ezt a `:- key(NAME(KEYVECTOR)).` deklarációval válthatjuk ki, ahol a `KEYVECTOR` szerkezet egy argumentumlista, ahol a „+” jel azt jelzi, hogy az argumentum kulcsként szerepel, a „-” pedig azt, hogy nem.

Az előre haladó következtetés alapproblémája, hogy a klózok feltételrészén több elemi feltétel is szerepelhet. Amikor ezek közül nem mindegyik elégül ki, a hiányzókat meg kell várni, és a következmény tüzelését csak akkor indítjuk, ha az utolsó feltétel is kielégült. Ezt úgy érjük el, hogy a már kielégülteket dinamikus állításokként tároljuk, és egy Contralog-szabály összes feltételliteráljához létrehozunk egy külön Prolog-szabályt, ami ellenőrzi, hogy a többi feltétel már korábban teljesült-e. Vegyünk egy egyszerű példát, tekintsük a következő Contralog-szabályt!

```
a:-b, c.
```

Ha a `b` vagy a `c` feltételek kielégültek, akkor az eredményként kapott tények a megfelelő `b/0`, ill. `c/0` dinamikus állításokban találhatók. Mindegyik feltételhez létrehozunk egy `fire_NAME` tüzelő, és egy `test_NAME` ellenőrző Prolog predikátumot. Az előbbi tárolja a kikövetkeztetett tényt, majd meghívja az utóbbit. Az utóbbi pedig ellenőrzi, hogy a többi Contralog-feltétel teljesül-e, és ha igen, akkor meghívja a következményhez tartozó tüzelő eljárást.

A fenti esetben ez a következő Prolog-kód létrehozását jelenti:

```
fire_b:- assert(b), test_b.
fire_c:- assert(c), test_c.

test_b:- c, fire_a.
test_c:- b, fire_a.
```

A fenti tárgymodellben továbbra is a Prologhoz hasonló *visszalépéses keresés* történik. Választási pontok többféleképpen is keletkezhetnek.

- Ha egy feltétel több Contralog-szabályban is szerepel, akkor annyi Prolog-alternatíva jön létre belőle, ahány szabályban a feltétel szerepel.
- Ha egy feltétel többször is teljesül, akkor ugyanannyi *dinamikus tény* jön létre belőle – feltéve, hogy az adott feltételre nem teljesülnek a következtetési ágak megnyirbálását célzó deklarációk.
- A modul összes statikus tényállításának a tárolása úgy történik, hogy a Prolog modul célállítása visszalépésesen meghívja az összes *statikus tény tüzelő eljárását*. Vagyis, ha valamilyen feltétel nem teljesül, akkor végső soron akár egészen a Prolog-célállításig is történhet egy visszalépés.

A nyitott választási pontokra a visszalépések során kerül a vezérlés. Visszalépés szintén többféleképpen bekövetkezhet

- Ha valamelyik feltétel az adott pillanatban *nem teljesül*. Ez lehet Contralog-feltétel, de a feltételek közé beszúrt Prolog-feltétel meghiúsulása is.
- Ha egy Contralog-célállítás elérésekor (a Prologhoz hasonlóan) újabb megoldások kérésével *visszalépésre kényszerítjük* a rendszert.

2.5 Előre haladó tárgymodell (Contralog)

Az előre haladó tárgymodell esetében a *szabályalkalmazási rohamokat* (burstout) az egyes mondatelemek, mint tények felvétele (beérkezése) indítja. A tények érkezhetnek *aszinkron módon*, időben elcsúsztatva, sőt akár tetszőleges sorrendben is: egy következtetési lépés akkor történik meg, ha minden feltétel megérkezett és rendelkezésre áll. Bár van lehetőség a *következtetési fa ágainak nyírbálására*, a következmények a teljes gazdagságukban előállnak, ha ezekből néhány illeszkedik a megadott célállításokra, akkor a következtetés leáll.

A modell előnye, hogy az egyszer bebizonyított tényeket tároljuk, és azokat akárhányszor fel lehet még használni.

Sajnos az előrehaladó modell abduktív módon szöveggenerálásra történő használata nem látszik kézenfekvőnek.

3 Értékelés

A tesztmondatok elemzése a bemutatott modellváltozatok alapján elegendő tapasztalatot szolgáltatott. A következő lépés a ℜeALLAN-Horn-klóz fordítóprogram megírása lehet. Károly Márton munkájában az elemzési modellt modalitások beépítésével egészíti ki. A modalitások kezelése pedig kijelöli az utat a háttérben alkalmazott tudástár összetevő megtervezéséhez – egy *multimodális többszereplős logikai következtető rendszer* képében.

A szerzőt e cikk alapjait jelentő kutatásaiban az OTKA T60595 sz. projektje támogatta, a konferencia-részvételt pedig a TÁMOP-4.2.1.B-10/2/KONV/2010/ KONV-2010-0002 (A Dél-dunántúli régió egyetemi versenyképességének fejlesztése).

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. Clockshin-Mellish: Programming in Prolog. Springer Verlag, Berlin, Heidelberg, New York (1994)
2. Alberti, G.: ℜeALIS. Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
3. Alberti, G., Kilián, I.: Vonzatkeretlisták helyett polaritásos hatáslánccsaládok - avagy a ℜeALIS σ függvénye. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2010) 113–127
4. Kilián, I.:: Contralog: egy előre haladó, Prolog-konform következtető motor, és alkalmazása ℜeALIS nyelvi elemzésre. In: SzámOkt 2011. konferencia kiadványa. Erdélyi Magyar Műszaki Tudományos Társaság, Kolozsvár (2011) 199–205
5 Nakashima, H.: Term Description: A Simple Powerful Extension to Prolog Data Structures Electrotechnical Laboratory, Umezono, 1-1-4, Ibaraki, Japan (1985)

Interpretáció, intenzionalitás, modalitás – avagy a ReALIS λ függvényének implementációja felé

Károly Márton[1]

[1] Pécsi Tudományegyetem, „Science, Please!" Projektiroda,
7622 Pécs, Vasvári Pál u. 4.
harczymarczy@gmail.com

Kivonat: Projektünk célja egy egyszerűbb diskurzusokat elemezni képes interpretáló rendszer implementálása. Ennek keretében 4 függvényt definiáltunk, ezek közül korábban a morfoszintaxist megragadó σ-ról volt szó. Most az intenzionalitást és modalitást leírni szándékozó λ szintfüggvény kerül terítékre, amely magának a világocskastruktúrának a kialakításáért felel. A függvény működését bemutatjuk néhány példán, majd, részben kódrészletek segítségével, eddigi eredményeinkre támaszkodva felvázoljuk az implementáció lehetséges útját, rávilágítva néhány problémára és lehetőségre. A λ-val kapcsolatban további elméleti cikkek megjelenése is várható, ezek főképpen a szintemelésért, szinttartásért felelős nyelvi elemekről, egyes partikulák jelentéséről (a λ tükrében) és általában a λ pragmatikai vonatkozásairól szólnak majd.

1 Bevezetés

A ReALIS projekt hosszú távú gyakorlati célja egy (később lehetőleg gépi fordításra is alkalmassá tehető) interpretáló rendszer implementálása. Kutatásunk az elméleti és számítógépes nyelvészet határterületén helyezkedik el, így része az elméleti modell felállítása, majd pedig annak implementálása.

Modellünk logikai és diskurzuselméleti alapokon nyugvó, totálisan lexikalista, kampiánus reprezentacionalista modell, melynek implementációjához egy szintén szabályalapú eszközt, a Prologot és kiterjesztéseit használjuk. Megközelítésünk azonban különbözik a klasszikus reprezentacionalizmustól annyiban, hogy az interpretáló elmét (benne a nyelvvel) is a világ részének tekintjük, ugyanazon eszközöket használva magának a világnak és az azt interpretáló elméknek a modellezésére. Ily módon – vagyis azáltal, hogy a reprezentáció „köztes" jellegét megszüntetjük és az egész világ leírásának egységes keretet adunk – tehát a legszigorúbb antireprezentacionalisták kívánalmainak is igyekszünk eleget tenni.

Szabályaink **lexikai** szabályok, magát az elemzett nyelv nyelvtanát is a lexikonban tároljuk, eltüntetve ezáltal a különbséget lexika és grammatika között. A [2]-ben definiált és hasonló generátorfüggvények a maglexikonból új lexikai egységeket állítanak elő. Így kezeljük pl. a magyar szórendet vagy a mondatban szereplő szabad határozókat: a generátorfüggvények előállítják az ige összes, szintaktikailag

lehetséges vonzatkeretét, a szórendi variánsokat, ill. a szabad határozókkal kibővített esetkereteket. Célul tűztük ki továbbá más nyelvekben található jelenségeknek a ReALIS keretei közé való beillesztését, mint pl. a német szórend, összehasonlítva a magyarral. Ezen részcél érdekében részben egynyelvű célnyelvi, részben kétnyelvű (bécsi egyetem, Finnugor Intézet) környezetben terepmunkát is folytatunk. Farkas [6] a finn nyelv szintaxisát is formalizálta (indexelt generatív módon), ugyancsak alapot teremtve ezzel a rendszerünkbe való beillesztésre.

Elméletünknek vagy egyes részeinek bizonyítása vagy cáfolata annak számba vételével lehetséges, mely nyelvi jelenségeket ragadunk meg, és melyeket nem. A helyesség bizonyításának legkézenfekvőbb módja azonban az, ha az elméletet „lefordítjuk" valamely programozási nyelvre, azaz programot írunk rá, és az az általunk elvárt eredményt adja. Ennek tükrében a ReALIS talán legfontosabb mérföldköve az lesz, ha a négy függvényt adekvát módon kezelő, legalább egy nyelvre, pl. a magyarra vagy eleinte annak egy korlátozottabb változatára jól működő, egyszerűbb szövegeket, minidiskurzusokat morfológiailag, szintaktikailag, szemantikailag és akár pragmatikailag is elemezni tudó programot fel tudunk mutatni.[1]

Bár kezdetben programozástechnikailag és részben ennek következtében a nyelvi szintek tekintetében is alulról felfelé haladtunk (kezdve a GeLexi projekttől), a nem kellően kidolgozott adatstruktúrák miatt az előrehaladás egyre nehézkesebbé vált. Járhatóbb útnak tűnik ugyanakkor a ReALIS négy (σ, α, λ és κ) függvényének fokozatos, egyenkénti kidolgozása, a folyamatos publikációk mellett részleges implementációkkal, tanulmányprogramok írásával egybekötve. Ezt követheti elvben a függvények „összeépítése" kész vagy könnyen késszé fejleszthető rendszerré.

A ReALIS modell részleteiről, az implementáció néhány kérdéséről és az eddig elkészült tanulmányprogramokról már korábbi publikációinkban is beszámoltunk ([1], [2] [3], [4], [5] stb.). A morfoszintaxist, a referensazonosítást és a fiktivitási/modális hierarchiát egy-egy függvénnyel (σ, α és λ) írjuk le, míg az időt, az eseményszerkezetet és az aspektust a κ kurzorral kezeljük. Mindennek eredménye egy kampiánus [7], DRS-ekből álló, de sajátos szintcímkerendszert használó összetett struktúra.

Az imént említett publikációk az általánosságokon túlmenően még döntően a σ függvényt tárgyalták. E cikk ugyanakkor már a fentebb leírt elgondolásba illeszkedik: a σ függvény után most a λ-ra – és a szintcímkék rendszerére – fókuszálunk. A lehetséges címkék halmaza véges és adott interpretáció vonatkozásában szigorúan meghatározott, bár céljainknak megfelelően bővíthető új nyelvészeti, logikai, pragmatikai elemekkel. A pontos definíciót (a másik három függvényével együtt) lásd [5:146-147].

[1] Utóbb Kilián [8] morfológiailag előzetesen elemzett szöveget vett ugyan alapul, az elméleti következetesség ugyanakkor megkívánja a morfológiai elemzés analóg módon történő implementálását. A projekt keretében morfológiai elemző is készült ugyan, ám, mint említettük, az adatstruktúrának az akkor még nem kellően kidolgozott szintaktikai és szemantikai adatszerkezettel való összefésülése már komoly gondot jelentett.

2 A ReALIS λ függvénye

A λ feladata egyes hatóköri viszonyok, valamint a propozicionális attitűdök és retorikai relációk megragadása. A szöveg elemzésekor a referenseket a σ függvénnyel konstruáljuk meg. Az α feladata az azonossági vélelmek meghatározósa, ám alkalmazása előtt a referensekhez hozzá kell rendelni a szintcímkéket, mert csak így tudjuk az α alkalmazási feltételeit vizsgálni. **Vagyis a λ szempontjából releváns nyelvi elemekhez hozzá kell rendelnünk azok szintmódosító tulajdonságát is.**

Előfordulhat persze, hogy a λ működését nyelvileg közvetlenül csak nehezen vagy egyáltalán nem megragadható tényezők vezérlik. Ekkor általában több pragmatikai értelmezés is lehetséges. A Prolog visszalépési mechanizmusa segítségével még ez is kezelhető (bár rásegítések nélkül nem túl hatékonyan). Szükségünk is lehet erre, mert a diskurzus további elemzése során kiderülhet, hogy az addig lehetségesnek tűnő elemzések közül néhány hibás.

A $\lambda[i,t]^{\Lambda}$: $\Lambda \times U[i] \rightarrow U[i]$ szintfüggvényt az i interpretáló belvilágában értelmezzük. U[i] elemei a **referensek**, ezek csakis valamely interpretáló belvilágában létezhetnek (míg a külvilágban infonokról, magrelációkról és entitásokról beszélünk). Csak az ún. **fiktív** referenseknek lehet képük, ezek pontosan egy szintcímke mellett képeződnek le egy másik referensre (, amire λ ismét alkalmazható stb.). A λ iterációja révén kapott (véges) címkesorozatot nevezzük a referens **világocskaindexének.** Az ún. gyökérreferensekre a λ soha nincs értelmezve, ezek világocskaindexét üresnek tekintjük. Viszont csak ezek horgonyzódhatnak ki a külső világba az α függvénnyel. (Ugyanakkor egy referens lehet külső képviselő nélküli is, pl. egy vágy tárgya.)

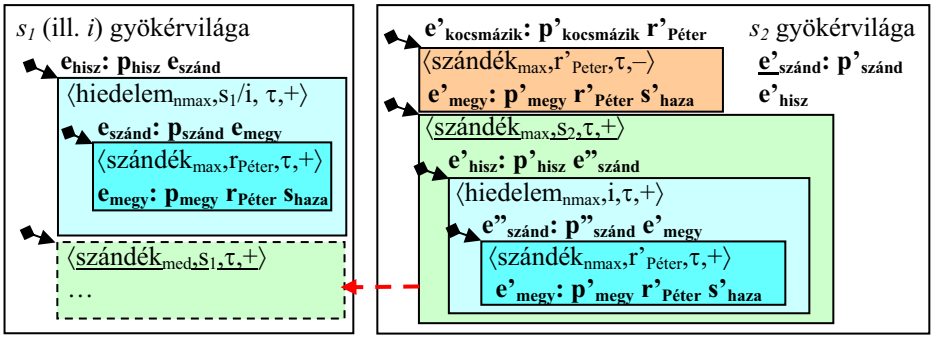
A szintcímkék Λ halmaza egy rendezett négyesekből álló halmaz: Λ_{modal} (⊂ [⬀⬉][.?!][supp|cons|bel|des|int|…], **modális címke**) × T_m (τ **időpillanat**) × U[i] (j **közvetlen gazda**, kihorgonyozva egy interpretáló-entitáshoz) × P (={+,0,–}, pozitív, semleges vagy negatív **polaritás**). Modális címkével jelöljük pl. a feltételezést (*supp*), következtetést (*cons*), a hiedelmet (bel_n), vágyat (des_n), szándékot (int_n, utóbbi háromnál *n* ranggal vagy egyéb módon jelezhetjük az erősséget), az öt érzékszervtől származó információt (*hear, see, smell, taste, touch*), a pragmatikai kifejtést (*elab*), narrációt (*narr*), valamint az utóbbi kettőre vonatkozó kérdést is (*?elab, ?narr*). Ezeken felül címkét kaphat magyarázat (*exp*), háttér-információ (*back*) vagy arra vonatkozó kérdés (*?back*), ellentét (*contr*), párhuzam (*par*), logikai művelet (*disj, conj* stb.). A felszólítás mint a szándék explicit kifejezőeszköze ugyancsak külön címkét (⬉*!*int_n) kap. Ebből világos az olvasó számára az is, hogy a modális címke három elemből áll: a nyíl lényegében a klasszikus mellé- és alárendelésnek (**szintemelő** és **szinttartó** jegy), a ponttól különböző írásjel a kérdésnek, ill. felszólításnak felel meg (**módjegy**), míg a harmadik elem a tulajdonképpeni modális tartalom.[2]

A λ értelmezésében a τ időpillanat is rögzített, de fontos, hogy a τ-k és i-k egymásba ágyazott világocskák esetén is különbözhetnek (pl. egy vélekedés esetén).

[2] Az eredeti definícióban [4] a szintcímkék funkcióinak angol nyelvű rövidítése szerepel, e cikkben viszont a továbbiakban a teljes magyar elnevezéseket használom.

2.1 Példa a λ függvényre

A λ működését először a *Péter hazamegy* mondat egyszerű példáján illusztráljuk „pragmatiko-szemantikai" szempontból. Ez persze másképp nézhet ki egy igazmondó s_1 és egy hazug s_2 beszélő (akinek célja a megtévesztés) szemszögéből, és a hallgató (interpretáló) csak az s_1-re és s_2 -re vonatkozó **háttértudása** alapján dönthet arról, elhiszi-e az elhangzott mondatot vagy sem, azaz: melyik világocskájába helyezi el azt. (Megj.: s_2-ről feltételeztük, hogy Péter alkoholizálási hajlamait próbálja eltitkolni.)



1. ábra A *Péter hazamegy* mondat kimondása mögötti két lehetséges elmeállapot ábrázolása a ReALIS modellben. s_1 valószínű tényt állapít meg, míg s_2 megtéveszti beszédpartnerét.

Az 1. ábrával kapcsolatban megjegyezzük: ahhoz, hogy elfogadhassuk igaznak az „s_2 feltett szándéka az, hogy i azt higgye, hogy Péter valóban hazamegy" statikus interpretációt, szükség van az **erre vonatkozó eventualitásokra is** a külső világban. (Ez jelentőségét akkor nyeri el, amikor az interpretálói információállapotban más interpretáló belső világáról való információk is szerepelnek; míg saját magáról mindenki tudja, mit hisz el, mit nem.) A világocskastruktúra mindenképpen létrejön, az eventualitás viszont csak akkor, ha már maga az interpretáló is viszonyulni próbál (elhiszi vagy nem stb.) a másik személyről birtokában lévő információhoz. A Hob–Nob-mondatoknál viszont pl. ezek az eventualitások **nem** jönnek létre, ezért lehet az erre vonatkozó statikus interpretáció eredménye negatív [5:283-285]. Itt viszont az attitűd (pl. hiedelem) világocskáján túl annak eventualitása is létre kell, hogy jöjjön.

Az 1. ábrán az is látszik, hogy ugyanazon referensnek egyidejűleg több példánya is lehet, ha ugyanazon megnyilatkozás révén jön létre. Ez egyebek mellett a λ függvény (?) adatstruktúrájának faszerkezetűvé alakítását tette szükségessé (l. később).

Az s_1 beszélőhöz tartozó ábrában nincs kifejtve s_2 szándék-világocskájának megfelelője. A *Péter hazament* mondat ugyanis pontos információ vagy erős hiedelem birtokában kimondható, mégpedig valós pragmatikai célunktól függetlenül. Az esetek többségében persze információt adunk át, tehát alapvetően s_1 szándéka is arra irányul, hogy i-ben legalábbis kialakuljon egy erős hiedelem Péter hazamenéséről, azaz a világocska szükséges. s_2 célja azonban nem lehet nagyon más, mint i megtévesztése: biztos forrásból tud Péter lerészegedéseiről, esetleg éppen aznap is találkozott már vele egy kocsmában. A naív i interpretáló pedig s_1-éhez

hasonló információállapotba kerül, persze immár a beszédszándékra utaló világocska nélkül.

Ha ez után i egy későbbi τ' időpillanatban értesül az igazságról, attól még a régi hiedelme τ időpont vonatkozásában megmarad. Ha tehát egy s_3 beszélő felvilágosítja i-t Péter alkoholizálási szokásairól, akkor i a régi hiedelmet (↖.hiedelem$_{nmax}$, τ) és az új, hallás útján szerzett információt tartalmazó (↖.hallás, τ') világocskák tartalmából, valamint s_2 és s_3 szavahihetőségére vonatkozó **háttérinformációi** alapján alakítja ki a τ' időpontban érvényes új hiedelmét (ami persze később ismét módosulhat). A háttérinformáció-világocska címkéje ⇗.háttér, a hozzá tartozó időparaméter mutathat akár τ'-re, akár τ-ra vagy még régebbre, amennyiben az információforrás szavahihetőségének vélelme időközben nem dőlt meg.

Mindezek után egy ún. **akkomodációs** lépés szükséges az új hiedelem kialakításához és a régi megdöntéséhez. Ez nem jelenti ugyanakkor azt, hogy a világocskáját is fel kell számolni: a *Mari korábban azt hitte, hogy Péter hazament* mondat igaz marad. Létrejön ugyanakkor egy új ↖.hiedelem$_{med}$, τ' világocska – benne Péter kocsmázásának eventualitásával. Minderről még pontosabb leírást kaphatunk, ha néhány **szabályleíró eventualitást** is behozunk háttértudásként, azaz akár ⇗.háttér címkéjű világocskába beágyazva: *Ha* (↖.feltételezés) *valaki kocsmázik, akkor nem otthon van* (↖.következtetés)*, ha valaki hazament, akkor nem kocsmázik.*[3]

Célunk tehát körülírni azt, hogy a ReALIS elméleti keretei között hogyan történhet a módbeli és intenzionális (segéd)igék, modális partikulák és morfémák elemzése. Összességében ezek tekinthetők a λ nyelvi pilléreinek. Modellünk kereteibe beilleszthetők ugyan olyan, az írott nyelven kívüli kifejezőeszközök is, mint a gesztusok és az intonációs sémák (ezek közül a hangsúlyozásról [2]-ben szót is ejtettünk), azonban a mi elsődleges célunk a billentyűzeten bevitt nyelvi input elemzése és az 1. ábrához hasonló doboz- (világocska-)struktúra mint interpretáció felépítése. Amennyiben ez kellően hatékonyan lehetséges, akkor egy következő lépésben a folyamatot megfordítjuk és egy másik nyelven generálunk egy olyan szöveget, amelyhez ugyanazok a struktúrák tartoznak.

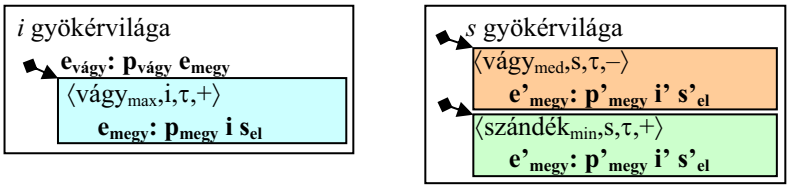
3 Adatok, adattárolás

3.1 A világocskák és referensek leírásához használt adatszerkezetről

A ReALIS implementációjának sikere vagy kudarca múlhat azon, hogyan ábrázoljuk a lexikon adatait, ideértve a feldolgozás során jelentkező, az `assert` predikátummal létrehozott tényeket is. A λ függvény esetén sincs ez másképp, sőt a modalitást és intenzionalitást kifejező szavak esetén meg kell találnunk annak a módját is, hogy a λ-szintcímkéket érintő lexikai szabályokat is egységes keretek között tároljuk.

[3] A „hazamenés" és a „kocsmázás" persze nem zárják ki egymást teljesen: ha Péter a szülőfalujába utazott, majd beült a helyi csapszékbe, akkor a két eventualitás egyszerre is fennállhat. Mi azonban a fenti okfejtés során végig egymást kizárónak tételeztük fel e két eventualitást, egyszerűsítési okokból leszűkítve a *hazamegy* jelentését.

Az eredeti, [5:146-147] alatti rekurzív definíció átvétele egyrészt implementációs szempontból nem hatékony, másrészt felvetődött egy olyan elméleti jellegű probléma is, amely a λ újragondolását tette szükségessé. Ez akkor jelentkezik, ha ugyanazon megnyilatkozás révén ugyanazon referenseket egyszerre több különböző világocskában helyezzük el. Erre a talán legegyszerűbb példát az *Egye fene, elmehetsz* magyar mondat elemzése szolgáltatja. A szereplők itt is s mint beszélő és i mint interpretáló. i erős késztetést (**vágyat**) érez arra, hogy elmenjen, s azonban csak többszöri ráhatásra hajlandó i-t elengedni. A lelke mélyén s továbbra is vágyik arra, hogy i maradjon, azonban meghallgatva i érvelését, végül – vágyán felülkerekedve – engedi őt távozni. Az engedélyt egy minimális erősségű szándék-világocskával jelezzük. s tehát **beletörődött** abba, hogy i távozásába, elfogadja azt (2. ábra).



2. ábra Az *Egye fene, elmehetsz* mondat kimondása mögötti elmeállapotok ábrázolása a ReALIS modellben i és s szemszögéből

A λ függvény implementációjához a fiktivitási hierarchiát jobban megragadó, eredetileg χ-vel jelölt címkesorozatot használjuk. Ez azon világocskacímkék egymásutánja, amelyeken keresztül a a gyökérvilágból a referenshez eljuthatunk. Tehát pl. az 1. ábra i interpretálóját nézve $\chi(p_{megy})$ = ⟨⟨$hiedelem_{nmax}$,i,τ,+⟩, ⟨$szándék_{max}$,$r_{Péter}$, τ,+⟩⟩. Technikai okokból, valamint hosszú távú célunkat (ami nem más, mint egy valódi multiágens rendszer építése) figyelembe véve szükséges még megjelölni azt az interpretálót, akinek elméjéhez tartozik a referens: ez esetünkben i.

A `lambda/3` tényállítások szerkezete ugyanakkor ezzel még korántsem végleges. A χ címkesorozat már könnyen átkonvertálható Prolog-listává, ugyanakkor a 2. ábrán szereplő megnyilatkozás kapcsán felvetett kérdés megoldásához χ többszörözésére és egy Γ (fa-)struktúra kialakítására van szükség. A bonyolultabb esetet mindazonáltal az 1. ábrán találjuk: s_2 p'_{megy} referensére $\Gamma(p'_{megy})$ = ⟨⟨⟨$szándék_{max}$,r'_{Peter},τ,–⟩⟩, ⟨⟨$szándék_{max}$,s_2,τ,+⟩, ⟨$hiedelem_{nmax}$,i,τ,+⟩, ⟨$szándék_{nmax}$,$r'_{Péter}$,τ,+⟩⟩⟩, míg a 2. ábrán $\Gamma(p'_{megy})$ = ⟨⟨⟨$vágy_{med}$,s,τ,–⟩⟩, ⟨⟨$szándék_{min}$,s,τ, +⟩⟩⟩.

További kérdés a gyökérvilág megjelenítése a reprezentációban. Még [1]-ben is megjelenítettük a gyökérvilágot, ám mivel definíció szerint üres χ-vel (és így Γ-val) rendelkezik, felvet egy igen komoly kérdést. Abból a feltevésből indulunk ugyanis ki, hogy egy interpretálónak összesen egyetlen gyökérvilága lehet.[4] Ha viszont mi a

[4] A ReALIS elméleti hátterét is ismerő pszicholingvistáink ugyanakkor úgy vélik, hogy pl. a skizofréniában szenvedő betegek – lefordítva a mi elméletünkre – legalább **két** gyökérvilággal rendelkezhetnek. Ekkor azonban (akár orvosi szempontból is) kérdés, mi alapján dől el az, hogy egy-egy új információ melyik gyökérvilágba, vagy ha úgy tetszik, melyik személyiségbe épül be. Annak kifejtése pedig, hogy pl. a gyökérvilág, ill. annak referensei (ideértve

mentálisan egészséges(!) interpretáló gyökérvilágát bármilyen módon felcímkézzük, nem jelenti-e ez esetleg annak önkényes megtöbbszörözését?

Úgy hisszük: igen. A gyökérvilág éppen attól válik azzá, hogy nincs modális címkéje. Valamely i' interpretálóval sem címkézhető, mert amit másról tudunk, ahhoz már vagy egy másik világocska tartozik (új címkével), vagy tudásunk legalábbis valami eventualitásként jelenik meg, amelynek egyik argumentuma az i'. A τ időpont egyvalami lehet, ez pedig a κ aktuális időkurzorértéke, vagyis a jelen. Minden más időpontról a tudásunk bizonytalan, a múltat elfelejtjük, ismereteink hamar töredékessé válnak, a jövőről pedig eleve a legritkább esetben állíthatunk biztosat. Végül a polaritás kérdését a háttértudásunkban található szabályleíró eventualitások oldják meg: ha valami *piros*, akkor az nem *zöld*. Itt azonban már ismét csak fiktív világocskákról: *háttértudásról, feltételezésről* és *konklúzióról* van szó.[5]

A fentiek tükrében tehát egy referenshez tartozó `lambda/3` tény a következőképpen nézhet ki:

```
lambda(REFID,OID,[[[COSUB,MOD,MODLEV,INT,T,P]| ...]|
...]).
```

Azaz: a `lambda/3` első argumentuma a referens azonosítója, a második az az interpretáló, akinek elméjéhez tartozik a referens, a harmadik pedig maga a Γ szintcímkelista, kétszeresen egymásba ágyazva. A belső listákban van egy-egy referenspéldányhoz tartozó címkehierarchia.

Az egyes hierarchiákon belül kérdés még a szintcímkék sorrendje: az új világocskák létrehozásának és így végső soron az elemzésnek a „belülről kifelé" sorrend kedvez, bár megnehezíti a világocskahierarchia ábrázolását.

Így tehát az 1. ábrán szereplő s_2 beszélőnél a p'_{megy} predikátumreferenshez pl. a következő Prolog-tények rendelhetők (a referensazonosítókat aláhúzással jelöltük):

```
lambda(66,11,[[[sub,int,1,55,now,-1]],[[sub,int,2,55,
now,+1],[sub,bel,2,1,now,+1],[sub,int,1,11,now,+1]]]).
%az 1. szint a legerősebb ('max') , mint általában.

ref(1,i,'_EGO',0). ref(11,i,'SPEAKER2',0).
ref(55,i,'Péter',1). ref(66,p,'megy',1).
%ref/4: azonosító, típus, lexikai egység, ref.-számláló
```

az eventuális referenseket), valamint az egyes fiktív világocskák pontosan milyen szerepet játszanak az interpretálói személyiség felépítésében, igen messzire vezetne.

[5] Itt ugyancsak messzire vezető, alapvető nyelvfilozófiai kérdésekbe botlunk. A *zöld* miért éppen *zöld?* Vagy ha egy másik bolygóról látogatók érkeznek a Földre, és hallják, hogy ugyanaz a szín egyszer *zöld*, másszor *green* vagy *vert*, akkor honnan fogják tudni, hogy éppen (nagyjából) ugyanazt fejezik ki sokféleképpen? Vagy azt, hogy a *zöld* meg a *rouge* viszont már nem ugyanaz a fogalom? Vélhetően valahogy úgy tanulnák meg, ahogy egy gyermek is elsajátítja az anyanyelvét (vagy akár egynél több nyelvet). Háttértudásukba pedig előbb-utóbb be fog épülni az *ami piros, az nem zöld* konstrukció és a kapcsolódó szabályleíró eventualitások.

A referenseket referenskonstruktorral hozzuk létre. Kérdéses még, hogy ennek integráns részét képezi-e majd pl. a λ-szintcímke hozzárendelése – az elmélet mindazonáltal ezt az elvet diktálja. A referenskonstruktor fő feladata a jelenleg négyargumentumú `ref/4` tények behozatala, amelyek egy-egy referenst reprezentálnak. A referensek jelenleg típusosak: adott típusú lexikai egységhez adott típusú referensek jönnek létre. Ugyanakkor könnyű olyan példát mondani (névszói állítmány), amelynél ugyanazon lexikai egységhez több típusú referenst is létre kell hozni, vagy a típusokat konvertálni kell. Ennek pontos megvalósítása a közeljövő egyik legfontosabb feladata.

3.2 A külvilági entitások leírása

A λ ily módon történő megragadása lehetővé teszi azt is, hogy magának a külvilágnak az entitásait (sőt, infonjait [9]) is leírhassuk, ill. hogy a `lambda/3` predikátumot felhasználhassuk a kihorgonyzásoknál is. Ha a referensek pozitív azonosítót kaptak, akkor a külvilághoz tartozó elemek számára a negatív egész számok fenntarthatók, a 0 pedig magának az orákulumnak az azonosítója. Azaz:

```
lambda(10,1,[]). lambda(-3,0,[]).
```

Ez után a 10-es gyökérreferens (amelynek λ-címkéje tehát üres) az α segítségével a –3-as entitáshoz horgonyozható ki. Ez csak a kihorgonyzás tényét hivatott megmutatni, és nem kell a rendszernek „tudnia” azt, hogy a külvilágban pontosan mi mivel azonos. Adott interpretáló vonatkozásában pedig az azonosíthatóság döntően annak háttértudásából vagy egyéb világocskáiból következtethető ki, és maga az azonosítás az α függvénnyel – de nem kihorgonyzással – történik.

3.3 Az adatbázis-kapcsolatról: újabb érv a Contralog [8] mellett

A skálázhatóság ma már a természetesnyelv-feldolgozó rendszereknél is alapvető követelmény. A Prologot használó rendszerek legnagyobb hátránya ennek nem kielégítő mértéke volt. A modern Prolog-megvalósítások (pl. Visual Prolog, SicSTus Prolog) azonban már rendelkeznek pl. viszonylag jól használható adatbázis-interfésszel (pl. a Visual Prolog ODBC-n keresztül kommunikál a Microsoft SQL rendszerrel).

Régebben azonban – a skálázatóság hiánya miatt – a Prolog-alapú megvalósítások ritkán jutottak tovább a prototípus szintjénél. Ennek persze volt egy másik oka is: ha egy részállítást a Prolog segítségével ismételten bizonyítunk, akkor az előző eredményt a rendszer nem tárolja el, hanem akár többször is bebizonyítja [8].

Sokan ezért áttértek hatékonyabb eszközök használatára – lemondva ezzel a Prolog két legfontosabb mechanizmusáról: a visszalépéses keresésről és az unifikációról.

A skálázhatósághoz szükséges adatbázis-kapcsolat miatt mi – legalábbis e cikkben – a tényállítások szerkezetére, vagyis lényegében az adatszerkezetre helyeztük a hangsúlyt. Az SQL-alapú rendszerek adatrekordjai könnyen átírhatók Prolog-tényekké és fordítva, így lényegesen egyszerűsödhet a Prolog-program és az SQL-

szerver közti kommunikáció, valamint a rendszer egyéb (pl. adatbiztonsági) szempontokból nézve is kezelhetőbb marad.

A többszöri bizonyítás problémájára Kilián [8] szolgáltat használható megoldást: ez a **következtetéses tárgymodellt** biztosító Contralog rendszer. Ebben lehetőség van a `{}/1` literál révén közvetlen Prolog-cél meghívására is, ekképpen mindig az éppen szükséges irányban „hajtva meg" a rendszert.

Látható még, hogy az adatbázis-kapcsolat szempontjából fontos tényállítások, amelyekről e cikk is szól, ugyancsak kiáltanak az előrehaladó következtetést alkalmazó rendszerért. Ily módon tehát pl. egy szöveg morfológiai elemzését követően az input ugyanolyan tényállításokká alakul, mint amilyenekből a maglexikon áll majd. (A maglexikon felépítését [2]-ben vázoltuk fel, míg a kiterjesztett lexikon előállításáért felelős lexikai szabályok szintén leírhatók Contralog-tényekkel.)

4 Példa a λ implementációjára

Az említett Contralog tárgymodell segítségével megkísérelhető pl. a *vágyik* ige (részleges) implementálása is. Ha valaki *vágyik* valamire, akkor ez az előző fejezet és [8] alapján két lépésben írható le. Az első:

```
sigma3(ID,S,X,TIME,SUB,OB,CLAUSE):-
regArg2(ID,S,XV,verb('vágy',[],_MODE,VTIME,_AGR),XS,
SUBJ,_PRS,XO,OBJ,_PRO),{TIME= .. [VTIME,_]},
sigma3(ID,S,XS,TIME,SUB,CLAUSE,(desire(TIME,SUB,OB):-
CONS)),sigma3(ID,S,XO,TIME,OB,CONS),
{newref(X,e,CLAUSE)}. %%newref: referenskonstruktor.
```

A [8]-ban szereplő kódot mi kiegészítettük egy provizórikus referenskonstruktorral. Ebben a rendszerben tehát a `CLAUSE` kimenő változó értéke egy ilyesfajta Prolog-klóz lesz: `desire(SUB,OB):-car(TIME,OB)` – amennyiben a vágy tárgya egy autó, és az *autó* lexikai egységéből kinyerjük a valaminek egy bizonyos időpontbeli *autó voltára* vonatkozó `car(TIME,OB)` predikátumot. Meg kell jegyeznünk továbbá, hogy míg Kilián következetesen `SUB,OB` stb. (az angol nyelvre specifikusan alany, tárgy) változókat alkalmaz, addig magam azt az irányvonalat képviselem, hogy az argumentumokat thematikusszerep-címkékkel kell ellátni (szélsőséges esetben akár igénként külön definiálva!), fenntartva ezzel a nyelvfüggetlenséget. Természetesen szükségünk van a GeLexi-hez hasonlóan kopredikációs szimbólumokra, ha később a ReALIS-t gépi fordításra szeretnénk használni, ahogy arra a 2. fejezet végén is már céloztunk. Mi többletként egyelőre azt kötjük ki, hogy a σ mellett a λ-ra, távlatban esetleg a megmaradó két függvényre (α és κ) vonatkozó lexikai szabályok nyelvfüggetlen részének pontos vagy közel pontos, oda-vissza történő alkalmazása szükséges a fordítási adekvátsághoz. Mindez persze a fordítástudománnyal foglalkozók számára túl szigorúnak tűnhet, de az esetleges enyhítés lehetőségeinek vizsgálata önmagában is megérne egy másik cikket. Ha a nyelvi inputból elő tudjuk állítani az interpretációs struktúrát, akkor abból miért ne tudnánk az input szöveget egy másik nyelven

visszaadni? Az ehhez szükséges háttértudás problémája humán fordítóknál is jelentkezik, de mi már az interpretációnál feltételeztük ennek bizonyos szintű meglétét. Komolyabb problémának tartom az egyes nyelvek (amelyeknél az információforrás befolyásolja az alkalmazott igemódot – ausztráliai nyelvek, török stb.) specifikus elemzésére kialakított világocskacímke-rendszer pontos adaptálását egy másik nyelvre. Ha pl. az információforrást a forrásnyelv nem különbözteti meg, akkor a célnyelven akár két vagy több különböző fordítás is megjelenhet: a törökben pl. nem mindegy, hogy a beszélő látott-e valamit, vagy csak mástól hallott.

A *vágyik* ige elemzésének 2. lépése, vagyis a vágy tárgyához a λ címke hozzárendelése a következőképpen zajlódhat:

```
lambda_des(STIREF,INT,[[[sub,des,1,XPREF,T,+1]|WLR]]):-
sigma3(_ID,S,EVREF,T,XPREF,STIREF,CLAUSE),ref(EVREF,e,
CLAUSE),desire(T,XPREF,STIREF),lambda(EVREF,INT,WLR),
bassert(lambda(STIREF,INT,[[[sub,des,1,XPREF,T,+1]
|WLR]])).
```

Azaz: ha az előzőekben a `desire/3` predikátumot kinyertük az elemzés során, és tartozik hozzá egy eventualitás (`EVREF`), akkor a vágy tárgya egy szinttel „mélyebbre" kerül a vágy-eventualitás szintjéhez képest, és kap még egy `des` (vágy-)címkét is. (**NB. Ebben a példában a világocskastruktúra még lineáris!** Faszerkezetet reprezentáló lista (3.1. fejezet) esetén minden allista elejére oda kell tenni az új világocskacímkét. Ennek miként jét, vagyis pl. az *Egye fene, csak vágyakozz az után a nő után* mondat elemzését az olvasóra bízzuk.)

5 Kitekintés – szinttartás, szintcsökkentés, akkomodáció: hogyan?

Ha továbbgondoljuk az előző, autóra történő vágyakozást taglaló példát, akkor óhatatlanul adódik a következő lehetséges folytatás: *Péter nagyon vágyik egy autóra. Nagyon sokat utazna vele. (De) csak egy rozoga biciklije van.*

Már szóltunk a *vágyik* ige szintemeléséről. E példából úgy tűnik, hogy a magyar feltételes mód használata ugyanakkor **szinttartó** jelleggel bír a *vágy* vonatkozásában. A vágyvilágból történő visszalépésért pedig a kijelentő mód felel, ez egy **törlő** lexikai szabállyal programozható le.

Szintén látható, hogy a *de* szócskát tartalmazó változat a valódi szituációt (ti. hogy *Péternek csak egy rozoga biciklije van*) szembeállítja a vágyvilágocskával és ez a tény egy mellérendelt ⇗.contr (*ellentét*) világocska létrehozását indokolja – valóban?

Most megnézünk még három továbbfolytatást: 1. *Ez nagyon bosszantja őt.* 2. *Pedig az autóval könnyebben közlekedne.* 3. *A bátyja felajánlott neki egy Toyotát.*

Az 1. esetén a mondat még a λ szempontjából sem egyértelmű. Bár a vágyvilágocskából kiléptünk, a bosszúság oka lehet maga a vágy is (régóta szeretné az autót, de nem tudja megvenni), vagy a bicikli rozoga volta, vagy mindkettő, azaz: a vágy és a valóság között régóta feszülő ellentét. Ezek közül az α-nak a lexikális szemantikára vonatkozó alkalmazásával tudunk majd dönteni: bosszúságot **negatív** dolog okozhat, az pedig ⇗.háttér világocskában dől el, hogy mi negatív és mi nem az.

Ha a bicikli rozoga volta okoz bosszúságot, akkor a düh eventualitása elvileg még a ⇗.contr világocskán belülre kellene, hogy kerüljön. De akkor miféle „ellentétben” áll a düh az autóra vonatkozó vággyal?

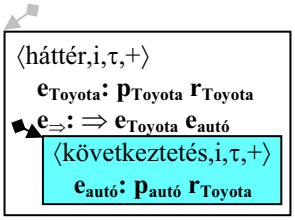
Kevésbé valószínű, hogy a vágy-eventualitáshoz köthető a düh: a szóban forgó magyar *ez* névmás pragmatikai hatóköre tipikusan az előző (tag)mondat eventualitása – az pedig a bicikli rozoga volta vagy hasonló.

A legvalószínűbb tehát az, hogy a bosszúság oka **maga az ellentét.** Ekkor azonban kérdés, hová tesszük annak az eventuális referensét. Ha világocska is tartozik hozzá (láthattuk, hogy a modalitásnak is lehet eventualitása), akkor abba – nézetünk szerint – nem kerülhet be maga a referens. Marad tehát a gyökérvilág. Akkor viszont mi jogosít fel minket arra, az $e_{\leftrightarrow}$: $\leftrightarrow$ $e_{vágy}$ e_{birt} eventualitást úgy használjuk, hogy a vágy eventualitása a gyökérvilágban, a rozoga bicikli birtoklásának eventualitása pedig a ⇗.contr világocskában legyen?

A problémát **akkomodációval** oldjuk meg: a ⇗.contr világocskát annak ideiglenes létrehozása után eventualitássá „zsugorítjuk”, és a fiktív világok közül csak a vágyvilágot hagyjuk meg. Vagy: a ⇗.contr világocskába ágyazzuk be a vágyvilágot a vágy- és a biciklibirtoklás-eventualitással együtt. A legjobban talán így írható le Péter valódi problémája, ami az ellentét-világocska puszta létében érhető tetten.

A 2. folytatás esetén a problémát a vágyvilágba való **visszalépés** jelenti. Ennek implementálása csak úgy lehetséges, ha a κ **kurzorfüggvényben** eltároljuk magukat az érintett világocskaszinteket is: tudnunk kell, hogy a feltételes mód előzőleg kinek a vágyához, feltevéséhez kapcsolódott. A κ-ról azonban ez ideig nem áll rendelkezésre akár csak kísérleti implementáció sem (a σ-val ellentétben).

Végül a 3. esetben azt kell megjegyeznünk: a vágy-világocskában szereplő autót nem szabad összehorgonyozni a Toyotával még akkor sem, ha az autó mindenben megfelel Péter vágyainak. Itt ugyanis egyszerű narrációnak tekinthető pragmatikai viszonyról van szó. A *Legjobban egy Toyotának örülne* mondatból viszont egyenesen következik az, hogy a vágyvilágba le kell képezni azt a háttérvilágocskát, amelyben *A Toyota egy autó* szabályleíró eventualitás szerepel, teljesen hasonlóan [5:273]-hoz (NB. ott viszont a vágybéli zongora és a Bösendorfer azonosítása is már valójában egy akkomodáció eredménye!).



3. ábra Példa egy szabályleíró eventualitásra: *A Toyota egy autó.*

Ami biztos: ha mindezt implementálni akarjuk, akkor egy komplett ontológiát kell a ReALIS mögé képzelni. Ez még megtehető ugyan, ha választunk egy kellően formalizált és könnyen implementálható modellt, és azt átfordítjuk a ReALIS nyelvezetére, viszont adódik az újabb kérdés: magukat az akkomodációs szabályokat hogyan írjuk le?

Talán a modális igék, melléknevek stb. eventualitásai jelenthetik erre a megoldást. Ha ezekre is kiterjesztjük a szabályleíró eventualitásainkat, elegendően erős eszközt kapunk az akkomodációs szabályok formalizálására is. De ez még a távoli jövő zenéje.

6 Összegezés

Bár a λ általunk felvázolt adatszerkezete meglehetősen egyszerűnek tűnik, nyelvi és nem nyelvi pillérei igen szerteágazóak. Cseppet sem magától értetődő tehát az az elméleti jellegű, de a gyakorlati megvalósítás szempontjából kulcsfontosságú kérdés, hogy mikor van mindenképpen szükség egy-egy új világocska létrehozására és mikor nincs. Főképp az előző fejezetben mutattunk rá néhány elméleti szempontból is alapos átgondolást igénylő kérdésre.

Láttuk azt is, hogy háttértudás ugyanazon eszközökkel ragadható meg, mint maga a nyelv. Erre elsősorban a *back* (háttértudás), *supp* (feltételezés) és *cons* (következmény) világocskák révén nyílhat mód. Lehetséges akár az ún. default következtetés mint operátor használata is.

Úgy hisszük, hogy egyes világocskák használatának, valamint az akkomodációnak a szabályai még nincsenek teljes körűen formalizálva. De miközben górcső alá vesszük a λ függvényt és megkíséreljük annak implementálását, efelé haladunk. A gyakorlati implementáció kísérletei tehát a ReALIS esetén még sokkal inkább visszahatnak a háttérelméletre, mint egy „átlagos” szoftver esetén, ideértve a természetesnyelv-feldolgozó szoftvereket is.

Köszönetnyilvánítás

A szerzőt e cikk alapjait jelentő kutatásaiban az OTKA T60595 sz. projektje, a konferencia-részvételt a TÁMOP-4.2.1.B-10/2/KONV/2010/KONV-2010-0002 (A dél-dunántúli régió egyetemi versenyképességének fejlesztése), a német nyelvvel kapcsolatban folyamatban lévő ausztriai terepmunkát pedig (mely később szintén több publikáció alapjául szolgálhat) – ösztöndíj formájában – az Osztrák-Magyar Akció Alapítvány támogatta.

Bibliográfia

1. Alberti G., Károly M.: The Implemented Human Interpreter as a Database. In: Cordeiro, J., Virvou, M. (eds.): Proceedings of IC3K the 5th International Conference on Software and Data Technologies Vol. 2. SciTePress, Funchal, Madeira (2011) 468–474
2. Alberti G., Károly M., Kleiber J.: From Sentences to Scope Relations and Backward. In: Sharp, B., Zock, M. (eds.): Natural Language Processing and Cognitive Science. Proc. 7th Int. Workshop on NLPCS. SciTePress, Funchal, Madeira (2010) 100–111

3. Alberti G., Károly M., Kleiber J.: The ReALIS Model of Human Interpreters and Its Application in Computational Linguistics. In: Cordeiro, J., Virvou, M. (eds.): Proceedings of the 5th International Conference on Software and Data Technologies Vol. 2. SciTePress, Funchal, Madeira (2010) 468–474
4. Alberti G., Kilián I.: Vonzatkeretlisták helyett polaritásos hatáslánccsaládok – avagy a ReALIS σ függvénye. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2010) 113–126
5. Alberti G.: ReALIS. Akadémiai Kiadó, Budapest (2011)
6. Farkas Judit: A finn nyelv indexelt generatív szintaxisa. Doktori disszertáció. Pécsi Tudományegyetem, Nyelvtudományi Doktori Iskola, Pécs (2011)
7. Kamp, H., van Genabith, J., Reyle, U.: Discourse Representation Theory. In: Handbook of Philosophical Logic Vol. 15. Springer-Verlag, Heidelberg (2011) 125–394
8. Kilián I.: Contralog: egy előre haladó, Prolog-konform következtető motor és alkalmazása a ReALIS nyelvi elemzésére. In: SzámOkt 2011. konferencia kiadványa, Erdélyi Magyar Műszaki Tudományos Társaság, Kolozsvár (2011) 199–205
9. Seligman, J., Moss, L. S.: Situation Theory. In: van Benthem, J., ter Meulen, A. (eds.): Handbook of Logic and Language. Elsevier, Amsterdam / MIT Press, Cambridge (1997) 239–309

Kvantifikált kifejezések hatóköri többértelműségének szabályalapú kezelése

Szécsényi Tibor

Szegedi Tudományegyetem
Általános Nyelvészeti Tanszék
szecsenyi@hung.u-szeged.hu

A magyar nyelvben az ige előtti kvantifikált kifejezések hatóköre követi a szórendet, az ige utániakra azonban jellemző a hatóköri többértelműség. Ezt a jelenséget a HPSG-ben a kvantortárolás segítségével lehet megmagyarázni. A cikk az elméleti megoldás gyakorlati megvalósítását végzi el. A Prolog-alapú, DCG nyelvtan képes kezelni a szabad szórendű magyar mondatokat, és helyes szűk és tág hatókörű olvasatokat rendeli a mondatokhoz.

1 A probléma

A természetes nyelvi kifejezések szemantikai homályosságának az egyik oka a kvantifikált kifejezéseket (*minden kalóz, háromnál több indián* stb.) tartalmazó mondatok hatóköri többértelműsége. A kötött szórendű nyelvekben, mint az angol, ezeknek a kifejezéseknek a mondatbeli pozíciója nem nyújt segítséget a kifejezések által bevezetett logikai kvantorok hatóköri viszonyainak a meghatározásához.

A magyar mint részben kötött szórendű, azaz diskurzuskonfigurációs nyelv [4], részben egyértelműsíti a kvantifikált kifejezések hatóköri viszonyait, ugyanis az ige előtti kifejezések sorrendje megegyezik a hatókörük sorrendjével (a '>' a nagyobb hatókört jelenti):

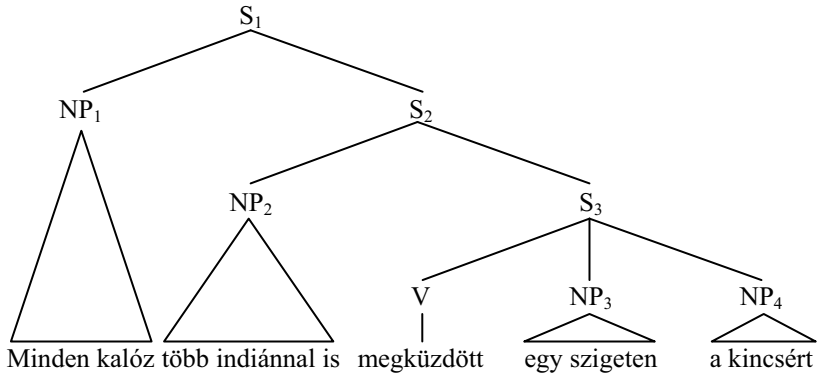
(1a) *Minden kalóz több indiánnal is megküzdött.*
minden kalóz > több indián
(1b) *Több indiánnal is minden kalóz megküzdött.*
több indián > minden kalóz

Az igét követő kvantifikált kifejezések hatóköre azonban nem meghatározott, azok hatóköre lehet kisebb is (2a: szűk hatókörű olvasat) vagy nagyobb is (2b: tág hatókörű olvasat), mint az őt megelőző kvantifikált kifejezéséé:

(2a) *Minden kalóz kibékült néhány indiánnal*.
(2b) minden kalóz > néhány indián, azaz
$\forall x \exists y$ (indián(y) $\wedge$ (kalóz(x) $\rightarrow$ kibékül(x,y))
(2c) néhány indián > minden kalóz, azaz
$\exists y \forall x$ (indián(y) $\wedge$ (kalóz(x) $\rightarrow$ kibékül(x,y))

2 Az elemzés

Korábbi [7], [8] és [9] tanulmányaimban ezt a természetes nyelvi jelenséget próbáltam leírni HPSG ([6]) elméleti keretben. Ezekben a tanulmányokban a klasszikus É. Kiss-féle ([4]) elemzés felszíni szerkezetét tulajdonítottam a magyar mondatoknak, elhagyva ugyanakkor a nála meglevő többi elemzési szintet. A mondat összetevős szerkezete tehát egy igével kezdődő, lapos frázisból és ehhez balról kapcsolódó, hierarchikus bal perifériából áll:



1. ábra: A magyar mondat összetevős szerkezete

A kvantifikált kifejezések hatókörének a meghatározásához a Head-driven Phrase Structure Grammar-ben (HPSG) használatos kvantortárolást használtam ([3]).

A kvantortárolás alapötletét az adja, hogy az olyan predikátumlogikai kifejezéseket, mint ami a (2b)-ben is látható, szétszedhetjük egy magjelentést kifejező részre (’kibékül(x,y)’) és a kvantifikált kifejezések jelentését leíró részekre: ’∀x(kalóz(x) → P(x)’, illetve ’∃y(indián(y) ∧ Q(y)’. A kvantorokban található P és Q egy-egy predikátumváltozó, lekötésükre egy-egy halmazképző lambda operátor szolgál: ’minden_kalóz’ = ’λP.∀x(kalóz(x) → P(x)’, illetve ’néhány_indián’ = ’λQ.∃y(indián(y) ∧ Q(y)’. Az így kapott tulajdonsághalmazokat (általánosított kvantorokat, kvantorokat) mint predikátumokat sorban alkalmazhatjuk a magpredikátumra, így megkaphatjuk ’minden_kalóz(néhány_indián(kibékül))’ logikai szerkezetű szűk hatókörű állítást. Ha a kvantorokat fordított sorrendben alkalmazzuk, akkor a ’néhány_indián(minden_kalóz(kibékül))’ tág hatókörű olvasatot. Ahhoz, hogy a kvantorokat tetszőleges sorrendben alkalmazhassuk a magra, először össze kell gyűjteni őket. A mondatban szereplő kvantorok összegyűjtése, majd sorbarendezése adja a kvantortárolási elemzést.

A HPSG-ben a kvantorok a kvantifikált kifejezésekből (pontosabban azok determinánsából) származnak, ott a kvantortárolóban (QSTORE) helyezkednek el. Az 1. ábrán minden NP bevezet egy-egy kvantort. Az NP kategóriák fölötti S kifejezések összegyűjtik az összetevőikben jelen levő kvantorokat. S_3 kvantortárolójában megtalálható NP_3 és NP_4 kvantorai: Q_3 és Q_4; S_2-ben NP_2 kvantora, Q_2, valamint S_3 összegyűjtött kvantorai, $\{Q_3, Q_4\}$; S_1 QSTORE-ja pedig a következő: $\{Q_1, Q_2, Q_3, Q_4\}$.

A logikai kifejezés magja az igei fejből származik, valamint az igén és az őt dominálló kifejezéseken jelöljük, hogy a kifejezésekben szereplő kvantorok milyen sorrendben alkalmazandóak a magra. Ez az igei fejű frázisok QUANTS listáján van megadva, tehát:

(3) Egy igei fejű S frázis esetén az összetevők QSTORE halmazában meglevő kvantorok vagy az S QSTORE-jában jelennek meg, vagy az S QUANTS listájának az elején (a QUANTS lista további része az S által közvetlenül dominált igei fejű összetevő QUANTS listájával azonos).

Az így kialakult mondatszerkezet esetén tehát – üres QSTORE halmazt feltételezve – a QUANTS lista megadja a kvantorok hatóköri sorrendjét.

A fent leírt módszer a HPSG általános kvantorértelmezési módszere, segítségével a kötött szórendű, konfigurációs nyelvek esetében is meg tudjuk magyarázni a hatóköri többértelműséget. A magyarban azonban, mint azt az (2) példák is mutatják, csak az ige utáni kvantifikált kifejezések hatóköre lehet szabad, az ige előtti kvantifikált kifejezések hatóköre egymáshoz képest kötött, az (1) példák szerint a kifejezések sorrendje meghatározza a hatóköri sorrendet. [7], [8], és [9] szerint a magyarban csak az ige utáni, komplementumpozícióból származó kvantorokra vonatkozik a (3) szabály, az igét megelőző, azaz filler-pozíciókból származó kvantorokra a (4) kiegészítő szabály is vonatkozik:

(4) Ha egy igei fejű S frázisnak van ige előtti, azaz filler-összetevője, akkor annak a QSTORE-jában megtalálható kvantorok nem jelenhetnek meg az S QSTORE-jában.

Az 1. ábrán látható szerkezetben így az ige előtti NP_1 és NP_2 összetevőkből származó Q_1, illetve Q_2 kvantorok nem az őket dominálló S_1, illetve S_2 frázisok QSTORE halmazában jelennek meg (4 szabály), hanem a megfelelő QUANTS listák élén (3 szabály). Mivel azonban S_1 QUANTS listájának a további része S_2 QUANTS listájával egyezik meg, amelynek viszont Q_2 volt az első eleme, a Q_1 kvantor mindig nagyobb hatókörű lesz, mint a Q_2 kvantor, vagyis az ige előtti kvantifikált összetevők sorrendje megegyezik a hatóköri sorrenddel. Az ige utáni kifejezésekből származó Q_3 és Q_4 kvantorokra viszont nem vonatkozik a (4) kiegészítő szabály, azok bármely S kifejezésnél átkerülhetnek a QUANTS listára, vagy tovább másolódhatnak a QSTORE kvantortárolóba.

3 Az implementáció

Az előző fejezetben ismertetett elméleti elemzés ellenőrzéseként szükséges a gyakorlatba is átültetni a megoldási javaslatot. Az elemzés nagyban épít a HPSG elméleti keretre. Létezik ugyan, és el is érhető a HPSG-nek számítógépes implementációja ([5]), azonban az egy fontos szempontból nem bizonyul kielégítőnek: nem tudja kezelni a magyar nyelvre jellemző szabad szórendűséget. Ezért arra vállalkoztam, hogy egy alapjaitól újra felépített elemző megalkotására teszek kísérletet. Ez, bár nem telje-

sen követi hűen a HPSG formalizmusát, szellemében megfelel annak, és lehetőséget nyújt arra, hogy egy jobban, pontosabban kidolgozott implementáció része, alapja legyen.

Mivel a jelenség elemzése frázisstruktúra-nyelvtannal történt, az alkalmazás Prolog nyelven történt, ahol a beépített DCG formalizmus nagy segítséget nyújt a frázisstruktúra nyelvtanok megfogalmazására.

Az alkalmazás több modulból álló nyelvtant feltételez. Az első modul a lexikai egységek lexikaiegység-specifikus tulajdonságait adja meg, úgymint hangalak, jelentés, ragozási paradigma stb. Ezekből építi fel a következő modul a tényleges alap lexikai egységeket, specifikálva az előző egység által csak jelzett tulajdonságokat – itt derül ki például, hogy egy tranzitív igének pontosan milyen vonzatszerkezete van. A harmadik modul a lexikai szabályokat tartalmazza, amelyek egy alap lexikai egység variánsait adják meg. A negyedik modulban találhatóak a tényleges szintaktikai/grammatikai szabályok, amelyekkel összeállíthatjuk a frázisokat, az összeállítással párhuzamosan azok szemantikai leírását is megadva. Ezzel a nyelvtan nemcsak elemezni képes magyar nyelvű mondatokat, hanem a mondatok jelentésreprezentációja is előáll. Ennek a jelentésreprezentációnak az olvashatóbb, predikátumlogikai formájúra átalakítását egy további modul végzi. Ez a modul teljes egészében a [1]-ben ismertetett megoldással azonos, amely elérhető [2]-n. A lexikaiegység-specifikus tulajdonságokat tartalmazó modul szintén [1] szellemében épült fel, bár nyelvspecifikussága miatt nyilvánvalóan nem változatlan átvétele annak.

3.1 A lexikaiegység-specifikus tulajdonságok

A lexikaiegység-specifikus tulajdonságokat tartalmazó modulban a lexikai egységeknek azon tulajdonságai, amelyek tipikusnak mondhatóak, csak jelzésszerűen vannak megadva, ilyen például a következő *minden* determináns esetében a szófaj: `det`. Azok a tulajdonságaik, azonban, amelyek egyediek teljes részletességükben, ahogyan ez a szemantikai leírásnál is látható.

```
lexentry(
  det,
  [def(indef), word([minden]), index(I),
   sem(lam(S,lam(Q,all(I,imp(app(S,I),app(Q,I)))))])).
```

Ugyanez a *kibékül* tranzitív igénél a következőképpen alakul. A szó igei kategóriájú, azon belül is tranzitív, mégpedig olyan, amelyiknek a második argumentuma *-val/vel* esetű kell hogy legyen (`tv2`), csakúgy, mint például a *találkozik* vagy a *megismerkedik* ige. A jelentésleírásában osztozik a tranzitív igékkel, mindegyik ugyanolyan séma alapján épül fel, csak a predikátum változik benne (`symbol(kibékül)`).

```
lexentry(
  tv2,
  [fin(fin), word([kibékül]), symbol(kibékül),
   agr(sg,3,indef)]).
```

3.2 Az alap lexikai egységek

A lexikai egységek a `lexentry` definíciók adatainak a felhasználásával állnak össze:

```
lex(
  synsem(
    cat(...),
    content(...)    ),
  qStore([bo(app(SemDet,SemN),I)]),
  slash([]))
  -->
    {lexentry(det,
      [def(Def),word(Word),index(I),sem(SemDet)])},
    Word.
```

A determinánsok (amelyeknek a szintaktikai (`cat`) és szemantikai (`content`) tulajdonságainak a részletezésétől eltekintek) `qStore` listáján egyetlen elem található, a determinánssal kezdődő főnévi csoport kvantorának a leírása. A `slash` lista leírása a lexikai szabályoknál lesz megtalálható.

A hatókör-értelmezés szempontjából érdekes még az igék szerkezete:

```
lex(
  synsem(
    cat(
      head(v(Fin)),
      comps([
        synsem(
         cat(head(n(nom)),args(_),deps(_),comps([]),_),
         content(agr(Num,Per,_),index(I1),restr(_))),
        synsem(
         cat(head(n(ins)),args(_),deps(_),comps([]),_),
         content(agr(_,_,_),index(I2),restr(_)))])),
    content(
      agr(Num,Per,Def),
      quants([]),
      nucleus(Sem))),
  qStore([]),
  slash([]))
  -->
  {lexentry(tv2,[fin(Fin),word(Word),symbol(Sym),
      agr(Num,Per,Def)])},
  Word.
```

Itt a `comps` lista tartalmazza a `tv2` típusú igék argumentumszerkezetét, ezen lista alapján tudjuk majd ellenőrizni a szintaktikai szabályoknál, hogy a mondatban megjelenő komplementumok megfelelőek-e az őket vonzó ige számára. A `nucleus` adja meg az ige jelentését, ami egyúttal a mondat magjelentése. A `quants` lista a magra alkalmazandó kvantorok sorrendjét, vagyis a kvantorok hatóköri sorrendjét tartalmaz-

za. Ez a lista üres az igék lexikai leírásánál, csakúgy, mint a `qStore` és a `slash` lista is.

3.3 Lexikai szabályok

A nyelvtan jelen pillanatban csak egyetlen lexikai szabályt tartalmaz.

A HPSG-ben az összetevős szerkezetek kialakításának két módja van. Az egyik az, amikor a szerkezet egyik összetevője, a szerkezet feje meghatározza, hogy milyen más összetevők, azaz komplementumok lehetnek még a szerkezetben. A fej lexikai leírásában szerepel a `comps` lista, amely a komplementumokat sorolja fel. Amikor egy komplementum összecsatlakozik a fejjel, akkor a komplementum unifikálódik a `comps` lista egyik elemével. A lista tehát azoknak az összetevőknek a leírását tartalmazza, amelyek még hiányoznak a fej mellől ahhoz, hogy teljes frázist – mondatot, főnévi csoportot stb. – kapjunk. Ha egy frázis tehát ilyen fej-komplementum szerkezetű, akkor a fej `comps` listája tartalmazza a komplementumot, a frázis `comps` listájáról azonban már hiányzik.

A másik frázisalkotási mód az olyan hiányos kifejezéseknek a hiányait szünteti meg, mint amilyen az elliptikus mondat, a kérdőszó-kiemeléses mondat vagy a datívuszi birtokos kimozgatásával hátra maradt hiányos főnévi csoport. Az ilyen jellegű hiányokat a kifejezések `slash` listái tárolják. Akkor jelenik meg egy kifejezés valaminek a `slash` listáján, ha az a kifejezés az elvárt komplementumpozíciójától távol kerül majd elő. A mondatszerkezet alján a listán megjelenő kifejezések a mondatszerkezetben fölfelé összegyűlnek, majd egy bizonyos ponton *filler* összetevőkként jelennek meg. A mi elemzésünk szempontjából ilyen filler összetevők az igét megelőző pozícióban található kvantifikált kifejezések.

Mivel egy kifejezés nem lehet egyszerre komplementum és filler is, a következő lexikai szabály az alap lexikai leírásban szereplő `comps` listát kettéválasztja valóban komplementumként megjelenő elemekre és filler összetevőként megjelenő elemekre, így egy új lexikai egységet hoz létre, ami az eredetinek egy argumentumszerkezeti variánsa:

```
sign(
  synsem(
    cat(head(v(fin)), comps(Comps)),
    Content),
  QStore,
  slash(Slash))
  -->
  lex(
    synsem(
      cat(head(v(fin)), comps(CompsHead)),
      Content),
    QStore,
    slash([])),
  {shuffle(Slash,Comps,CompsHead)}.
```

A szabályban szereplő `shuffle` predikátum a `Slash` és a `Comps` lista elemeit csúsztatja össze oly módon, hogy az eredeti listák elemeinek egymáshoz viszonyított sorrendje ne változzon – mint amikor két pakli kártyát csúsztatunk össze.

3.4 Szintaktikai szabályok

A kvantifikált kifejezések hatókörének a meghatározásához szükséges a kifejezések mondatban elfoglalt pozíciójának meghatározása, úgyhogy elsődlegesen a tényleges mondatelemzéshez szükséges szabályokat vizsgáljuk meg, a megfelelő pontokon rámutatva, hogy a kvantorok hatókör-értelmezésénél az adott ponton milyen részletek játszanak szerepet.

A magyar mondatok szerkezete az 1. ábrán bemutatottak szerint két fő részből áll. Az egyik az igét és az őt követő mondatszakasz összetevőit tartalmazza, és mindegyik összetevő a lexikai ige testvére.

Az igét követő összetevők az ige komplementumai. Ebben a mondatszakaszban az összetevők sorrendje szabad, jelentéskülönbséget (és hatóköri különbséget) nem okoz az összetevők felcserélése. Az igei fejű, lapos, szabad komplementumsorrendű szerkezetet a `sign2` kategória generálásával hozzuk létre:

```
sign2(
  synsem(
    cat(head(v(fin)), comps(CompsVP)),
    Content),
  qStore(QStoreVP),
  Slash)
  -->
  {shuffle([SynsemArg], CompsVP, CompsHead)},
  sign2(
    synsem(
      cat(head(v(fin)), comps(CompsHead)),
      Content),
    qStore(QStoreV),
    Slash),
  sign(SynsemArg,qStore(QStoreArg),_),
  {append(QStoreArg,QStoreV,QStoreVP)}.
```

`sign2` rekurzívan előállítható egy igei fejből és az igei fej egy véletlenül kiválasztott komplementumából, és az eredményül kapott kifejezés `comps` listája eggyel rövidebb, mint az ő igei fejéé: `shuffle([SynsemArg],CompsVP,CompsHead)`. Az így létrehozott kvázi lapos szerkezet generálásakor semmi más nem történik, csak a `comps` lista kiürül, és összegyűlnek a komplementumok `qStore` listáján tárolt kvantorai: `append(QStoreArg,QStoreV,QStoreVP)`.

Az így kapott, üres `comps` listájú igei kifejezés már megfelel az 1. ábra legalsó S kategóriájának:

```
sign(
  synsem(
    cat(head(v(fin)), comps([])),
    content(Agr, quants(QuantsVP), Nucleus)),
  qStore(QStoreVP),
  slash(SlashVP))
  -->

  sign2(
    synsem(
      cat(head(v(fin)), comps([])),
      content(Agr, quants(QuantsV), Nucleus)),
    qStore(QStoreV),
    slash(SlashVP)),
  {quantorRule(QStoreVP,QStoreV,[],QuantsVP,QuantsV)}.
```

Ezen a ponton történhet meg először az eltárolt kvantorok bármelyikének a hatókörének a meghatározása, azaz itt kerülhetnek át elemek a `qStore` halmazból a `quants` listára. Ezt a (3) szabályban leírtaknak megfelelően a `quantorRule` predikátum végzi el:

```
quantorRule(QStoreMother, QStoreHead, QStoreSister,
            QuantsMother, QuantsHead):-
  append(QStoreSister,QStoreHead,Temp1),
  deleteSubList(Temp2,Temp1,QStoreMother),
  append(Temp2,QuantsHead,QuantsMother).
```

A definícióban szereplő `deleteSubList` az első argumentum elemeit törli a második argumentumról, és a maradékot a harmadik argumentumba teszi.

A magyar mondatszerkezet másik fő részében az igét megelőző összetevők egyenként csatlakoznak az előzőekben kialakított, komplementumaival már teljes mértékben kiegészített kifejezéshez:

```
sign(
  synsem(
    cat(head(v(fin)), comps([]),),
    content(Agr, quants(QuantsS), Nucleus)),
  qStore(QStoreS),
  slash(SlashMother))
  -->
  {shuffle([SynsemFiller],SlashMother,SlashHead)},
  sign(SynsemFiller,qStore(QStoreFiller),_SlashFiller),
  sign(
    synsem(
      cat(head(v(fin)), comps([])),
      content(Agr, quants(QuantsVP), Nucleus)),
    qStore(QStoreVP),
    slash(SlashHead)),
  {quantorRule(QStoreS,QStoreVP,QStoreFiller,
               QuantsS,QuantsVP),
   subSet(QStoreFiller,QuantsS)}.
```

A balról csatlakozó filler összetevők a fej `slash` listájáról kerülnek ki egyenként, tetszőleges sorrendben. Az összetevők kvantorai, csakúgy, mint az előző újraíró szabály esetében is, választhatóan kerülhetnek a szülőcsomópontnak a `qStore` halmazába vagy a `quants` listájára. Pontosabban ez az opció csak az igei fejről származó kvantorok számára nyitott, a filler összetevő kvantora kizárólag a quants listára kerülhet: `subSet(QStoreFiller,QuantsS)`. Ez a (4) szabály Prolog-megfelelője.

3.5 A mondat szemantikai tartalmának predikátumlogikai formulává alakítása

A tényleges mondatelemzési folyamat ezzel készen is van, a nyelvtan képes generálni és elemezni a feltételeknek megfelelő magyar mondatokat: szintaktikailag azokat a nyelvi jeleket (`sign`) tekinti mondatnak, amelyiknek a kategóriája ige (`cat(head(v(fin)))`), komplementumai mind szerepelnek a kifejezésben (`comps([])`), és a filler összetevői is megjelentek a bal periférián (`slash([])`). A mondat szemantikai értelmezhetőségéhez még az is szükséges, hogy valamennyi kvantornak meg legyen határozva a hatóköre (`qStore([])`).

A kvantorok hatóköreinek az erőviszonyait, mint azt a 2. szakaszban láthattuk, a kvantorok `quants` listán elfoglalt helye egyértelműen meghatározza. Hogy ezt szemléletesen is belássuk, alakítsuk át a kapott kvantorlistát könnyebben olvasható, predikátumlogikai formulává!

A *Minden kalóz kibékült néhány indiánnal* mondat elemzése után a `nucleus` és a `comps` tartalmazzák a logikai kifejezés magját és a kvantorok listáját, a tág hatókörű olvasat esetén például ez a lista a két elemű, a lista első tagja a *néhány indián* kvantora, a második eleme pedig a *minden kalóz* kvantora. Először egyetlen formulává alakítjuk a magjelentést és a kvantorokat úgy, hogy a kvantorokat a legkisebb hatókörűtől a legnagyobb hatókörű felé haladva egymás után alkalmazzuk a magjelentésre. Ekkor kapunk egy λ-formulát:

```
app(app(lam(_G298,lam(_G301,exist(_G304,
and(app(_G298,_G304),app(_G301,_G304))))),lam(_G276,
indián(_G276))),lam(_G276,app(app(lam(_G116,lam(_G119,
all(_G122,imp(app(_G116,_G122),app(_G119,_G122))))),
lam(_G72,kalóz(_G72))),lam(_G72,kibékül(_G72,_G276)))))
```

Ugyanez konvencionális formában (a @ a függvényalkalmazás jele):

(5) ((λR.λS.∃v(R@v ∧ S@v) @ λy.indián(y)) @ λy.((λP.λQ.∀w((P@w) → (Q@w)) @ λx.kalóz(x)) @ λx.kibékül(x,y)))

Ezen végrehajtva az [1]-ben használt, [2]-ben elérhető β-konverziót, megkapjuk a szokásos elsőrendű formulát:

```
exist(_G304,and(indián(_G304),all(_G999,
imp(kalóz(_G999),kibékül(_G999,_G304)))))
```

Ugyanez konvencionális formában:

(6) $\exists y\,(\text{indián}(y) \wedge \forall x\,(\text{kalóz}(x) \rightarrow \text{kibékül}(x,y))$

(6) logikailag ekvivalens (2c)-vel. A Prolog-implementáció megadja a szűk hatókörű olvasatot is, amely a szükséges konverziókkal (2b)-vel ekvivalens formulává alakítható. Az elméleti megoldás számítógépes implementációja tehát helyesen működik, képes megadni az elvárt hatóköri többértelműséget.

4 További lehetőségek

Az implementáció, mivel egy kidolgozott elméletre, a HPSG-re alapul, kibővíthető további grammatikai szabályokkal, amelyek például szabályozhatják, hogy az ige előtt pontosan milyen elemek és hol jelenhetnek meg, gondolva itt a fókuszértelmezésre és a topikalizációra. A már meglevő implementációrészek azonban ebben a kibővített elemzőben is megfelelően működnek.

További bővíthetősége az implementációnak, hogy a rendszer az [1]-ben bemutatott elemekkel kiegészítve az elsőrendű logikai kifejezések alapján képes egy mondatot interpretálni egy megadott világmodellben, vagyis egy olyan lekérdező rendszert készíthetünk, amelyben a kérdések természetes nyelven vannak megfogalmazva.

Bibliográfia

1. Blackburn, P., Bos, J.: Representation and Inference for Natural Language: A First Course in Computational Semantics. CSLI Press (2005)
2. Blackburn, P., Bos, J.: Representation and Inference for Natural Language: Software Requirements and Downloads: http://homepages.inf.ed.ac.uk/jbos/comsem/software1.html
3. Cooper, R.: Quantification and Syntactic Theory. Reidel, Dordrecht (1983)
4. É. Kiss, K.: Configurationality in Hungarian. Akadémiai Kiadó, Budapest (1987)
5. Penn, G.: The ALE Homepage: http://www.cs.toronto.edu/~gpenn/ale.html
6. Pollard, C., Sag, I A.: Head-Driven Phrase Structure Grammar. CSLI – University of Chicago Press, Stanford – Chicago (1994)
7. Szécsényi T.: Sorrend és hatókör a magyarban: HPSG elemzés. Nyelvtudomány Vol.1 (2005) 171–205
8. Szécsényi T.: Lokalitás és argumentumöröklés. A magyar infinitívuszi szerkezetek leírása HPSG keretben. Doktori értekezés. Szeged, SZTE (2009)
9. Szécsényi T.: Magyar mondatszerkezeti jelenségek elemzése HPSG-ben. In: Bartos Huba (szerk.): Általános Nyelvészeti Tanulmányok XXIII (2011) 99–138

VII. Poszterek és laptopos bemutatók

Interaktív formánsérték-módosító fejlesztése

Abari Kálmán[1], Olaszy Gábor[2]

[1] Debreceni Egyetem, Pszichológia Intézet
abari.kalman@arts.unideb.hu
[2] BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Kivonat: A cikkben bemutatjuk egy webalapú interaktív formánsérték-módosító program felépítését és használatát. Az alkalmazás kötött szerkezetben várja a kiinduló formánsértékeket, melyeket egy Flash-ben készült program segítségével tudunk kényelmesen módosítani, azaz hozzáigazítani a hangszínképhez. A kiindulási és módosított értékeket is MySQL adatbázisban tároljuk, melyek fel- és letöltésről külön funkció gondoskodik. A formánsmódosítás során használt hangszínkép megjelenítéséhez a WAV formátumú hangfájlok feltöltése is szükséges. A fejlesztést a magyar formánsadatbázis készítése és továbbfejlesztése ihlette.

1 Bevezetés

Az elmúlt 2 évben már bemutattuk az első magyar formáns adatbázist, amely a BME Távközlési és Médiainformatikai tanszékén kezdeményezett félautomatikus formáns-elemző eljáráson alapul [1,2,3]. A formánsmeghatározáshoz használt szóadatbázis a következő adatokat tartalmazza minden szóra: ortografikus szöveg, fonetikai átirat, a szó hullámformája (férfi és női ejtésben), hanghatár-jelölések a hullámformában és a mért formánsok. Az adatbázis szabadon hozzáférhető, webalapú keresőfelülettel rendelkezik (*http://magyarbeszed.tmit.bme.hu/formans*). A teljes formánsadatbázisban közel 3000 szó és összesen 10 391 magánhangzó szerepel. Egy magánhangzón belül 3 mérési pontot jelöltünk ki: a teljes hang időtartamának 25, 50 és 75%-os pontját. Kivételt képeztek a kezdő és befejező magánhangzók, ahol csak két mérési pontot vettünk fel: kezdőhöz 50% és 75%, befejezőhöz 25% és 50%.

A formánsadatbázis létrehozása során hozzávetőleg a magánhangzók negyedében volt szükség a formánsértékek kézi korrekciójára. Már ekkor felmerült, hogy szükség lenne egy interaktív formánsérték-módosító eszközre, amely a grafikus felhasználói felület előnyeit kihasználva, kényelmes formánsérték-leolvasást tesz lehetővé a szó színképe alapján, és így az esetleges korrekciók is rugalmasabban megoldhatók. Jelen cikkben ennek az eszköznek egy továbbgondolásáról számolunk be, amely megnyitja az utat további formánsadatbázisok készítése előtt azzal, hogy lehetővé teszi tetszőleges beszédadatbázisból származó – többnyire automatikus módszerekkel meghatározott – formánsértékek egyszerű, vizuális alapú kézi javítását.

2 Az interaktív formánsérték-módosító felépítése

Az interaktív formánsérték-módosító eszköz egy szabadon hozzáférhető webes alkalmazás, melynek fő komponensei a MySQL adatbázis, a PHP/HTML forráskódú állományok és a Flash-ben készült „animáció”. Egyelőre az alkalmazás béta verziója készült el, várhatóan az év végére az alkalmazás minden funkciója elérhető lesz a *http://magyarbeszed.tmit.bme.hu/ifem* címen.

A használat szempontjából az alkalmazás 3 fő részt tartalmaz: (1) a formánsadatbázis-feltöltőt, (2) a formánsértékeket módosító Flash alkalmazást és (3) a javított beszédadatbázist eltároló modult. A következőben ezeket tekintjük át részletesebben.

2.1 A formánsadatbázis feltöltése

A formánsmódosító programunk a saját adatbázisába feltöltött formánsfrekvencia értékeket ajánlja fel korrigálásra. Ezt a beszédadatbázist nevezzük a továbbiakban *formánsadatbázisnak*, mely alapvetően címkézési adatokat és hangfelvételeket tartalmaz. A formánsadatbázisba feltöltendő adatok forrása egy ún. *nyers formánsadatbázis*, mely legtöbbször valamilyen automatikus formánsmeghatározó algoritmus segítségével áll elő. A nyers formánsadatbázisból kell előállítanunk a feltöltéshez szükséges két állományt: (1) egy kötött szerkezetű, tabulátorral tagolt szöveges állományt és (2) a bemondásokat tartalmazó WAV fájlok (ajánlott 22 kHz, 16 bit, de nem követelmény) tömörített állományát.

A nyers formánsadatbázisban a bemondások alapegysége lehet szó, de a szónál kisebb (akár egy magánhangzó) vagy szónál nagyobb nyelvi egység is, erre nézve nincs megkötés a feltöltés szempontjából. A címkézéssel kapcsolatos adatokkal szemben azonban elvárás, hogy álljon rendelkezésre minden bemondáshoz (1) az ortografikus szöveg, (2) a fonetikus átirat, (3) a hanghatárok és (4) valamilyen formánsmeghatározó algoritmussal megmért idő- és formánsfrekvencia-érték párok halmaza.

2.1.1 A tagolt szöveges állomány előkészítése

Feltöltés előtt a rendelkezésre álló – tetszőleges nyers formánsadatbázisból származó – adatainkat konvertálni kell egy tabulátorral elválasztott szöveges állományba. A tagolt szöveges állomány minden sora egy-egy beszédhangra vonatkozó információt tartalmaz. Ezek tipikusan magánhangzók vagy zöngés mássalhangzók lesznek, de formailag erre semmilyen megkötés nincs.

A tagolt szöveges állomány kötelezően tartalmaz fejlécet, azaz az első sor az oszlopnevek tabulátorral elválasztott listája lesz. Javasoljuk a következő oszlopnevek használatát:

WAV	HANGSORSZAM	FORMANS	HANGHATAROK	BETUSOR	HANGSOR	BESZELO	ID_2

A második sortól kezdve azoknak a beszédhangoknak az adatai következnek az oszlopnevek fenti sorrendjében, amelyeknek a formánsértékeit szeretnénk vizuálisan

ellenőrizni, esetleg manuálisan módosítani. A WAV mezőbe a hangfájl neve kerül, elérési út nélkül. Feltételezzük, hogy két azonos nevű hangfájl nem fordul elő a szöveges állományban. A HANGSORSZAM mező a hangállományban tárolt bemondás szegmentáltságát tételezi fel, és annak a hangnak a sorszámát tartalmazza, amelynek a formánsait vizsgáljuk, illetve módosítani akarjuk. Egy hangfájlhoz (azaz WAV állományhoz) természetesen több hangsorszám mező is tartozhat, ebben az esetben ez a szöveges állományban új sorként fog megjelenni. Új sorban meg kell ismételni a WAV fájl nevét és a HANGSORSZAM mezőbe a következő, formánsmódosításban részt vevő hang sorszámát kell megadnunk. A FORMANS mezőbe a beszédhang iniciális formánsértékei kerülnek, pl.:

0.103:746;0.122:788;0.1538:810;0.1856:759;0.2047:647@0.103:1359; 0.122:1382; 0.1538:1380; 0.1856:1398; 0.2047:1364@0.103:2698; 0.122:2670; 0.1538:2700; 0.1856:2532; 0.2047:2464@0.103:2900; 0.122:3363; 0.1538:3299; 0.1856:3455; 0.2047:3455

A fent felsorolt összes idő- és formánsérték egy adott beszédhangra vonatkozik (tehát sortörés nélkül egy sorba kellett volna őket írnunk). Először az F1, majd az F2, F3, F4 értékei következnek. Az egyes formánsokat kukac (@) karakterrel választjuk szét, a formánson belüli, időben elkülönülő méréseket pontosvessző tagolja. Egy méréshez két adatra van szükség egy időkoordinátára (s, másodperc) és egy frekvenciaértékre (Hz). Ezt a két koordinátát kettőspont (:) választja el. A fenti példában minden formánsra 5 mérési pontot adtunk meg és minden formáns esetén azonos időkoordinátákat használtunk (0.103 s, 0.122s, 0.1538 s, 0.1856 s és 0.2047 s). Most ezek a beszédhang hosszának 10, 25, 50, 75 és 90%-os pontjait jelentik, így most rendelkezésre áll 5 olyan mérési hely, amely a teljes hangot lefedi. Mivel abszolút időértékeket kell megadnunk a FORMANS mezőben, a mérési helyek rendszerének kialakítása tetszőleges lehet. A formánsérték módosító alkalmazásunkban az összes itt tárolt frekvenciaértéket meg tudjuk változtatni.

A következő négy mező a WAV állományban tárolt bemondást jellemzi és nem a sort meghatározó beszédhangot. Ennek megfelelően értékük minden olyan sorban azonos lesz, ahol a WAV mezőben is azonos érték található. Ez redundáns tárolást jelent, de így egyszerűbb, könnyebben kezelhető szerkezetet kapunk. A HANGHATAROK mező pontosvesszővel elválasztva tartalmazza a teljes bemondás másodpercben (s) meghatározott hanghatár-jelölőinek időkoordinátáját. A BETUSOR mezőbe a bemondás ortografikus szövege kerül, tagolás nélkül. A HANGSOR mező pontosvesszővel elválasztva a fonetikai átiratot tárolja. Itt tetszőleges jelölést használhatunk, bármit, amit az ASCII szöveges állomány tárolni enged (pl. TMIT, SAMPA). A BESZELO mezőbe a hangfelvételt adó személyről adhatunk egy leírást (pl. azonosító, neme stb).

Az ID_2 mezőbe egy tetszőleges karaktersorozatot szerepeltethetünk, amely az adatok visszatöltését segíti a saját adatbázisunkba a formánsmódosítás után. Ez a mező tipikusan a forrás adatbázis valamilyen azonosítóját tartalmazza, mely vonatkozhat bemondásra vagy akár beszédhangra is. Szerepe egyértelműen a formánsmódosítás eredményének egyszerű visszavezetése a korábban használt adatbázisunkba.

Amennyiben előállítottuk a tabulátorral tagolt szöveges állományt, érdemes néhány ellenőrzést elvégezni. (1) A WAV és HANGSORSZAM mezők együtt egyértelműen azonosítják a szöveges állomány sorait (azaz elsődleges kulcsok). (2) Azok-

ban a sorokban, ahol a WAV értéke megegyezik, ott a HANGHATAROK, a BETUSOR, a HANGSOR és a BESZELO mezők értéke is megegyezik. (3) A HANGHATAROK mezőben a pontosvesszővel elválasztott értékek száma mindig eggyel több, mint az ugyanazon sor HANGSOR mezőben lévő pontosvesszővel elválasztott elemeinek száma. (4) A HANGSORSZAM mező a HANGSOR pontosvesszővel elválasztott elemeinek valamelyikének sorszámát tartalmazza (1-től induló sorszámmal).

A fenti ellenőrzéseket, néhány további kíséretében, maga a program is elvégzi, miközben a szöveges állományt eltárolja az adatbázisban. A weboldalon ez egy egyszerű állománykiválasztást követően automatikusan végbemegy, az esetleges hibák, illetve a feltöltött sorok statisztikája szintén a weboldalon követhető nyomon.

2.1.2 A WAV állományok feltöltése

A formánsértékek kézi módosításnak alapja a hangszínkép. Ezek létrehozásához a bemondásokra is szükség van. Az összes WAV állományt gyűjtsük össze egy könyvtárba, majd csomagoljuk őket össze ZIP tömörítővel. Az összecsomagolt állományt a weboldal megfelelő funkciójának kiválasztásával tölthetjük fel a szerverre. A hangszínképek létrehozása után a WAV állományok a szerverről automatikusan törlődnek, a továbbiakban nincs szerepük.

A formánsadatbázis két komponensének (tagolt szöveges állomány, ZIP fájl) feltöltése után a weboldalon tájékoztatást kapunk a MySQL adatbázisba felmásolt formánsérték adatokról (különböző bemondások száma, a formánsaiban módosítható beszédhangok száma, beszédhangonként a formánsok száma, illetve a mérési pontok száma formánsonként). Az elkészült hangszínképekről is kapunk egy statisztikát, ellenőrizhetjük, hogy az adatbázisunk minden bemondásához elkészült-e a hangszínkép.

2.2 A formánsértékek módosítása

Az adatbázisba kiindulásképpen feltöltött formánsértékek módosítását egy Flash alkalmazás végzi el. Megmutatja a bemondás hangszínképét és pontokkal jelzi az adott mérési helyeken a nyers formánsadatbázisból származó, korábban meghatározott formánsértékeket. Az azonos formánshoz tartozó, de különböző mérési pozíciókban megjelenő pontokat egyenes vonal köti össze. Az 1. ábra a *lábmelegítő* szó (mint bemondási egység) elejének formánsmódosítását szemlélteti. A kép tetején lévő szürke görgetősáv tájékoztat minket, hogy a képernyőn nem látjuk a teljes bemondást. A görgetősáv alatt TMIT hangjelöléssel a bemondás fonetikai átiratát láthatjuk, mely a HANGSOR mezőből származik. Az éppen formánsmódosítás alatt álló beszédhang szimbólumát halványkék háttérszínnel jelzi a program.

A Flash alkalmazás legnagyobb részét a hangszínkép teszi ki. A hangszínképeket az R statisztikai program [4] *seewave* [5] csomagjával készítettük, és a HANGHATAROK adatbázismező segítségével rajzoltuk meg a hanghatár jelölő függőleges vonalakat. A frekvencia tengelyt 0-5000 Hz-ig jelenítjük meg. A hangszínkép mint képállomány magasságát figyelembe véve (347 pixel), a formánsértékek módosításának pontossága 14 Hz, azaz egyetlen pixelnyi mozgatás az *y* tengely men-

tén kb. 14 Hz-et jelent a frekvenciatengelyen. Az időtengely mentén egyszerre kb. 0,8 másodpercet láthatunk a bemondásból 540 pixel széles területen. Egy rövid, 50 ms-os magánhangzóra ekkor kb. 34 pixelnyi széles terület jut. A formánsméréseket reprezentáló piros pontok szélessége 6 pixel, így maximum 5 mérési helyhez tartozó pontot tudunk egyszerre úgy megjeleníteni, hogy az a kézi módosítás során ne legyen zavaró. A most felsorolt, megjelenítésből adódó korlátozásokat a program használata előtt vegyük figyelembe, a mérési pontok számát és a módosított adatokból levonható következtetéseket ez alapján határozzuk meg!

A hangszínkép alatt szöveges mezőket láthatunk, amelyekben az éppen módosított formánsfrekvencia érték jellemzőit olvashatjuk: formánssorszám (pl. F4), időkoordináta és frekvenciakoordináta. A bemondó személyéről is kapunk tájékoztatást a BEMONDO adatbázismező alapján.

A formánsfrekvenciák módosítását billentyűzet segítségével végezhetjük el. Egy szokásos munkamenet a weboldalon megjelenő Flash alkalmazással a következő lehet:

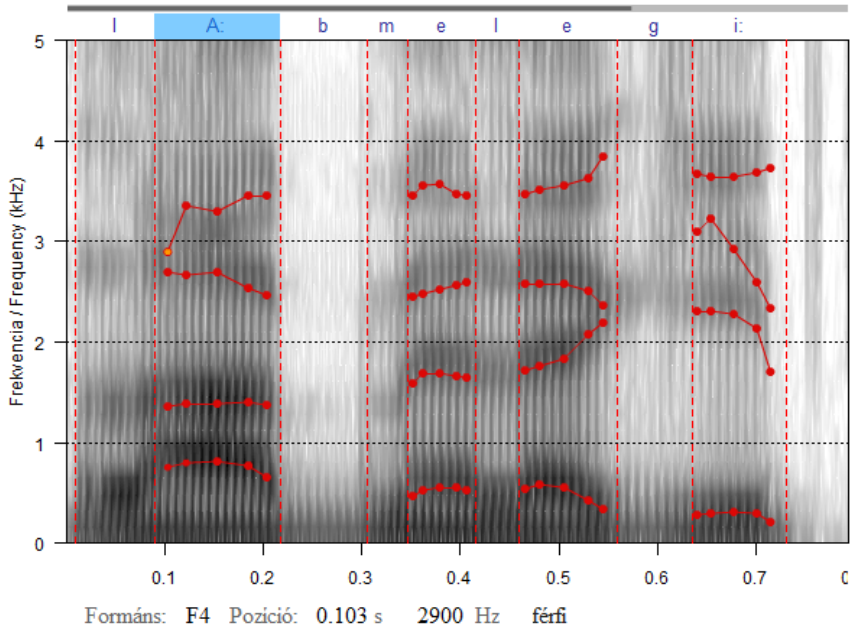
1. Egérrel kattintunk Flash alkalmazás területén, pl. a hangszínképen. Ezzel aktiváljuk a programot, amely most már fogadja billentyűparancsainkat.
2. Eldöntjük, hogy a szó mely beszédhangját szeretnénk vizsgálni, módosítani. A hangok közötti választást a Ctrl+JOBBRA NYÍL és a Ctrl+BALRA NYÍL segíti. A hangok közötti mozgás a hangszínkép görgetését is maga után vonhatja, amit a felső görgetősávon követhetünk nyomon. A hangok közötti váltásnál a program biztosítja, hogy a vizsgált hang környezetét is láthassuk.
3. A magánhangzón belül a módosítandó formáns kiválasztására a kurzormozgató nyilakat használhatjuk (LE NYÍL, FEL NYÍL, BALRA NYÍL, JOBBRA NYÍL). Az aktuális pontot eltérő színezés különbözteti meg a többi ponttól. A pontok közötti mozgás hatása az alsó információs mezőkben is nyomon követhető.
4. Az aktuális pont – és így a formánsérték – mozgatására a Q és A billentyűket használhatjuk. A Q-val növeljük az A-val csökkentjük a formánsértéket. Az információs mezőben ezt is követhetjük.
5. A módosítások mentésére az ENTER billentyűt használjuk. Ez azonnal az adatbázisba rögzíti a módosításokat.

Összefoglalva a Flash alkalmazásban használatos billentyűparancsok:

Ctrl+JOBBRA NYÍL és a Ctrl+BALRA NYÍL: az aktuális beszédhang kiválasztása, a hangszínkép vízszintes görgetése
LE NYÍL, FEL NYÍL, BALRA NYÍL, JOBBRA NYÍL: az aktuális pont kiválasztása az aktuális hangon belül
Q és A billentyűk: a pont mozgatása fel és le
ENTER: a változtatások mentése.

2.3 A javított beszédadatbázis mentése

A formánsmódosítás elvégzése után a javított adatokat tartalmazó tabulátorral tagolt szöveges állomány mentése következik. A mentés során letöltött adatok mindenben megegyeznek a feltöltés során használt adatbázissal, kivéve, hogy az kiegészül a korrigált formánsértékeket tartalmazó FORMANS_JAV oszloppal. Ez hasonló szerkezetben tárolja az idő- és formánsfrekvencia-értékeket, de természetesen már a korrigált adatokat tartalmazza.



1. ábra. Az interaktív formánsmódosító Flash alkalmazás képe a vizuális megfigyeléshez. A *lábmelegítő* szó első magánhangzójában az F4 első (0.103 s) pontban mutatott értékét módosíthatjuk.

3 Összefoglalás

Jelen cikkben egy webalapú formánsérték-módosító program felépítését mutattuk be. Az automatikusan meghatározott formánsfrekvencia értékek kézi módosítása a bemondás hangszínképe alapján történik, amelyet szintén az alkalmazás állít elő. Erre a hangszínképre vetíti rá a program az automatikus mérésből származtatott Hz értékeket (kis pontok formájában). Ez adja a vizuális ítélkezés alapját. Amennyiben az automatikusan meghatározott formánsérték kiugróan eltér a hangszínképen leolvashatótól, akkor a mért értéket a hangspektrogram alapján módosítjuk, és ezt eltároljuk a

formánsadatbázisunkban. A Flash alkalmazásban billentyűparancsokkal határozhatjuk meg a módosítás helyét (hang), a hangon belül a formánst és végül a formáns függőleges pozícióját, azaz a formánsfrekvencia értéket. Tetszőleges hosszúságú bemondás (hang, szó, mondat) formánsait tudjuk kezelni, a hangszínkép vízszintesen görgethető. A megjelenítés ennek ellenére rendelkezik korlátokkal, ezeket a mérések megkezdése előtt figyelembe kell venni. Az alkalmazás segíti a nagy pontosságú formánsadatbázisok létrehozását.

Bibliográfia

1. Abari K., Olaszy G.: Magyar formánsadatbázis az interneten. In: Gósy, M. (szerk.): Beszédkutatás. MTA Nyelvtudományi Intézet, Budapest (2011) 73–82
2 Olaszy G., Rácz Zs. Zs., Bartalis M.: Formánsmérések automatizálása, formánsadatbázisok létrehozása. In: Gósy M. (szerk.): Beszédkutatás 2009. MTA Nyelvtudományi .Intézet, Budapest (2009) 134–147
3. Rácz Zs., Abari K. , Olaszy G.: A formant trajectory database of Hungarian vowels. In: Németh G., Olaszy G. (eds.) The Phonetician 97 (2011) 6–13 (http://www. isphs.org)
4. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
5. Sueur, J., Aubin, T., Simonis, C. Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics Vol. 18 (2008) 213–226

Korpuszalapú entrópiamértékek gating- és lexikai döntési kísérletekben

Fazekas Judit[1], Németh Kornél[1], Pléh Csaba[1], Varga Dániel[2]

[1] BME Kognitív Tudományi Tanszék, Budapest, Egry József utca 1.
e-mail: {jfazekas,knemeth,pleh}@cogsci.bme.hu
[2] BME MOKK, Budapest, Egry József utca 1.
e-mail: daniel@mokk.bme.hu

Nagyméretű gyakorisági szótár birtokában lehetőségünk nyílik információelméleti mértékeket definiálni, amelyek olyan kérdéseket formalizálnak, mint például hogy egy adott szó-prefix a korpuszon belül milyen mértékben korlátozza a szó lehetséges befejezéseinek halmazát.

Cikkünkben ezen mértékek felhasználásával megkíséreljük, hogy összefüggést tárjunk fel az emberi morfológiai feldolgozás és szófelismerés teljesítménye, valamit a szóalakok információelméleti struktúrája között.

Cikkünk bővített változatában három olyan kísérlet eredményeit mutatjuk meg, melyek a fenti kérdéseket járják körül szisztematikus módon.

Az első két, gating feladaton [5] alapuló vizsgálat anyagát 60 darab kétszótagú főnév képezte. A 30 gyakori és a 30 ritka szó közül 15-15 korai egyediségi ponttal rendelkezett (*japán*), 15-15 pedig későivel (*cinke*). A varianciaanalízis egyedül a gyakoriságról mutatta ki, hogy szignifikáns hatása van a felismerés hatásfokára.

A második vizsgálatban bevezettünk egy megszorítást, a szófelismerést befolyásoló top-down hatások vizsgálatának céljából. A résztvevők fele a következő instrukciót kapta: "Csak kétszótagú főneveket fog hallani toldalékok nélkül.", a többi kísérleti személy nem kapott semmilyen információt. Mind a gyakoriság, mind pedig a megszorítások hatása kimutatható volt. Az egyediségi pontok hatása csak a gyakori szavaknál volt egyértelmű.

A mérési adatok birtokában az egyértelműségi pont fogalmának korpuszalapú finomítása céljából a Magyar Webkorpuszra épülő morfológiailag elemzett Szószablya Gyakorisági Szótárhoz [3] fordultunk, és a szótár prefix-fájának információelméleti analízisét végeztük el. Ennek során entrópiamértéket vezettünk be a szóalakok prefixein, az alábbi módon: A gyakorisági szótár a magyar nyelv szóalakjain értelmezett valószínűségeloszlást definiál. Egy szó-prefix entrópiáját ezután úgy definiáltuk, mint e valószínűségeloszlásnak a feltételes entrópiáját azon feltétel mellett, hogy a megfigyelt szó az adott prefixszel kezdődik. A feltételes entrópia tehát a fennmaradó bizonytalanságunk mértéke az adott szóról, miután a prefixét a tudomásunkra hozták. Intuitíve, a mérték azt számszerűsíti, hogy mennyire változatos módon fejeződhet be az adott prefix a korpuszunkban.

Megemlítjük, hogy Antal László [2] már 1964-ben felvetette azt a hipotézist, hogy a morfológiailag összetett szavak morfémahatárai statisztikai értelemben összefüggésbe hozhatók azon pontokkal, ahol az így definiált entrópiamérték zu-

han. A Szószablya Gyakorisági Szótáron végzett méréseink igazolták ezt a hipotézist.

Egy adott kapuhoz az ott felvett mérési pontokat három osztályba soroltuk, aszerint, hogy 1. éppen abban a pontban történt meg a felismerés, 2. éppen a következő pontban történt meg a felismerés, illetve 3. egyéb esetek. Azt tapasztaltuk, hogy valamely kaput rögzítve, a prefixek entrópiamértéke szignifikáns mértékben eltér az 1. és 2. kategóriájú adatpontok között, vagyis a felismerést még a kapura kontrollálva is entrópiacsökkenés előzi meg. Ez a jelenség még akkor is fennáll, ha a gyakoriságra és az egyediségi pont helyére mint kétértékű változókra kontrollálunk. Mi ezt a megfigyelést úgy értelmezzük, mint amely demonstrálja, hogy az entrópia szándékainknak megfelelően az egyediségi pont naív fogalmának kvantitatív finomítása. Ez az eredmény összhangban van Moscoso, Kostic és Baayen [4] modelljével.

Nemcsak az entrópia, hanem az entrópia szomszédos kapuk közötti megváltozása is mutatta a fenti jelenséget, annak ellenére, hogy ez egy erősen nemmonoton viselkedést mutató függvény.

Egy következő kísérletünk Pléh és Juhász [6] szófelismerésre vonatkozó vizsgálatainak folytatása volt. Itt rontott szavak azonosítása volt a kísérleti személyek feladata. A szavak egyes vagy többes számúak voltak, tőalakban, vagy a -nak, -ban, -ra ragokkal. A rontás a szótő, a jel, illetve az esetrag valamelyikében történt, és típusukban lehettek magánhangzó-harmónia hibák, vagy a szótőben történő fonémarontások.

A gyakoriságnak és a rontás típusának egyaránt szignifikáns hatása volt az azonosítás pontosságára. A gyakoribb szavakat gyorsabban kategorizálták a kísérleti személyek, de alacsonyabb pontossággal. Erős korreláció volt a rontás pozíciója és a sikeres visszautasítások aránya között; a későbbi rontások gyorsabb és pontosabb visszautasításhoz vezettek.

Gyakorisági szótárunk segítségével korpuszalapú vizsgálatnak is alávetettük ezen mérések kimeneteit. Hipotézisünk az volt, hogy könnyebben felismerhetőek azok a rontások, melyek szokatlan fonéma n-gram kombinációkhoz vezetnek. A hipotézis formalizálásához meghatároztuk a fonéma trigramok gyakoriságait a korpuszunkban, majd metrikánkat úgy definiáltuk, mint a rontás fonéma trigram környezetének gyakorisága arányítva az eredeti, rontatlan fonéma trigram környezet gyakoriságával. Hipotézisünknek megfelelően a sikeres visszautasítás valószínűsége és sebessége egyaránt erős korrelációban volt az így definiált rontásitrigram-metrikával.

Hivatkozások

1. Aitchison, J.: Words in the mind. London, Blackwell (1987)
2. Antal, L.: A formális nyelvi elemzés, Budapest, Gondolat (1964)
3. Kornai, A., Halácsy. P., Nagy. V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Proceedings of the EACL 2006 Workshop on Web as a Corpus (2006)

4. Moscoso, F., Kostic, A., and Baayen, R. H.: Putting the bits together: an information theoretical perspective on morphological processing. Cognition, 94, pp. 1-18 (2004)
5. Grosjean, F.: Spoken word recognition processes and the gating paradigm. In: Attention, Perception, & Psychophysics, Springer (1980)
6. Pléh, Cs., Juhász, L. Processing of multimorphemic words in Hungarian. Acta Linguistica Hungarica, 43, pp. 211-230. (1995)

Automatikusan előállított protoszótárak közzététele

Héja Enikő, Takács Dávid

MTA Nyelvtudományi Intézet
{eheja, takdavid}@nytud.hu

A három éve folyó EFNILEX projekt célja (l. [1]) annak vizsgálata, hogy a modern nyelvtechnológiai eszközök mennyiben alkalmasak a szótárkészítés támogatására. Jelen demonstráció célja, hogy bemutassa az automatikusan előállított prototípus-szótárak (a továbbiakban protoszótárak) lekérdezhető változatát.

A protoszótárak újdonságát az adja, hogy párhuzamos korpuszokon automatikusan, szóillesztéssel állítjuk elő őket. Bár már majdnem két évtizede használnak különféle statisztikai algoritmusokat forrásnyelvi és célnyelvi szópárok kinyerésére, hogy így bővítsék a gépi fordítás bemenetéül szolgáló szótárakat (pl. [2]), érdekes módon a lexikográfusok között a mai napig sem eldöntött kérdés, hogy használhatóak-e a párhuzamos korpuszok emberi felhasználásra készülő szótárak előállítására.

Az így létrejövő szótárak természetesen több ponton is lényegesen különböznek a hagyományos, lexikográfusok által létrehozott szótáraktól. A legfontosabb különbség, hogy a protoszótárak alapstruktúrájában más típusú adatokkal találkozunk: a protoszótárak mikrostruktúrája kevésbé kidolgozott, de a fordítási jelölteken kívül korpuszgyakorisági adatokat, valamint az illesztő algoritmus által kalkulált fordítási valószínűséget ($P(\text{szó}_{\text{cél}}|\text{szó}_{\text{forrás}})$) is tartalmazza. Nagy mennyiségű természetes nyelvi kontextus áll rendelkezésre, valamint könnyen kiszámíthatóak a fordított irányú protoszótár fordítási valószínűségei is ($P(\text{szó}_{\text{forrás}}|\text{szó}_{\text{cél}})$) is. A protoszótár hátránya, hogy utószerkesztési munkálatok hiányában szükségszerűen tartalmaz hibás jelentésmegfeleltetéseket is. Általánosan elmondható, hogy a protoszótár fedése és pontossága fordítottan arányosak: a fent említett paramétereken alapuló szűréssel növelhető a jó fordítási jelöltek aránya, ennek az ára viszont a szótár fedésének a csökkenése.

Célunk egy olyan online felület fejlesztése, amely kiaknázza a módszer előnyeit és minimálisra csökkenti a hátrányait. Fedés és pontosság vonatkozásában ez azt jelenti, hogy a lekérdező felülettel a protoszótárak személyre szabhatóak lesznek: a fedés-pontosság görbe különböző pontjai eltérő felhasználói igényeknek feleltethetőek meg. Például egy kezdő nyelvtanuló esetében az alapszókincsre van szükség, és az is elvárás, hogy a célnyelvi megfelelő a legjobb (legtöbbet használt) fordítás legyen. Ebben az esetben tehát a protoszótárat úgy vágjuk, hogy a gyakoribb szavakat vesszük csak figyelembe mind a forrásnyelvi, mind a célnyelvi oldalon, és a fordítási párok közül is csak azokat, amelyeknek magas a fordítási valószínűsége. Ezzel szemben egy fordító képes a rossz fordítások közül a jót kiszűrni, különösen, ha rendelkezésére állnak a javasolt fordításokat támogató párhuzamos szövegrészletek. Így az ő esetében egy nagyobb lefedettségű, ám alacsonyabb pontosságú protoszótár megfelelő. Ezért követelmény, hogy az online felületen a felhasználó határozhassa meg, hogy a protoszótár melyik szeletével kíván dolgozni.

A protoszótár paramétereinek beállításával határozható meg a szótár mérete. Eddigi kiértékelési eredményeink szolgálhatnak ugyan némi fogódzóul arra nézve, hogy

hogyan érdemes ezeket a paramétereket beállítani, ám ezzel pont a valódi testreszabás lehetőségét veszítjük el: sokkal célszerűbb lehetővé tenni, hogy a felhasználó egyénileg kisérletezhesse ki, melyek a számára optimális paraméterbeállítások.

A ritkán használt fordítások értelmezésénél nyújt segítséget a nagy mennyiségű természetes példamondat, amely a kérdéses fordításra kattintva kilistázható.

A felület kialakításánál célunk, hogy a rendelkezésünkre álló információkat vizuálisan reprezentáljuk. A fordítási jelölteket szófelhőben, illetve grafikonon is megjelenítjük. Az ábrázoláshoz az alábbi változók közül választhatunk: oda- és visszirányú fordítási valószínűség, forrásnyelvi és célnyelvi szó abszolút gyakorisága.

Hipotézisünk szerint ezek mentén a paraméterek mentén a fordítási jelöltek különböző osztályokba sorolhatók, aszerint, hogy milyen szemantikai viszony áll fenn a fordítási pár két tagja között, illetve a fordítási jelöltek jelentése szerint. Például, ha mindkét irányú fordítási valószínűség magas és a gyakoriságok megközelítőleg megegyeznek, a fordítási jelöltek nagy valószínűséggel jól meghatározott, konkrét dolgokra referáló kifejezések lesznek (pl. terminusok, tulajdonnevek). Ezzel szemben, ha az odairányú fordítási valószínűség magas, de a célnyelvi kifejezés sokkal gyakoribb, valószínű, hogy a célnyelvi kifejezés jelentése sokkal általánosabb, illetve a forrásnyelvi kifejezés használata jelölt. Pl. egy magyar-litván párhuzamos tesztkorpuszban a magyar *tüzetes* szó 5-ször fordul elő, míg a litván *įdėmiai* 100-szor úgy, hogy a fordítási valószínűségük magas: 0.76. Valóban, egy angol-litván szótár alapján a litván szó jelentése sokkal általánosabb: *attentively*, *carefully* – 'figyelmesen', 'óvatosan', 'gondosan' jelentései egyaránt lehetnek.

A protoszótárak elérhetőek a http://efnilex.nytud.hu/efnilex alatt.

Bibliográfia

1. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography. In: Proceedings of the LREC2010 Conference, La Valletta, Malta, May (2010) 2798–2805
2. Wu, D.: Learning an English-Chinese Lexicon from a Parallel Corpus. In: Proceedings of AMTA'94 (1994) 206–213

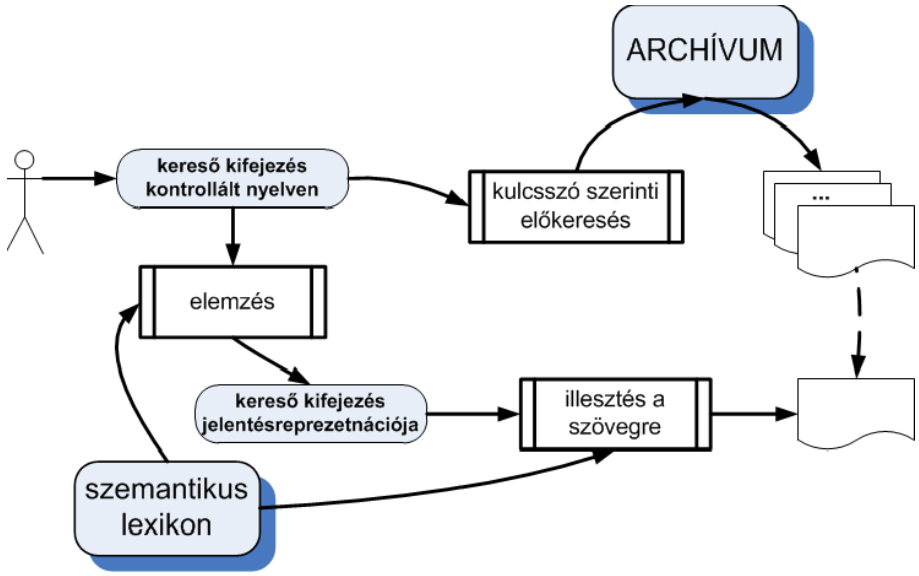
MASZEKER: szemantikus keresőprogram

Hussami Péter[1]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7
hussami@all.hu

A Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtár- és Humán Információtudományi Tanszéke közös projektet (TECH_08_A2/2-2008-0092) indított az Nemzeti Fejlesztési Ügynökség támogatásával. A projekt célja egy olyan, új elveken alapuló integrált keresőrendszer kifejlesztése, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül a keresést végző felhasználó szemantikai kompetenciáját az eddigieknél nagyobb mértékben kiaknázva teszi lehetővé a természetes nyelvi dokumentumtárakban (szövegekben) történő valóban tartalmi keresést. Egyszerűen szólva: a felhasználó jól formált frázisokkal, mondatokkal specifikálhatja, milyen tartalmú dokumentumokat keres.

A 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 keresőkifejezést,
2. szintaktikus és szemantikus elemzés előállítja keresőkifejezés jelentésreprezentációját,
3. szavak szerinti keresés előszűri az archívumot,
4. azokra a szövegszegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a keresőkifejezés jelentésreprezentációját.

Az MSzNy VII konferencián tartott előadáson [1] ismertetésre kerültek a fenti elemek megvalósítására vonatkozó elméleti alapelvek, elsősorban a szemantikus reprezentáció felépítése mint sarokkő köré szervezve. Idén be kívánjuk mutatni a megvalósulás jelenlegi állapotát egy demó prezentálásával.

A demóban az archívumot szabadalmi leírások főigénypontjaiból összeállított dokumentumgyűjtemény alkotja[1]. A felhasználó a kontrollált nyelven megadhat keresőkifejezést. A keresőkifejezés több mondatból, ill. főnévi kifejezésből állhat, a megszorítások az egyértelműséget biztosítják – például korlátozzák az igeneves szerkezeteket. A felsorolásokat 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. 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.

A végleges kiépítéshez képest a demó a következő egyszerűsítéseket alkalmazza:

- a kisméretű „archívum” miatt a kulcsszó szerinti előkeresés felesleges,
- a szemantikus lexikon kiépítettsége még messze van a kívánatostól, ezért a jelentésreprezentációk hiányosak lehetnek,
- a szintaktikus elemzés szemantikus kontrollja még nem teljes,
- a hasonlóság felismerésénél vannak figyelembe nem vett tényezők,
- a szabadalmi igénypontok szerkezetéből és a témakörből adódó heurisztikus megoldásokat kielégítően még nem alkalmaztuk[2],
- a relevancia meghatározása még nem eléggé kifinomult.

Mind a felismerés pontosságát, mind a performanciát a további kísérletek alapján javítani kívánjuk.

Bibliográfia

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 Tudományegyetem, Szeged (2010) 159–167

[1] A projekt egyik kiemelt felhasználási területe a szabadalmi keresés, s a demóban „gyógyhatású készítmények és kozmetikai szerek” témaköréből származó szabadalmakat használunk.

[2] Mind a szintaktikus, mind a szemantikus elemzést, mind a hasonlóság megállapítását nagyban befolyásolja, hogy milyen témakörben, milyen típusú dokumentumok közt keresünk.

Interaktív fonetikai eszköz az artikulációs csatorna keresztmetszet-függvényének meghatározására

Jani Mátyás[1], Björn Lindblom[2], Sten Ternström[3]

[1] Pázmány Péter Katolikus Egyetem, ITK,
Budapest, Práter utca 50/A, e-mail: janma@digitus.itk.ppke.hu
[2] Department of Linguistics, Stockholm University
106 91 Stockholm, Sweden
[3] Department of Speech, Music and Hearing, School of Computer Science and Communication, Kungliga Tekniska Högskolan (Royal Institute of Technology)
100 44 Stockholm

Kivonat A projekt célja annak az eldöntése volt, hogy a SuperCollider programozási környezet mennyire alkalmas egy interaktív artikulációs modell implementálására. Az elkészült szoftver az APEX nevű, kétdimenziós modellt használja, amit az artikulációs csatorna alakja és a formánsok közötti összefüggés vizsgálatára hoztak létre.

Kulcsszavak: artikulációs modell, supercollider, beszédszintézis

1. Bevezetés

Manapság a konkatenatív beszédszintetizálásra használt módszer a legelterjedtebb, annak ellenére, hogy az összefűzéssel készített beszédhang minősége elmarad az artikulációs módszer által elméletileg előállítható beszédhang minőségétől. Emiatt újabban egyre nagyobb figyelmet kap az artikulációs beszédszintetizálás és egyre több artikulációs modell jön létre [1]. Ezen modellek feladata nem mindig a beszédszintetizálás, használhatók kutató és pedagógiai eszközöknek is. Segítségükkel többek között meg lehet figyelni a formáns frekvenciák és az artikulációs csatorna alakja közötti összefüggést. Jelen munka fő célkitűzése egy meglévő kétdimenziós artikulációs modell implementálása, valamint a SuperCollider környezet ilyen jellegű feladatra való használhatóságának kiderítése.

2. APEX modell

Az eredeti APEX program célja adott artikulációból formáns adatok (frekvencia, sávszélesség) kinyerése volt [2]. A modell egy virtuális kétdimenziós artikulációs csatornát használ, ennek geometriáját tesztalanyról készített röntgenképekből nyerték ki. A formáns adatok előállításához több lépésre van szükség. Először

az ajkak, a nyelvcsúcs és nyelv törzs állapotaiból, az állkapocs és a gégefő helyzetéből egy artikulációs profil készül egy mesterséges középvonallal, ami az artikulációs csatorna első és hátsó oldala között félúton helyezkedik el. Ezután le lehet mérni a középvonal mentén tetszőleges pontokban az artikulációs csatorna keresztmetszetét. A keresztmetszetek hosszát egy adott szabály felhasználásával keresztmetszeti területekké kell konvertálni, ez már lényegében az artikulációs csatorna csőmodelljének felel meg. Hangszintézis megvalósításának egyik módja a formánsszintézis, ehhez a csőmodellből ki kell nyerni a formánsparamétereket. Az APEX modell az orrüreget nem modellezi, így a nazális hangokat nem tudja megfelelően szintetizálni.

2.1. Adatok kinyerése

A körvonalak és egyéb geometriai adatok kinyeréséhez röntgenfelvételekre volt szükség [3]. A röntgenfelvételek fő problémája, hogy a tesztalanyokat sugárzás éri és a biztonság érdekében bizonyos biztonsági előírások korlátozzák a felvételek hosszát és az elszenvedett sugárzási mennyiséget. A hangképzőszervek körvonalai 0,5 - 1 mm pontossággal határozhatók meg.

A keresztmetszetek számításához szükséges együtthatók meghatározásához keresztmetszeti MR (mágneses rezonancia) képeket készítettek az artikulációs csatorna mentén több helyen [4]. A felvétel alatt használt szöveganyag svéd magánhangzókat tartalmazott, és az MR képek mellett videó- és hangrögzítés is történt.

2.2. Keresztmetszetek területekké alakítása

A kétdimenziós módszerek közvetlenül csak az artikulációs csatorna oldalnézeti keresztmetszetét tudják felhasználni. A valódi alakzatok nem állnak rendelkezésre, így az artikulációs csatorna irányára merőleges szeletek területét az oldalnézeti keresztmetszethosszakból kell kiszámolni.

Többféleképpen is lehet becsülni ezeket a területeket [5], általában mérésekből adódó együtthatókat felhasználva. A leggyakrabban Heinz és Stevens (1964, 1965) által publikált hatványfüggvényt használják:

$$A = K \cdot d^{\alpha}$$

ahol A az artikulációs csatorna irányára merőleges metszet területe, d a mért hossz, K és α pedig együtthatók, melyek értéke függ a tesztalanyon és a vizsgált metszet pozícióján.

2.3. A nyelv alakjának meghatározása

A nyelv alakjának paramétereit főkomponens-analízis segítségével határozták meg. Körülbelül négyszáz nyelvkörvonalat nyertek ki röntgenképekből, majd

ezeket a körvonalakat 25 pontban mintavételezve tárolták [6]. A főkomponens-analízis eredménye néhány bázisfüggvény súlyozott lineáris kombinációja:

$$V(x) = N(x) + c_1(v) \cdot PC_1(x) + c_1(v) \cdot PC_1(x) + ...$$

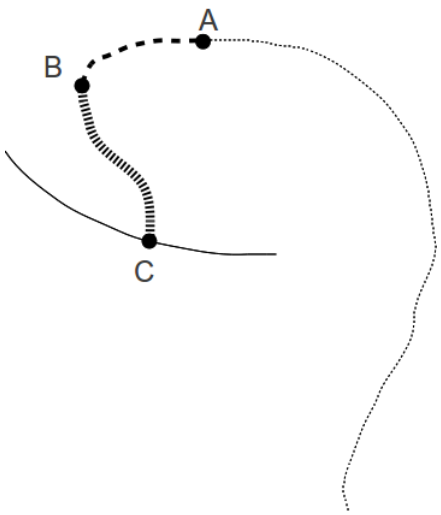
ahol x a kontúr mintavételezett pontjának indexe, $V(x)$ a kiszámolt nyelvalakzat, $N(x)$ egy semleges nyelvkontúr (a megfigyelt körvonalak átlaga) és $PC_i(x)$ az i. bázisfüggvény. Az egyes c_i együtthatók a bázisfüggvények súlyai. c_i egy kétdimenziós vektor, értéke a megszólaltatott magánhangzótól függ, amit bemeneti paraméterként használ a modell.

Pontosság: egyetlen PC bázisfüggvénnyel 85,7% pontosságot lehetett elérni, két bázisfüggvénnyel már 96,3%-ot [6].

2.4. Artikuláció

A modellben használt artikuláció egyszerűsített változata a tényleges artikulációnak. Csak a programban megvalósított részeket mutatjuk be. A hangképző szervek közül néhányat rögzített alakzatként kezeltünk, ilyen például az artikulációs csatorna hátulsó fala és a szájpadlás. A mozgatható alakzatok közé tartozik a gége a hangszalagokkal, a nyelv és az egész alsó állkapocs.

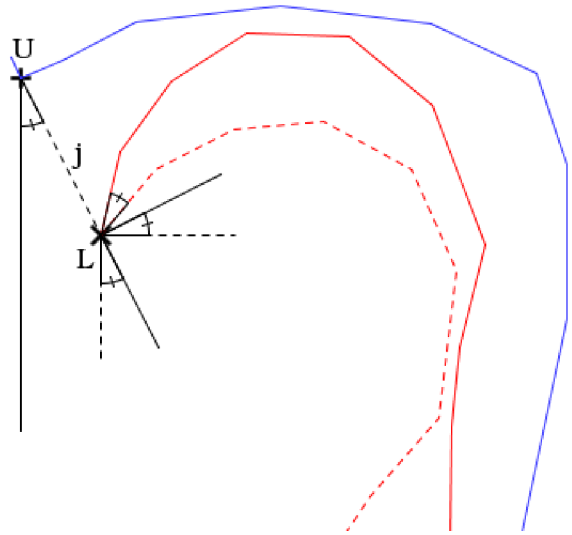
A gége fix kontúrral rendelkezik, azonban függőleges irányban mozgatható, ezzel lehet rövidíteni, illetve hosszabbítani az artikulációs csatornát.



1. ábra. A nyelv alakja három részből tevődik össze.

A nyelv alakja 3 részből áll (1. ábra). A hátulsó részének formáját a főkomponens-analízissel nyert egyenlettel számoljuk ki. A nyelv csúcsának helyzete (B pont) külön állítható, a csúcspontot Hermite interpolációval készített görbe

köti össze a hátsó nyelvformával. Ahhoz, hogy a kapcsolódás törésmentes legyen, az első derivált használatára is szükség volt a kapcsolódási pontban (A pont). A nyelv csúcspontja a szájüregben a száj alsó részén egy rögzített ponthoz (C pont) csatlakozik. Ennek a harmadik görbének az alakjához megfigyelt adatokat használtunk fel.



2. ábra. Az alsó állkapocs mozgatása.

Alsó állkapocs mozgása az alsó állkapocs koordináta rendszerének eltolását és forgatását foglalja magába. Ezzel együtt mozog az alsó fogsor, a szájüreg alsó fele és a nyelv. Az elforgatás szögét az alábbi egyenlettel számoljuk:

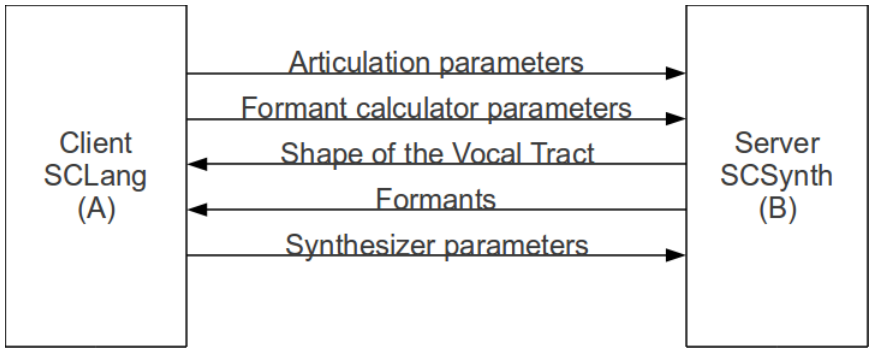
$$\alpha_{deg} = \frac{j}{2} + 7$$

ahol α_{deg} a szög fokban, j pedig az állkapocs nyitottsága (a távolság az alsó és felső metszőfogak között, mm-ben). A 2. ábrán a kék görbe az artikulációs csatorna hátulsó fele, az U pont a felső állkapocs koodináta rendszerének origója. Ha a nyitottság j-re van állítva, akkor U és L közötti távolság j. Az ábrán jelölt összes szög α. A belső szaggatott piros vonal a j-vel eltolt nyelv, a folytonos piros vonal az eltolt, majd elforgatott nyelv.

3. Megvalósítás

A modellt a SuperCollider környezetben implementáltuk. A SuperCollider egy programozási környezet algoritmikus zeneszerzésre és hangfeldolgozásra. Kliens-

szerver architektúrájú a felépítése, a kliensben található interpretált, objektum-orientált small-talk-szerű programozási nyelv felel a szerver vezérléséért. A szerver feladata a gyors jelfeldolgozás, valamint a hang be- és kimenet kezelése, natív bővítmények segítségével [7].



3. ábra. Kommunikáció a SuperCollider szerver és a kliensalkalmazás között.

A megvalósítandó program első verziója csak a kliens oldalon helyezkedett el, a szerver részt csak a hangszintetizáláshoz használta. A sok geometriai művelet sajnos nem volt elég hatékony az interpretált nyelvben, így később a számításigényes részek átkerültek a szerverre. A kliens-szerver közti aszimmetrikus kommunikáció szinkronizálása sok nehézséget okozott (3. ábra).

4. Eredmények

Az APEX modellnek létezik egy korábbi implementációja is, de annak fejlesztése félbemaradt, és a program elavult. Az új program még további fejlesztésre szorul, mivel hiányzik a szájüregi rész helyes kezelése (ajkak, fogak, nyelv alatti terület). Ezt leszámítva a modell megvalósítása sikeresnek mondható. Előrelépés a korábbi változathoz képest, hogy a használt környezetnek köszönhetően könnyebb a programot átírni más platformokra (Linux rendszeren készült, Macen is sikerült futtatni).

A hangszintézis az elkészült új verzióban interaktív, a bemenetet változtatva azonnal hallható a változás eredménye. A bemenő paraméterekből listát készítve lehet több hangot is összefűzni. A többi artikulációs modellhez hasonlóan az APEX-ben is megfigyelhetőek a hangok közötti átmenetek, a koartikuláció. Az artikulációs modell alkalmas a hangátmenetek beszédszervek tényleges fizikai jellemzőin alapuló interpolációjára.

5. Továbblépési lehetőségek

Több irányban is tovább lehet folytatni a fejlesztést. A hiányzó rész elkészítésével a teljes modell meg lenne valósítva. A teljes modell leprogramozása után a modell által kiszámolt formánsfrekvenciákat össze lehetne vetni valóságos mérésekkel.

A program jelenlegi felépítése a szerver-kliens közötti kommunikáció miatt nem ideális. Ennek egyik kiküszöbölési módja, hogyha a SuperCollider kliens helyett saját, natív klienst készítenénk. Ekkor nem lennénk korlátozva az interpretált nyelv sebességével, másrészt a SuperCollider szerver csak a hang kiadásáért lenne felelős, és csak a formánsadatokat kellene továbbítani.

A számítások sebességet tovább lehetne gyorsítani SIMD (Single Instruction Multiple Data) utasításkészlettel, mivel a keresztmetszetfüggvény kiszámításánál például minden keresztmetszeti szeleten ugyanazt az algoritmust kell végrehajtani.

A munka Erasmus ösztöndíj keretében, MSc diplomaterv formájában lett elfogadva a Kungliga Tekniska Högskolan Stockholm Speech, Music and Hearing tanszékén.

Hivatkozások

1. Shadle, C.H., Damper, R.I.: Prospects for articulatory synthesis: A position paper. In: 4th ISCA workshop, Pitlochry, Scotland. (2001)
2. Stark, J., Ericsdotter, C., Branderud, P., Sundberg, J., Lundberg, H.J., Lander, J.: The apex model as a tool in the specification of speaker-specific articulatory behavior. In: Proc XIVth Int'l Congr Phonetic Sci (ICPhS 99), San Francisco. (1999)
3. Branderud, P., Lundberg, H.J., Lander, J., Djamshidpey, H., Wäneland, I., Krull, D., Lindblom, B.: X-ray analyses of speech: Methodological aspects. In: FONETIK 98. (1998)
4. Ericsdotter, C.: Articulatory-Acoustic Relationships in Swedish Vowel Sounds. PhD thesis, Stockholm University (2005)
5. Soquet, A., Lecuit, V., Metens, T., Demolin, D.: Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with mri. Speech Communication **36**(3-4) (2002) 169–180
6. Lindblom, B.: A numerical model of coarticulation based on a principal components analysis of tongue shapes. In: 15th Int'l Congr Phonetic Sci, Barcelona. (2003)
7. Wilson, S., Cottle, D., Collins, N.: The SuperCollider Book. The MIT Press (2011)

Szabadalmak igénypontgráfjának automatikus előállítása és hibaelemzése

Kiss Márton[1], Vincze Veronika[1], Nagy Ágoston[1], Alexin Zoltán[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
{mkiss, vinczev, nagyagoston}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
alexin@inf.u-szeged.hu

Kivonat: Az alább ismertetett kutatásaink középpontjában az angol nyelvű szabadalmak igénypontjai állnak. A szabadalmak a részletes leíráson túl, az igénypontokban szabatosan foglalják össze a kért szabadalom lényegét, azt, hogy a védelem pontosan mire terjedjen ki. Egy szabadalom igénypontjai között vannak kitüntetett főigénypontok és aligénypontok, az aligénypontok főigénypontra és egymásra hivatkozhatnak. Ez az igénypontstruktúra minden esetben egy gráfot alkot. Nyelvtechnológiai eszközökkel előállítottuk az igénypontgráfot. Az előállított gráfot ábrázoltuk, hogy megkönnyítsük a hibadetektáláshoz szükséges szabályrendszer kialakítását, valamint a további kutatásokat. Mivel tanuló- vagy referenciakorpusz nem állt rendelkezésünkre így másik rendszerrel hasonlítottuk össze eredményeinket. A gráfok elemzése közben kialakítottunk egy szabályrendszert, amely megsértése legtöbbször rossz hivatkozásra, a főigénypont hiányára vagy más hibára utalt. A szabályrendszer segítségével a főigénypontok detektálására is lehetőség nyílik. A módszerrel az Amerikai Szabadalmi Hivatal által elfogadott és nyilvánosan elérhető szabadalmak között kerestünk és találtunk hibásakat.

1 A szabadalmak felépítése

A szabadalmak egységes szerkezettel bírnak [1]. A főigénypont mindig azzal kezdődik, hogy milyen kategóriába tartózik a levédetni kívánt szabadalom, például módszer, eljárás, eszköz, összetétel. Eztán következik ezek kifejtése: milyen lépésből/anyagokból áll a főigénypont elején említett dolog, és ezeket az alpontokat rekurzívan továbbfejtik az úgynevezett aligénypontokban. Fontos megjegyezni, hogy egy szabadalomnak speciális esetben több főigénypontja is lehet. A mi kutatásaink csak a főigénypont szerkezetére és az egymásra való hivatkozásaikra korlátozódtak.

2 Az igénypontgráf előállítása

Miért volt szükségünk az igénypontgráf előállítására, hiszen már van működő rendszer [2, 3], mely ezt a problémát megoldja? - tehetnénk fel joggal a kérdést. Sajnos az a rendszer, melyet mi találtunk (pattools.com/claim_tree.html) csak a gráfot állítja elő, a hivatkozások típusát viszont nem adja meg. Nekünk pedig szükségünk volt erre az információra is a további kutatáshoz.

Az igénypontok közötti kapcsolatot az igénypontokban lévő, reguláris kifejezésekkel felismerhető, hivatkozások/utalások segítségével határoztuk meg. Ezen hivatkozások felhasználásával építettük fel az igénypontgráfokat. A kutatáshoz írt programokat az UIMA keretrendszerben [4, 5] írtuk.

2.1 Az igénypontgráf előállításakor használt hivatkozástípusok

Kutatásunk során 997 db A24F alosztályba tartozó szabadalmat vizsgáltunk. A szabadalmak igénypontszekciói összesen 16812 darab igénypontot tartalmaztak. Az alábbi táblázat tartalmazza, hogy milyen hivatkozástípusokat különböztettünk meg és ezeknek milyen volt az eloszlásuk az általunk vizsgált 997 szabadalomban.

1. táblázat:A hivatkozástípusok megoszlása az általunk vizsgált 997 szabadalom esetében.

Hivatkozástípus	Előfordulás
root/nem hivatkozik	2 787
in claim #	3 277
of claim #	9 102
according to #	2 833
összes hivatkozás	17 999

2.2 A előállított igénypontgráfok ellenőrzése

Nem állt rendelkezésünkre referenciakorpusz, így egy meglévő rendszerrel hasonlítottuk össze eredményeinket. A pattools.com/claim_tree.html címen elérhető rendszer által generált gráfokkal vetettük össze a mi kimeneteinket. Így kézi ellenőrzésre csak akkor volt szükség, amikor különbséget fedeztünk föl a két kimenet között.

3 Főigénypontok meghatározása az igénypontgáf felhasználásával

Későbbi kutatási témát jelenthet, hogy a gráfokat felhasználva automatikusan detektálhatjuk a főigénypontokat. Erre nagy szükségünk lesz, mert a K+F projektünkben a későbbi szemantikus elemzés kiindulópontjai minden esetben a főigénypontok.

4 Hibaelemzéshez szükséges szabályrendszer kialakítása

Az igénypontgráf megalkotása után a kapott gráfokat elemezve 3 fő hibatípust tudtunk megkülönböztetni: 1) saját magára hivatkozik az igénypont, 2) a hivatkozott igénypont nem létezik, 3) ugyanaz két igénypont száma. Valamint felderítettünk lehetséges hibákat is, melyek nem minden esetben bizonyultak hibának, így ezek jelzése után kézi ellenőrzéssel kellett eldönteni, hogy valós volt-e a jelzés. Ilyen volt például, ha egy igénypont az utána következő igénypontra hivatkozik, vagy ha a hivatkozott főigénypont és a hivatkozó igénypont között van főigénypont.

A vizsgált 997 Amerikai Szabadalmi Hivatal által elfogadott szabadalomban az alábbi táblázatban felsorolt hibákat derítettük föl.

2. táblázat: A szabadalmakban felderített hibák.

Hibatípus	Előfordulás
Saját magára hivatkozik az igénypont	6
A hivatkozott igénypont nem létezik	2
Ugyanaz két igénypont száma	4
összes detektált hiba	12

5 Összefoglalás

Módszerünk más rendszerrel való összehasonlítása és a felderített hibák elemzése azt mutatja, hogy indirekt módon bizonyítható, hogy a rendszer kis hibával működik, ezért a későbbiekben jól használható szabadalmak igénypontgráfjainak előállítására. Az igénypontgráfot felhasználva több hibatípus kiszűrhető és megállapíthatóak a főigénypontok is.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Vincze V., Nagy Á., Klausz Á., Almási A., Kiss M.: Nyelvészeti problémák a szabadalmak feldolgozásában. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 168–179
2. Milton, H. W.: Method for preparing a claim tree in the preparation of a patent application. In: Patent. Bloomfield Hills, MI, US (2008)

3. Kahn, M. R.: Patent claim visualization system and method. In: Patent, Westampton, NJ, US (2009)
4. Osenga, K.: Linguistics and patent claim construction. Rutgers Law Journal Vol. 38, No. 61 (2006) 61–108
5. D. Ferrucci, A. Lally: UIMA by Example. IBM Systems Journal 43, No. 3 (2004) 455–475
6. D. Ferrucci, A. Lally: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. In: Journal of Natural Language Engineering. (2004) 327–348

Magyar NP-felismerők összehasonlítása

Miháltz Márton[1]

[1] MTA Nyelvtudományi Intézet, 1068 Budapest, Benczúr u. 33.
mmihaltz@gmail.com

Kivonat

Az előadásban szeretnénk bemutatni egy vizsgálat eredményét, melynek célja a cikk írásakor elérhető magyar nyelvű szintaktikai elemzőprogramok kiértékelése és összehasonlítása. Az elemzést a mondatokban található maximális főnévi csoportok határainak felismerésére korlátoztuk, összehasonlítási alapként a Szeged Treebank 2.0 [1] anyagát használtuk fel. A következő NP-felismerőket vetettük vizsgálat alá:

- MetaMorpho fordítóprogram szintaktikai elemzője [3]
- NooJ [5] magyar NP-nyelvtan
- Hunchunk gépi tanulásos NP-felismerő [4]

A MetaMorpho magyar-angol fordítóprogram forrásnyelvi szintaktikai elemző komponense kézzel írt szabályokkal működő jegystruktúrás környezetfüggetlen nyelvtant használ. A Nyelvtudományi Intézetben fejlesztett NP-nyelvtan a NooJ keretrendszerben készült véges állapotú automaták kaszkádja. A lexikai (morfológiai) elemzési szinthez több különböző megoldással is teszteltük. A Hunchunk rendszer a Szeged Treebanken tanított, maximum entrópiás Markov-modell NP-felismeréshez.

A Szeged Treebank 6 különböző témakörből (szépirodalom, iskolai fogalmazások, újságcikkek, számítástechnikai szövegek, jogi szövegek, gazdasági és pénzügyi rövidhírek) 1,2 millió szövegszót tartalmaz 82 ezer mondatban, részletes morfológiai és szintaktikai annotációval. A vizsgálathoz egyesítettük a mondatok halmazát, majd az ismétlődéseket kiszűrve 80,877 különböző mondathoz jutottunk. Minden mondatot külön, az eredeti szövegkörnyezete nélkül elemeztünk a vizsgált elemzőprogramokkal, a többször szereplő mondatokhoz az első előfordulásukhoz megadott annotációt használtuk fel (anélkül, hogy megvizsgáltuk volna, hogy a különböző előfordulások elemzései különböznek-e egymástól.)

A kiértékelés során minden mondatban megvizsgáltuk, hogy az egyes elemzők által megadott maximális NP-k közül hány szerepelt a treebankben (pontosság), illetve a treebank maximális NP-i közül hány található az elemző kimenetében (fedés), valamint megadtuk a két érték szokásos kombinációját is (F1-mérték). Egyezésnek csupán a teljesen megegyező kezdő- és záró terminálissal rendelkező NP-ket fogadtuk el, a részleges egyezéseket ebben a vizsgálatban ugyanúgy hibaként kezeltük, mint a teljesen rossz találatokat. A méréseket minden elemzővel elvégeztük külön-külön a 6 korpusz-témakör, illetve a 15 különböző forrás mindegyikére is.

Az 1. táblázatban közöljük a NooJ keretrendszerben írt szintaktikai elemző két különböző morfológiai elemzőt használó változatának összehasonlítását. Az 1. változat a Magyar Nemzeti Szövegtár [7] és a morphdb.hu [6] anyaga alapján készült morfo-

lógiai lexikont használja, míg a 2. változat egy, a NooJ rendszerben kézzel írt morfológiai elemző automatát. A 2. táblázatban a MetaMorpho és a NooJ elemző MNSZ-morphdb.hu-s változatának összehasonlítása látható.

1. táblázat: A NooJ elemző két változatának összehasonlítása a teljes treebank anyagán.

	NooJ 1.			NooJ 2.		
Témakör	P	R	F	P	R	F
Iskolai	43.61%	68.31%	53.23%	47.09%	67.52%	55.48%
Szám.tech.	34.19%	52.25%	41.34%	27.86%	43.18%	33.87%
Gazdasági	28.85%	48.80%	36.26%	23.92%	41.32%	30.30%
Szépirodalom	45.93%	68.19%	54.89%	43.87%	62.52%	51.56%
Hírek	35.16%	56.19%	43.25%	31.83%	50.43%	39.03%
Jogi	28.20%	51.34%	36.40%	22.58%	45.82%	30.25%
Teljes korpusz:	36.51%	58.72%	45.02%	33.34%	53.47%	41.07%

2. táblázat: A MetaMorpho és a NooJ elemzők összehasonlítása a teljes treebank anyagán.

	MetaMorpho			NooJ 1.		
Témakör	P	R	F	P	R	F
Iskolai	65.50%	71.92%	68.56%	43.61%	68.31%	53.23%
Szám.tech.	46.45%	56.72%	51.07%	34.19%	52.25%	41.34%
Gazdasági	43.78%	53.59%	48.19%	28.85%	48.80%	36.26%
Szépirodalom	63.91%	67.27%	65.55%	45.93%	68.19%	54.89%
Hírek	53.03%	58.43%	55.60%	35.16%	56.19%	43.25%
Jogi	35.21%	45.37%	39.65%	28.20%	51.34%	36.40%
Teljes korpusz:	52.14%	60.25%	55.90%	36.51%	58.72%	45.02%

A 3. táblázat a Hunchunk NP-felismerő és a másik két rendszer összehasonlítását foglalja össze. Mivel a Hunchunk rendszert a Szeged Treebank mondatainak egy részén tanították be, ehhez az összehasonlításhoz nem a teljes korpuszt, csak a tanításhoz fel nem használt, a szerzők által a kiértékelésre elkülönített 16.989 mondatot használtuk fel. Ezek közül kihagytunk 142 ismétlődő mondatot, illetve 494 mondatot a Hunchunk kimenetéből technikai okok miatt nem tudtunk az eredeti korpuszban azonosítani, így az összehasonlítás a maradék 16.353 mondat segítségével történt.

3. táblázat: A Hunchunk, a MetaMorpho és a NooJ elemzők összehasonlítása a treebank kiértékelésre elkülönített részén.

HunChunk			MetaMorpho			NooJ 1.		
P	R	F	P	R	F	P	R	F
78.67%	84.99%	81.71%	54.39%	61.52%	57.73%	37.57%	59.28%	45.99%

A NooJ elemző két változatának összehasonlításából egyértelműen kitűnik, hogy az MNSZ-morphdb.hu morfológiai anyagát használó változat teljesít jobban (1. táblá-

zat). A MetaMorpho elemző ennél a változatnál szignifikánsan jobban teljesít (2. táblázat). A Treebank szempontjából további érdekesség, hogy mindkét rendszer az iskolai fogalmazások és a szépirodalmi alkotások szövegein teljesít a legjobban és a jogi szövegeken a legrosszabbul.

A gépi tanulásos rendszer kiértékelő halmazán végzett mérések (3. táblázat) ugyanezt a sorrendet mutatják a két szabályalapú rendszer között, az élre viszont a Hunchunk rendszer kerül szignifikáns előnnyel. Mindenképpen szükséges azonban megemlíteni, hogy a gépi tanulásos rendszer teljesítménye szempontjából az alkalmazott technológián túl nem elhanyagolható szempont, hogy ez a rendszer a Szeged Treebank – a kiértékelő halmaz mondataihoz hasonló – mondatain tanulva a kiértékelő korpusz inherens sajátosságaira jobban rá volt hangolódva, mint a másik két, a korpusz anyagától függetlenül fejlesztett rendszer.

A bemutatott NP-felismerők kiértékelésében további lehetséges munka, ha a korrektebb összehasonlítás érdekében az elemzők és a Treebank különböző koncepciókkal készült nyelvtanai között megtaláljuk a legnagyobb közös részhalmazt, és az ezzel megadható elemzésekre redukálva ismételjük meg a kiértékelést. Néhány példa ilyen nyelvtani különbségekre: a MetaMorphoban a főnévi igeneves szerkezetek NP-knek számítanak, a Szeged Treebankben nem; a névutók a MetaMorphoban részei az NP-knek, a Treebankben nem; az olyan birtokos szerkezetek, ahol a birtok közvetlenül követi a birtokot, a Treebankben két NP-nek számítanak, a MetaMorpho és a NooJ nyelvtanában viszont van a kettőt egyesítő NP; a MetaMorphoban a főnévi fejhez kapcsolódó vonatkozó mellékmondat része a maximális NP-nek, a Treebankben nem stb. A részleges találatok súlyozott figyelembevétele és a hibatípusok vizsgálata szintén további lehetőségek.

Bibliográfia

1. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005) kiadványa, Szeged, december 8-9. (2005) 409–412
2. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas (2002) 710–717
3. 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, Foundation for International Studies. La Valletta, Malta (2004) 138–142
4. Recski G., Varga A., Zséder A., Kornai A.: Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2009)
5. Silberztein, M.: NooJ : an Object-Oriented Approach. In: Muller, C., Royauté, J., Silberztein M. (Eds): INTEX pour la Linguistique et le Traitement Automatique des Langues, Cahiers de la MSH. Presses Universitaires de Franche-Comté, Ledoux (2004) 359–369
6. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Simon, E., Vajda, P.: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In: III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2005)
7. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation. Las Palmas (2002) 385–389

Javában taggelünk

Novák Attila[1], Orosz György[2], Indig Balázs[2]

[1]MorphoLogic Kft., 1116 Budapest, Kardhegy utca 5.
novak@morphologic.hu
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai Kar,
oroszgy@itk.ppke.hu
dlazesz@gmail.com

Kivonat: A szófaji egyértelműsítés (POS tagging) a számítógépes nyelvfeldolgozás egyik alapfeladata. A feladat megoldására számtalan algoritmus sok különböző programozási nyelven megírt implementációja létezik. Az egyes szövegszavakhoz rendelendő morfológiai címkék megállapítása azonban csak az egyik részfeladat, amelyet a szöveg morfológiai annotációjakor el kell végezni: a címkén kívül a szótövet is azonosítani kell. A nem túl gazdag morfológiájú analitikus angol nyelv esetében egy szófaji egyértelműsítő és egy egyszerű tövesítő egymás után kapcsolása elfogadható eredményt ad. A magyarhoz hasonló ragozó nyelvek esetében azonban sokkal jobb eredményt kapunk, ha a szófaji egyértelműsítést és a szótő megállapítását egyaránt elvégző morfológiai elemzőt tartalmazó integrált eszközt használunk.

1 Bevezetés

Cikkünkben egy olyan új nyílt forráskódú eszközt mutatunk be, amely egyszerre végzi el a szófaji egyértelműsítést és a szótő megállapítását, tehát teljes egyértelműsített morfológiai annotációt ad. Az eszköz szófaji egyértelműsítő algoritmusa a TnT és HunPoS taggerekben implementált rejtett Markov-modell (HMM) algoritmuson alapul. Emellett tartalmaz egy olyan felületet, amelynek használatával morfológiai elemző illeszthető hozzá, amely nemcsak a tanítóanyagban nem látott szavak morfológiai címkéjének egyértelműsítését teszi sokkal pontosabbá, hanem a szavak szótövét is megadja. Az eszközt Java nyelven implementáltuk.

2 A korpusz reprezentativitása

Ha a magyarhoz hasonló agglutináló nyelveket az angollal hasonlítjuk össze abból a szempontból, hogy egy adott méretű korpusz milyen arányban tartalmazza az adott nyelv lehetséges szóalakjait, akkor azt tapasztaljuk, hogy míg egy azonos méretű korpuszban sokkal több különböző szóalak szerepel az agglutináló nyelvek esetében, mint az angolban, ezek ugyanakkor mégis sokkal kisebb részét fedik a korpuszban szereplő szótövek lehetséges alakjainak. A korpusz tehát sokkal kevésbé reprezenta-

tív a szókincs szempontjából, mint az angol esetében. 10 millió szavas korpuszméret esetében például az angolban általában 100 000-nél kevesebb különböző szóalakot találunk, ugyanakkor a magyarban jóval 800 000 feletti a különböző szóalakok száma. Ugyanakkor míg az angolban egy nyílt szóosztályba tartozó szónak legfeljebb 4–6 alakja van, a magyarban több száz vagy több ezer különböző alakot kapunk attól függően, hogy a produktív szóképzés eseteivel is számolunk-e. Természetesen a sokkal több lehetséges szóalak azt jelenti, hogy a lehetséges szófaji címkék száma is jóval magasabb a magyar esetében (több ezer szemben az angol néhány tucat címkéjével). Ezért egy magyar korpusz a szóalakok szintjén több szempontból is sokkal hiányosabban reprezentálja a nyelvet, mint az angol esetében: a szövegekben szereplő lemmák lehetséges ragozott alakjainak túlnyomó többsége teljesen hiányzik; az előforduló szóalakok is sokkal kevesebbszer szerepelnek; sokkal kevesebb példa van az adott konkrét morfológiaicímke-sorozatokra, sőt a lehetséges címkék nagy része egyáltalán nem szerepel a korpuszban.

A tanítóanyagban nem látott szavak kezelésére (illetve pl. a maximum entrópia modellt használó taggerek esetében a tanítóanyagban látott szavak esetében is) a szófaji egyértelműsítő eszközök általában tartalmaznak valamilyen mechanizmust, amely a szavak végződéseit vizsgálja a címke megjósolásához. A magyar esetében az előforduló hosszú toldaléksorozatok miatt jóval hosszabb szóvégek figyelembevételére van szükség, mint a nem agglutináló nyelvek esetében (ez különösen így van, ha a ragok mellett bizonyos produktív képzőket is azonosítani szeretnénk).

3 A morfológiai elemző hatása

A magyarhoz hasonló nyelvek esetében a rendszer tanítóanyagában nem szereplő szóalakok nagy része olyan szó, amelynek más ragozott alakjai előfordulnak a tanítóanyagban. Oravecz és Dienes [5], valamint Halácsy és mtsai. [4] bemutatták, hogy morfológiai elemző felhasználásával az általa ismert szóalakok esetében sokkal pontosabban meg lehet állapítani a tanítóanyagban nem szereplő szavak címkéjét, mint pusztán a tanítóanyagon betanított nyelvfüggetlen szóvégződés-felismerővel. Az utóbbi téves javaslatait a morfológiai elemző kimenetével megszűrve a tanítóanyagban nem látott szavakra a szófaji egyértelműsítés pontossága hatékonyan javítható. A morfológiai elemző pontosságot javító hatása annál jelentősebb, minél kisebb a rendelkezésre álló kézzel egyértelműsített tanítóanyag.

Az imént idézett eredmények nem olyan rendszerrel készültek, amely valóban integrált morfológiai elemzőt tartalmazott volna, hanem az annotálandó szövegen offline lefuttatott morfológiai elemző által visszaadott címkéket táblázat formájában betöltve szimulálták a morfológiai elemző hatását. Ez a fajta megoldás azonban nem használható bizonyos alkalmazásokban, például ha a taggert webszolgáltatásként szeretnénk üzemeltetni.

Többek között ezért döntöttünk úgy, hogy olyan eszközt implementálunk, amely integrált morfológiai elemzőt tartalmaz. A morfológiai elemzőt nemcsak arra használjuk, hogy a tanítóanyagban nem látott szavak címkézésének pontosságát javítsuk, hanem szükségünk van rá a szótövek megállapításához is. A morfológiai elemző számára sem ismert szavak kezelése (legfőképpen a szótövük megállapítása) morfo-

lógiai guesser (toldalékelemző) beépítésével oldható meg. Ezért az eszköz két csatolófelületet tartalmaz: egyet a morfológiai elemző, egyet pedig a guesser illesztésére.

4 Az optimális tő kiválasztása

A morfológia és főleg a sokkal lazább megszorításokkal dolgozó guesser gyakran több olyan lehetséges tőjelöltet is visszaad, amely a tagger által választott címkével kompatibilis. Sokszor tehát nem triviális a helyes szótő kiválasztása. A magyarban az egyik ilyen többértelműségi osztály az az azonos tövű ikes–iktelen igepároké. A lexikális *tör/törik, (fel)dolgoz/dolgozik* típusú párok mellett a produktív *-z/-zik* képzőpár szinte korlátlan mennyiségben hozza létre az ilyen típusú többértelműségeket. Emellett a két ragozási paradigma lényegében csak abban az egyetlen E/3 jelen idejű kijelentő módú alakban tér el, amely a lemmát adja, az összes többi igealak többértelmű a tő szempontjából, ezért egyben ez a leggyakoribb olyan tőtöbbértelműség-típus, amely a morfológiai elemző által felismert szóalakok körében fellép.

A tő egyértelműsítésére legegyszerűbb alapmodellként egy egyszerű unigram modellt használtunk. Ebben a modellben a szóalakként leggyakrabban előforduló alakot választjuk a lehetséges tövek közül. Ennek az egyszerű modellnek előnye, hogy nincs szükség a statisztika alapját képező korpusz semmiféle annotációjára. Ezért nem kell a rendelkezésünkre álló annotált korpuszra szorítkoznunk, hanem tetszőleges méretű anyagot használhatunk, még maga az annotálandó szöveg is hozzáadható a statisztika alapját képező anyaghoz. Ez a modell magyarra elég jó teljesítményt ad az ismeretlen szavak túlnyomó részét adó névszók esetében, mert ezeknek a leggyakoribb alakja a toldalékolatlan alanyeset.

Az egyik leggyakoribb többértelműségi osztály, ahol az egyszerű tőválasztási algoritmus hibázik, a magas hangrendű ikes–iktelen igepárok esete (ahol az *–ik* nélküli ige tárgyas). Ezeknek az *–ik* végű alakja is többértelmű: T/3 alanyú határozott tárgyas alak is lehet, és az ennél az igeosztálynál sokszor gyakoribb az *–ik* nélküli lemmánál (pl. a *nevezik* alak 4-szer olyan gyakori, mint a *nevez*). Ezt a problémát részben lehet kezelni egyrészt úgy, hogy a morfológiai elemzőben letiltjuk a *nevez*-hez hasonló gyakori igék produktív képzéssel előállított felbontását (ezzel a *név+ezik = nevezik* képzett alakot). Emellett az egyszerű unigram szóalak-gyakorisági modell annotált korpuszból vett adatokkal nyelvspecifikus módon kombinálva, illetve a tövek megelemzése után a tagger által választott elemzéssel inkompatibilis tövek kiszűrésével a tőmeghatározás pontossága növelhető.

5 Morfológiailag annotált korpusz építése nulláról

Azon nyelveknek jelentős része, amelyekre nem léteznek kézzel annotált tanítóanyagok, a magyarhoz hasonlóan bonyolult morfológiával rendelkezik. Ezen nyelvekre morfológiailag annotált egyértelműsített korpusz létrehozására egy olyan iteratív eljárás tűnik a leghatékonyabb módszernek, amelynek során morfológiai elemző létrehozását követően a rendelkezésre álló korpusz egy kis részhalmazát elemeztetjük,

és ezt kézzel egyértelműsítve a taggert betanítjuk. A korpusz következő részletét az így betanított taggerrel előegyértelműsítjük, majd az annotációt kézzel javítjuk, ezt a folyamatot addig ismételve, amíg elegendő annotált korpuszhoz nem jutunk. Nulláról épített annotált korpuszok esetében a minimális méretű tanítóanyag miatt a korábban vázolt adathiány-probléma még súlyosabb. Minél kevesebb tanítóanyag áll rendelkezésre, annál jelentősebb az integrált morfológiai elemző jótékony hatása az automatikus morfológiai címkézés pontosságára. Az annotáció kézi javítása is sokkal hatékonyabban végezhető, ha a morfológiai elemző egyéb elemzései is rendelkezésre állnak a tagger által választott elemzés mellett, és egyszerűen választani lehet az elemzések közül, mint ha ténylegesen mindig kézi javítgatásra van szükség.

Az iteratív korpuszannotációs eljárás használhatóságának fontos feltétele, hogy a tagger újratanítása ne vegyen igénybe túlzottan hosszú időt. A betanítás sebességének szempontjából a rejtett Markov-modell alapú szófaji címkéző eszközök nagyságrendekkel felülmúlják a bonyolultabb maximum entrópia vagy CRF-alapú algoritmusokat, amelyeknek betanítási ideje jóval hosszabb. (Konkrétan a HMM-alapú HunPoS [4] betanítása a Szeged korpuszon [6] kevesebb, mint egy percet vesz igénybe, szemben a maximum entrópia alapú OpenNLP hat órás betanítási idejével ugyanazon a gépen.) Mindemellett a HMM-alapú eszközök számos nyelvre – többek között magyarra is – az egyértelműsítés pontosságában is élen járnak.

Bár a magyar nyelvre már létezik egy olyan nyelvspecifikus eszköz, amely tartalmaz morfológiai elemzőt, és platformfüggetlen implementációval rendelkezik: a magyarlanc [7], ennek azonban nyelvspecifikus mivolta mellett komoly hátránya az alapjául szolgáló Stanford POS tagger nagy erőforrásigénye és a betanítás lassúsága.

6 Az új eszköz

Az elérhető HMM-alapú megoldások nem tartalmaznak beépített morfológiai elemzőt. A népszerű és megengedő licenszű HunPos tagger kiegészíthető lenne a kívánt funkcionalitással, de az implementációjához használt programozási nyelv csekély ismertsége ennek (és a tagger integrálásának) korlátját jelenti. Egy, az iparban elterjedtebb nyelv használata könnyebb szerves integrációt tesz lehetővé olyan nyelvfüggetlen keretrendszerekhez, mint az UIMA vagy a GATE. Ezért döntöttünk egy új, a tanítási sebességét tekintve jól használható, nyelvfüggetlen, morfológiai elemzővel könnyen integrálható szófaji egyértelműsítő implementációja mellett. Az új, nyílt forráskódú, Java nyelven implementált, rejtett Markov modellen alapuló POS-tagger, melynek alapjául a TnT [1] és a HunPos rendszerek szolgálnak, a korábban említett problémák kiküszöbölése érdekében a szófaji egyértelműsítés és a szótövezés problémáját egy feladatként kezeli. A rendszer képes morfológiai elemző és guesser aktív használatára a szófaji egyértelműsítés közben, továbbá az elemzés kimenetét a szótő meghatározására is felhasználja. Az eszközt olyan alkalmazásprogramozási felülettel láttuk el, amelyen keresztül egyszerűen illeszthető hozzá tetszőleges morfológiai elemző. Mivel gyakran az egyértelműsített taghez tartozó tő sem egyértelmű (különösen nem az azoknak a szóalakoknak az esetében, amiket a morfológiai elemző nem ismer, hanem a lehetséges töveiket a guesser állítja elő), olyan

mechanizmussal is kiegészítettük a rendszert, amely a lehetséges többértelmű tövek közül is hatékonyan választ.

Bibliográfia

1. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the sixth conference on Applied natural language processing (2000)
2. 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
3. Halácsy, P., Kornai, A., Oravecz, Cs., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC (2006) 2245–2248
4. Halácsy, P., Kornai, A., Oravecz, Cs.: HunPos: an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (2007) 209–212
5. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: Third International Conference on Language Resources and Evaluation (2002) 710–717
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 (2010)
7. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 394–395

A HunOr magyar-orosz párhuzamos korpusz

Szabó Martina Katalin[1], Schmalcz András[2], Nagy T. István[2], Vincze Veronika[3]

[1]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
szabomartinakatalin@gmail.com
[2]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
schmalcz.andras@stud.u-szeged.hu, nistvan@inf.u-szeged.hu
[3] SZTE-MTA Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: A jelen dolgozatban a HunOr, egy eleddig hiányzó digitalizált magyar–orosz párhuzamos korpusz létrehozásáról számolunk be. A dolgozat a korpuszépítési munka céljáról, jelenlegi állásáról, az eddigi munka során szerzett tapasztalatokról, a munka folyamatáról és eszközeiről, valamint a HunOr korpusz adatairól igyekszik átfogó képet adni. Az ismertetés során részletesen szólunk azokról az elméleti és gyakorlati jellegű problémákról, amelyek az eddig elvégzett és a jelenleg folyó feldolgozási munkák (mondatra bontás, mondatszintű párhuzamosítás, NE-annotálás) során elméleti vagy gyakorlati szempontból megoldásra váró feladatként léptek fel.

1 Bevezetés

A HunOr korpusz autentikus magyar nyelvű szövegeket, valamint azok orosz fordításait, illetve autentikus orosz nyelvű szövegeket, valamint azok magyar fordításait tartalmazza. A korpusz létrehozásának elsődleges célja, hogy vizsgálati anyagot teremtsünk a magyar–orosz, illetve az orosz–magyar fordításkutatás számára. Ugyanakkor, mivel a korpusz nem csupán fordított, hanem autentikus szövegeket is tartalmaz mindkét nyelven, számos, egyéb tudományterület kérdéskörébe tartozó nyelvészeti probléma számítógéppel támogatott vizsgálatát is lehetővé fogja tenni. A korpusz mindemellett különféle számítógépes nyelvészeti alkalmazásokhoz, például a gépi fordításhoz is kitűnő segédletet biztosíthat.

2 A HunOr korpusz szöveganyaga

A korpusz feldolgozott szövegállománya jelenleg valamivel több mint 75 000 szövegszót tartalmaz, azonban folyamatos bővítés alatt áll. A korpusz szövegei különböző típusú forrásból (internetes kiadvány, könyvformátum stb.) származnak.

A HunOr a szövegműfajokat illetően három kisebb egységre bontható: szépirodalmi, tudományos, valamint hivatalos alkorpuszra. Hamarosan azonban reményeink

szerint sajtónyelvi, a Russzisztika Központ *Orosz Negyed* című kiadványainak szövegeivel is bővül a korpusz.

A szépirodalmi alkotások közül a korpusz jelenleg a *Kladbiščenskie istorii* című művet tartalmazza, amelynek szerzője a Magyarországon egyelőre csak álnéven, Borisz Akunyinként ismert Grigorij Cshartisvili. A novellákat és esszéket tartalmazó könyv 2005-ben jelent meg. A művet 2008-ban *Temetői történetek* címmel Bagi Ibolya és Sarnyai Csaba ültették magyar nyelvre. A korpuszban található tudományos szövegek a szépirodalomhoz kapcsolódó, orosz forrásnyelvű elemző tanulmányok: Nyikolaj Bergyaev egy hosszabb lélegzetű, 1990-ben, *O „večno-babjom" v russkoj duse* címen publikált művének egy részlete, valamint Vitalij Orlov *Hranitel „nenužnih veščej"* című, 1999-es tanulmánya. A fordításokat 2007-ben Régéczi Ildikó, valamint 2009-ben Józsa György Zoltán készítették. A hivatalos alkorpusz a Magyar Külügyminisztérium honlapján közzétett, *Tények Magyarországról* című kiadvány következő szövegeiből áll: *A magyar kultúra ezer esztendeje; Nemzeti jelképek, nemzeti ünnepek; Magyar Nobel-díjasok egy jobb világért.*

Az alábbi táblázat bemutatja a HunOr jelenlegi feldolgozott állományának összefoglaló adatait:

1. táblázat: A HunOr korpusz adatai.

Szövegtípus	Szövegszavak		Mondatok		Fordítási irány
	orosz	magyar	orosz	magyar	
Szépirodalom	52 798	57 980	3 255	3 313	orosz → magyar
Tudományos	7 014	7 483	360	348	orosz → magyar
Hivatalos	15 924	14 412	710	561	magyar → orosz
Összesen	75 736	79 875	4 325	4 222	

3 A korpusz feldolgozása

A korpusz későbbi hasznosíthatósága érdekében szükségesnek bizonyult a szövegek mondatokra bontása, mondatszintű párhuzamosítása, illetve – ez utóbbival összefüggésben – a szövegek tulajdonnévi annotálása.

3.1 A szövegek mondatokra bontása és mondatszintű párhuzamosítása

A korpusz mondatokra bontása, valamint mondatszintű párhuzamosítása szükségessé tette a mondatnak mint a két művelet alapegységének a pontos meghatározását.

A mondat meghatározásának a feladata korántsem triviális; problematikusak ugyanis az olyan kifejezések, amelyekben a kettősponttal záródó szerzői szavakat egy nagy kezdőbetűvel kezdődő idézet (egyenes beszéd), egy dialógus, egy önálló mondatokból álló felsorolás vagy egy kifejtő magyarázat követi. E szövegtípusok közül az idézés és a dialógus a szépirodalmi, a felsorolás és a kifejtő magyarázat pedig a tudományos és a hivatalos stílusú szövegek gyakori szerkesztésbeli sajátsága. A HunOr korpusz műfaji összetétele okán fontos feladat volt tehát, hogy egységes rendszert

dolgozzunk ki a kettősponttal szerkesztett kifejezések annotálásához. A probléma megoldásának céljából elvégeztük az említett szövegtípusok magyar és orosz helyesírási gyakorlatának összevető vizsgálatát, valamint áttekintettük a vonatkozó orosz és magyar irodalom megjegyzéseit [3, 11, 13, 14]. A tapasztaltak részletes bemutatásától a dolgozat keretei miatt most eltekintünk.

A kettőspont után kis kezdőbetűvel kezdődő kifejezések annotálása nem volt problematikus számunkra, azokat egységesen egy mondatba tartozónak jelöltük az előtte álló, kettősponttal végződő szerzői bevezetővel. A nagy kezdőbetűvel kezdődő, kettőspont után álló idézetek, dialógusok, felsorolások és leírások annotálása azonban már kérdéses volt. A kínálkozó lehetőségek a következők voltak:

a) a kettősponttal záródó kifejezést egy mondatként kezeljük az általa bevezetett mondattal; amennyiben a kettősponttal záródó kifejezést több mondatból álló szövegrész követi, úgy a szerző szavait egy mondatként kezeljük annak első mondatával, majd a többi mondatot önálló mondatokként annotáljuk;

b) a kettősponttal záródó kifejezést, valamint az általa bevezetett, egy vagy több mondatból álló szövegrészt együtt egyetlen mondatként kezeljük;

c) a kettősponttal záródó kifejezést önálló mondatként annotáljuk csakúgy, mint az általa bevezetett mondatot, vagy a több mondatból álló szövegrész minden egyes mondatát.

Vizsgáljuk meg a fenti szegmentálási lehetőségeket az alábbi példán [3] keresztül!

E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. A másik vizsgálati forma a fizikális terheléses teszt. Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására.

A lehetséges mondatra bontási megoldások tehát a következők:

a) **<S>** E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. **</S> <S>** A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. **</S> <S>** A másik vizsgálati forma a fizikális terheléses teszt. **</S> <S>** Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

b) **<S>** E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. A másik vizsgálati forma a fizikális terheléses teszt. Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

c) **<S>** E vizsgálatoknak két formája terjedt el: **</S> <S>** Az egyik vizsgálati forma az oxitocinterheléses teszt. **</S> <S>** A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. **</S> <S>** A másik vizsgálati forma a fizikális terheléses teszt. **</S> <S>** Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

Az (a) és a (b) megoldást támogatja a magyar és az orosz korpuszannotálási gyakorlat [4, 7, 12, 15], amely szerint minden kettőspontot tagmondatok közötti írásjelként annotálnak a készítők. A módszer azonban ellentmondásosnak tűnik, amennyiben szem előtt tartjuk Rozental [13] megjegyzését, miszerint az egyenes beszéd megfelel az önálló mondat szintaktikai kritériumainak, illetve azt, hogy mind a magyar, mind az orosz szerzők [3, 11, 14] különbséget tesznek az önálló mondatokból, valamint a nem önálló mondatokból álló felsorolások között. Amennyiben a korpuszannotálási gyakorlatot követnénk tehát, úgy kettő vagy több, szintaktikai szempontból önálló mondatot egyetlen mondatként jelölnénk be a korpuszban.

Az (a) megoldást támogatja továbbá az orosz helyesírási gyakorlat; az orosz szerzők ugyanis – a magyar gyakorlattal ellentétben [3] – nem ismerik el a kettőspontot mondatvégi írásjelként: a mondatzárók között rendre a pontot, a felkiáltójelet, a kérdőjelet, valamint a három pontot sorolják fel [11, 13, 14]. Amennyiben tehát az orosz helyesírási gyakorlathoz ragaszkodnánk, úgy a pontokat mondatvégi, a kettőspontokat pedig tagmondatok közötti írásjelként kezelnénk, azaz az (a) megoldást alkalmaznánk a korpuszban. Az eljárásmód vitatható volta azonban kiütközni látszik azokban az esetekben, ahol a szerző szavai több mondat vezetnek be. Véleményünk szerint ugyanis semmiféle különbség nem mutatkozik a szerző szavai és az azokat közvetlenül követő mondat, valamint a szerző szavai és az azokat nem közvetlenül követő mondat (vagy mondatok) között, ami alapul szolgálhatna ehhez a sajátos annotálási módhoz.

A (c) megoldást támogatják az (a) és a (b) megoldással szemben tett kritikai észrevételek, ugyanakkor a (c) annotálási mód ellen szól az említetteknek megfelelően a korpuszannotálási gyakorlat, valamint az, hogy az orosz nyelvben nem ismerik el a kettőspont esetleges mondatvégi státusát. Ugyanakkor grammatikáinkban nem találni olyan kritériumot, amely lehetetlenné tenné a kettősponttal végződő mondat feltevését, pl: „[A mondatot] a szerkesztés különféle nyelvtani eszközeinek viszonylagos lezártsága jellemez” [8]; „formai szempontból elsősorban az intonáció egysége, lezártsága jellemzi a magyar mondatot” [6]; „A mondat egy vagy több szóból áll, zárt intonációs szerkezet jellemzi” [2].

Az ismertetett érveket és ellenérveket megfontolva a HunOr korpuszban végül a (c) megoldás alkalmazása mellett döntöttünk. Az általunk választott eljárásmód tehát a következő: azokat a kettőspontokat, amelyek nagy kezdőbetűvel kezdődő, egy vagy több mondatból álló szövegrészt vezetnek be, mondatvégi írásjelekként kezeljük a korpuszban, s a kettősponttal végződő szerzői bevezető utáni mondatot vagy mondatokat önálló egységekként annotáljuk.

Az annotáció az elmondottak alapján tehát szakít a hazai és az orosz korpuszannotálási gyakorlattal. Ugyanakkor, mivel elméleti megfontolásokon alapszik, teoretikus szempontból a többi lehetséges megoldásnál helytállóbbnak tekinthető. Mindemellett érdemes kiemelni azt is, hogy a módszer az egységessége folytán nem teremt kérdéses eseteket, amelynek köszönhetően annak korpuszbeli alkalmazása mind az annotátori döntéshozatal, mind az automatikus munka szempontjából problémamentesen megoldható.

A mondatok párhuzamosításában a fordítási egység hatféle megfeleléstípusát szokás megkülönböztetni [1, 5, 10], a HunOr korpusz építése során azonban egy hetedik típust is detektáltunk ((g)-vel jelölve). A hét megfeleléstípus tehát a következő:

a) 1-1 megfelelés: egy forrásnyelvi mondat egy célnyelvi mondatnak felel meg;
b) 0-1 megfelelés, azaz a beszúrás;
c) 1-0 megfelelés, azaz a kihagyás;
d) 1-N megfelelés, azaz a részekre bontás;
e) N-1 megfelelés, azaz az összevonás;
f) N-M megfelelés, amely a mondathatár eltolódásából fakad;
g) N=M megfelelés, amely a mondatok sorrendjének a cseréjéből fakad: a forrásnyelvi szöveg két, (a) (b) sorrendű mondatának megfelelője a célnyelvű szövegben (b) (a) sorrendben található meg.

A hetedik megfeleléstípust az alábbi, a HunOr korpuszból származó példa szemlélteti:

Dombrovszkij ezt a verset igen szerette.

Kit vulkán edzett jó előre
S a Nemezis kezébe tett:
A bosszú kése vagy szabadság titkos őre,
Bírák bírája bűn és jogtiprás felett!

Лемносский бог тебя сковал
Для рук бессмертной Немезиды,
Свободы тайный страж, карающий кинжал,
Последний судия Позора и Обиды.

Это стихотворение Домбровский очень любил.

3.2 A tulajdonnévi annotálás

Az automatikus párhuzamosítást segítik a szövegben található horgonyelemek, például a számok és tulajdonnevek [9], így a szövegekben két független annotátor bejelölte a tulajdonneveket. Az annotáció során a négy klasszikus tulajdonnévosztályt alkalmaztuk: személy, szervezet, hely és egyéb. Az annotációk közti egyetértési ráta a magyar anyagon 0,8695 és 0,9609, az oroszon pedig 0,7995 és 0,9318 volt (κ-mértékben és mikro F-mértékben megadva). A tulajdonnevek kézi annotálása lehetővé teszi továbbá különféle magyar és orosz tulajdonnév-felismerő rendszerek teljesítményének mérését.

A 2. táblázatból kiderül, hogy a két nyelvben eltérő gyakorisággal fordulnak elő a tulajdonnevek, ami valószínűleg egyrészt nyelvek közti különbségeknek köszönhető: léteznek sajátos, csak az adott nyelvben tulajdonnévnek számító elemek, mint például az orosz *человечество*, melynek magyar megfelelője (*emberiség*) nem számít tulajdonnévnek. Másrészt a fordításnak köszönhetően stilisztikai különbségek is lehetnek a szövegek között: például az egyik nyelvben szereplő tulajdonnév helyett állhat névmás a másik nyelvű szövegben.

2. táblázat: A HunOr korpuszban található tulajdonnevek.

	orosz	**magyar**
Személy	1535	1487
Hely	608	479
Szervezet	137	105
Egyéb	291	224
Összesen	2571	2295

A HunOr korpusz esetében a horgonykeresést illetően több jelentős nyelvi tényezőt kell szem előtt tartanunk: Először is, az általunk feldolgozni kívánt szövegek nem azonos karakterkészletű nyelvekből származnak, hiszen a magyar nyelv a latin, az orosz nyelv a cirill ábécét használja. A tulajdonnevek tehát nem azonos írásmódban fordulnak elő, ami jelentős nehezítő körülmény például egy magyar–angol párhuzamos korpusz létrehozásához képest. További jelentős nehezítő körülmény, hogy az orosz nyelvben az idegen tulajdonneveket nem azok forrásnyelvi betűzése, hanem részben azok kiejtése alapján írják át cirill betűkre, pl. *New York Times* (angol) → *Нью-Йорк Таймс* [Nju Jork Tajms]; *François de la Chaise* (francia) → *Франсуа де ла Шез* [Fransua de la Šez]. E problémákra tehát fokozott figyelmet kell fordítanunk az automatikus párhuzamosítás során.

Ugyanakkor jelentős könnyebbség, hogy a köz- és a tulajdonnevekben a kezdőbetűk nagyságát illetően a két nyelvben nincs alapvető eltérés, illetve, hogy a két nyelv központozási készlete és annak használati sajátságai alapvetően azonosak.

4 A HunOr korpusz hasznosíthatósága

Az elkészült korpuszt a jövőben szeretnénk morfológiai és szintaktikai elemzésnek is alávetni. A morfológiailag és szintaktikailag elemzett párhuzamos korpusz minden bizonnyal kiemelkedő szerepet tölthet majd be a transzferalapú gépi fordítórendszerek fejlesztésében, de többnyelvű információkinyerésben is hasznosítható lesz, ugyanakkor a többszintű annotációnak köszönhetően (morfológia, szintaxis, névelemek) a két részkorpusz a magyar, illetve orosz nyelvű számítógépes nyelvészeti kutatásokat egyaránt ösztönözheti.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg. Szabó Martina Katalin konferencián való részvétele a Szegedi Tudományegyetem Hallgatói Önkormányzata segítségével vált lehetségessé.

Bibliográfia

1. Klaudy K.: A fordítás elmélete és gyakorlata. Angol / francia / német / orosz fordítástechnikai példatárral. Scholastica Kiadó, Budapest (1997)
2. Kugler N.: A mondattan általános kérdései. In: Keszler B. (szerk.): Magyar Grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 369–393
3. Laczkó K., Mártonfi A.: Helyesírás. Osiris Kiadó, Budapest (2006)
4. Magyar Nemzeti Szövegtár [http://corpus.nytud.hu/mnsz/]
5. Pohl G.: Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás. In: Alexin Z., Csendes D. (szerk.): MSzNy 2003 – I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 254−259
6. Rácz E.: Mondattan. In: Rácz E. (szerk.): A mai magyar nyelv. Nemzeti Tankönyvkiadó, Budapest (1968) 205–458
7. Szeged Korpusz [http://www.inf.u-szeged.hu/projectdirs/hlt/]
8. Tompa J.: A mondat és a mondattan általános kérdései. In: Tompa J. (szerk.): A mai magyar nyelv rendszere. Leíró nyelvtan II. Akadémiai Kiadó, Budapest (1962) 7–22
9. Tóth, K., Farkas, R., Kocsor, A.: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. Acta Cybernetica Vol. 18, No. 3 (2008) 463–478
10. Vincze V., Felvégi Zs., R. Tóth K.: Félig kompozicionális szerkezetek a SzegedParalell angol−magyar párhuzamos korpuszban. In: Tanács A., Vincze V. (szerk.): MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 91−101
11. Лопатин, В.В. ред.: Правила русской орфографии и пунктуации. Полный академический справочник. Издательство «Эксмо», Москва (2007)
12. Национальный корпус русского языка [http://www.ruscorpora.ru/]
13. Розенталь, Д.Э.: Русский язык. Пособие для поступающих в вузы. Издание второе, дополненное и переработанное. Московский университет, Москва (1988)
14. Соловьев, Н.В.: Орфографический словарь. Комментарий. Правила. 3-е издание. Издательство «Норинт», Сацкт-Петербург (2000)
15. ХАНКО [http://www.ling.helsinki.fi/projects/hanco/]

Magyar szóalak- és morfológiaielemzés-adatbázis

Szidarovszky Ferenc P.[1], Tóth Gábor[1], Tikk Domonkos[2,3]

[1] F12 Kft., 1025 Budapest, Szépvölgyi út 191.
{ferenc.szidarovszky, gabor.toth}@f12.com
[2] Gravity Research&Development Kft., 1101 Budapest, Expo tér 5–7.
domi@gravityrd.com
[3] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tsz, 1117 Budapest, Magyar Tudósok krt. 2.
tikk@tmit.bme.hu

Kivonat: Célunk egy olyan morfológiai elemző megoldás létrehozása, mely átlagos felhasználás mellett a szavak nagy arányát tudja elemezni, megengedve a helytelen szavak „közeli" értelmezését is. Ennek a megoldásnak műszakilag platformfüggetlennek és kevés szó elemzése esetén is hatékonynak kell lennie. Ennek érdekében egy olyan statikus MySQL adatbázist építünk, mely tartalmazza a szóalakokat és azok elemzését, így a szavak elemzése adatbázis-lekérdezéssel történhet. Kellő feltöltöttséggel ez az adatbázis megvalósíthatja célunkat.

1 Bevezetés

Az elmúlt években sikerrel és nagy megelégedésünkre használtuk az OcaMorph morfológiai elemzőprogramot [1]. Funkcionalitási szempontból magyar szavak morfológiai elemzésére a legjobb megoldások egyike. Technikai szempontból azonban vannak hátrányai:

- Csak külön folyamatként lehet elindítani, nehezen és/vagy nem hatékonyan integrálható más rendszerekbe.
- Magas a kezdeti inicializálás időigénye, gyakori, de kevés szót tartalmazó elemzési feladatokra nem hatékony. (Ilyen használat merül fel pl. ajánlórendszerek esetében.)

Célunk egy olyan morfológiai elemző megoldás létrehozása, mely a fenti technikai problémákat kiküszöböli. Ezt egy olyan statikus adatbázis létrehozásával igyekszünk elérni, mely tárolja a szóalakokat és azok morfológiai elemzéseit.

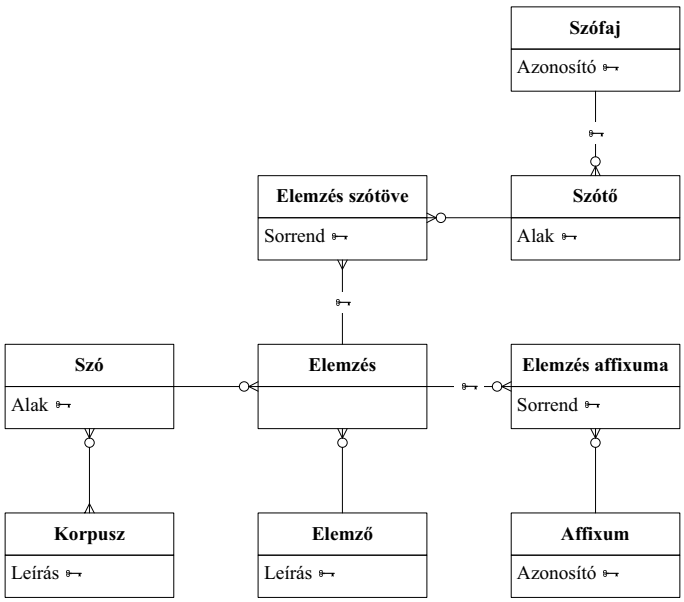
A megoldással kapcsolatos elvárásainkról fontos megjegyezni:

- A megoldástól nem várjuk, hogy teljes legyen, de törekvésünk, hogy átlagos felhasználás esetén a szóalakok minél nagyobb arányát tartalmazza.
- A megoldástól elvárjuk, hogy egy helyes szóalakra jó elemzéseket adjon, de helytelen szóalakok esetén csak annyit várunk el, hogy ha ad elemzést, akkor az alakhoz „közeli" elemzéseket adjon.
- A megoldástól nem várjuk, hogy tartalmazza az összetett szavakat. (Ezek elemzése jól visszavezethető több nem összetett szó elemzésére.)

2 Az adatbázis létrehozása

2.1 Adatstruktúra

Az adatbázis adatmodelljét az 1. ábra szemlélteti:



1. ábra. Az adatbázis adatstruktúrája

A **Szófaj** tábla tartalmazza a szófajok listáját (jelenleg 18 sor), kulcsa a szófaj azonosítója. Az **Affixum** tábla tartalmazza az affixum fajták listáját (jelenleg 137 sor), kulcsa az affixum azonosítója.

A **Korpusz** tábla tartalmazza a korpuszok listáját (jelenleg 3 sor), kulcsa a korpusz leírása. A **Szó** tábla tartalmazza az eddig talált elemezhető szóalakokat (jelenleg 2 300 717 sor), kulcsa az alak. A korpuszokat és a bennük megtalálható szavakat összekapcsoljuk.

A **Szótő** tábla tartalmazza az eddig talált szótövek listáját (jelenleg 199 822 sor), kulcsa a kapcsolódó szófaj és az alak párosa.

Az **Elemző** tábla tartalmazza a morfológiai elemzők listáját (jelenleg 1 sor), kulcsa az elemző leírása. Az **Elemzés** tábla tartalmazza a tárolt elemzések listáját (jelenleg 3 881 689 sor), kapcsolódik hozzá az elemző, és az elemzett szó.

Az **Elemzés szótöve** tábla (jelenleg 4 671 757 sor) tartalmazza a kapcsolódó elemzés által megadott szótöveket sorrendben. Az **Elemzés affixuma** tábla (jelenleg 9 543 740 sor) tartalmazza a kapcsolódó elemzés által megadott affixumokat sorrendben.

Mint látható, az adatmodellt felkészítettük a korpuszok szétválasztására és a jövőbeli esetlegesen előforduló többféle morfológiai elemző együttes kezelésére.

2.2 Feltöltés

Az adatbázis feltöltése az OcaMorph [1] felhasználásával történt úgy, hogy különböző korpuszok szavait leelemeztettük az OcaMorph-fal, és a kapott elemzéseket betöltöttük az adatbázisba.

Az alábbi korpuszok kerültek feldolgozásra:

- Web korpusz 2.0 [2, 3]
- Magyar wiki korpusz [4]
- Saját, 368 könyvből/regényből álló, az internetről letöltött korpuszunk.

3 Eredmények

3.1 Az adatbázis

Létrejött egy statikus (MySQL) adatbázis, mely:

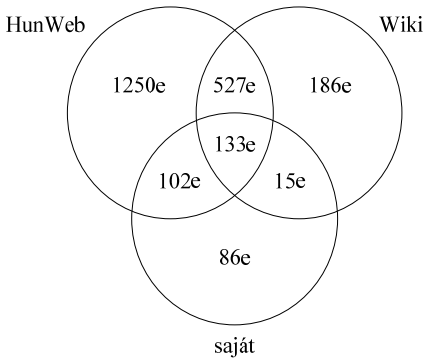
- platformfüggetlen;
- jól integrálható meglévő rendszerekbe;
- gyakran végrehajtásra kerülő, de kevés szó elemzését igénylő feladatokra is hatékony.

További előnye, hogy az elemzések egyszerre, hatékonyan állnak rendelkezésre, így alkalmassá váltak statisztikai elemzések elvégzésére, pl. szociolingvisztikai elemzésekhez.

3.2 Statisztikák

A fenti három korpusz feldolgozásával kb. 2,3 millió szóalak összesen kb. 3,8 millió elemzését tároltuk le. Ezek az elemzések közel 260 ezer szótőre hivatkoznak.

Az alábbi ábra szemlélteti a szóalakok korpuszokon belüli előfordulását:



2. ábra. Szóalakok korpuszokon belüli előfordulása.

Az egy szó alternatív elemzéseinek számának eloszlását az alábbi táblázat tartalmazza:

1. táblázat: Egy szóra eső alternatív elemzések számának eloszlása.

A szó alternatív elemzéseinek száma	Ilyen szavak száma
1	1 353 265
2	578 828
3	211 574
4	105 065
5	17 166
6	25 463
7	2 627
8	4 198
9	1 164
≥10	1 365

Az elemzésekben szereplő affixumok számának eloszlását az alábbi táblázat tartalmazza:

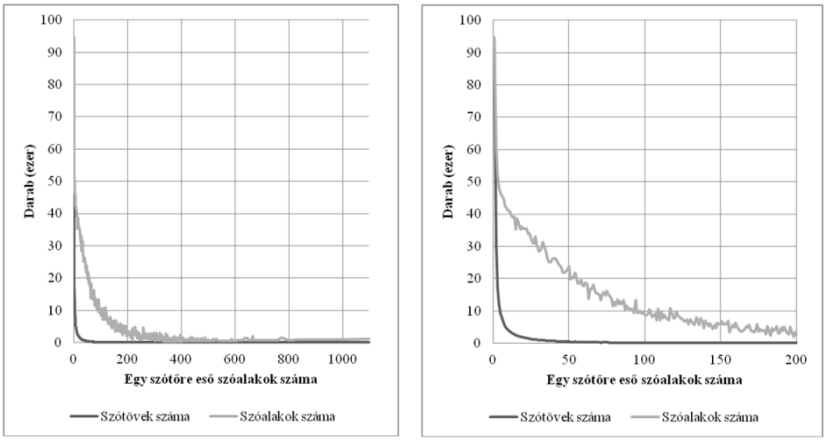
2. táblázat: Az elemzésekben szereplő affixumok számának eloszlása.

Elemzésben szereplő affixumok száma	Ilyen elemzések száma
1	1 106 984
2	798 212
3	896 217
4	468 085
5	238 277
6	119 013
7	30 890
8	15 925
9	2 183
10–12	1 034

Az elemzésekben közel 20 ezer különböző affixumsorozat szerepel.

A legtöbb különböző szóalak az *út* szótőhöz tartozott, összesen 1098. Az öt legtöbb különböző szóalakkal rendelkező szótő az *ad*, *gond*, *név*, *szó* és *út* voltak.

A 3. ábra mutatja, hogy hogyan alakul a szótövek, illetve szóalakok száma az egy szótőhöz talált különböző szóalakok számának függvényében:



3. ábra. Szótövek, illetve szóalakok száma az egy szótőhöz talált különböző szóalakok számának függvényében.

4 Jövőbeli tervek

4.1 További korpuszok bedolgozása

Tervezzük az adatbázis bővítését további korpuszok 1.2 pontban leírtak szerinti feldolgozásával.

Ennek első lépéseként learattuk az Országos Széchenyi Könyvtár online elérhető anyagait, ezek feldolgozásának előkészületei jelenleg folynak.

4.2 Szóalakok generálása

Vizsgáljuk egy ragozómotor kialakításának lehetőségét, mely egy szótőből és egy affixumsorozatból szóalakot képezne. Egy ilyen motorral korpusz nélkül lehetne célzottan bővíteni az adatbázist. A ragozómotor kialakítását segíti, hogy – amint a Bevezetőben is említettük – nem teljességre törekszünk, hanem a gyakorlati felhasználhatóság támogatására.

Az eddigi statisztikák alapján az adatbázis bővítése az eddig talált összes szótővel és alkalmazható affixumsorozattal jelentős, de megfelelő informatikai háttérrel kezelhető feladatnak tűnik.

4.3 Performancia mérése

Az Országos Széchenyi Könyvtár letöltött anyagainak bedolgozása után meg kívánjuk mérni az adatbázis teljességi mutatóit, továbbá működési sebességét. A jelenlegi mé-

retek mellett aggregációs segédtáblázat segítségével egy szálon kb. 9 ezer szó/másodperc sebességet tudtunk elérni.

5 Konklúzió

Az előzőekben ismertetett statikus MySQL adatbázisra épülő megoldás kellő feltöltöttség esetén megvalósítja a kitűzött célokat. Jó kilátások vannak arra, hogy nagy találati arányt adó adatbázist tudjunk építeni.

Bibliográfia

1. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. (2005) 77–85
2. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: Creating open language resources for Hungarian. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (2004)
3. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.:. Web-based frequency dictionaries for medium density languages. In: Proceedings of the 2nd International Workshop on Web as Corpus (ACL-06) (2006) 1–9
4. Héder, M., Farkas, M., Oláh, T., Solt, I.: Sztakipedia – Mashing Up Natural Language Processing, Recommender Systems and Search Engines to Support Wiki Article Editing. In: Proceedings of the AI Mashup Challenge 2011 at Extended Semantic Web Conference (ESWC). Iraklion, Greece (2011)

Lemmaasszociáció és morfológiai jegyek mesterséges neurális hálózatokban

Tóth Ágoston[1], Csernyi Gábor[1]

[1] Debreceni Egyetem, Angol Nyelvészeti Tanszék
{toth.agoston, gabor.csernyi}@arts.unideb.hu

1 Bevezetés

Kutatásunk célja egy lemmatizálást és korlátozott morfológiai elemzést mintaasszociáció segítségével megvalósító mesterséges neurális hálózat implementálása, továbbá a neurális modellezés erősségeinek és nehézségeinek dokumentálása.

2 A kísérleteink

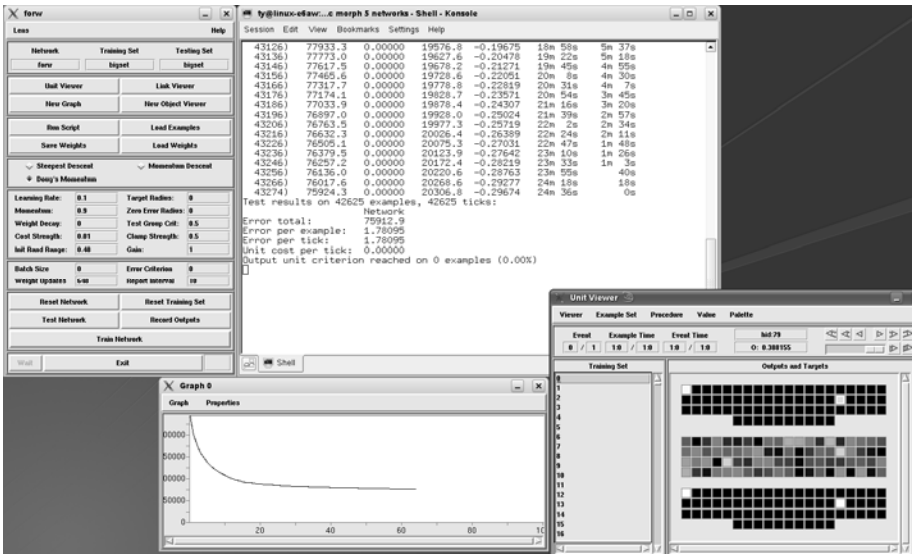
A kísérletekhez használt tanító adatokat a *Magyar Webkorpusz* [1] 100000 leggyakoribb szóalakját tartalmazó listáról nyertük, melyet feldolgozás előtt szűrtünk. Az így előállt, körülbelül 82 ezer szavas szólista 63531 elemére adott a *Hunmorph* [4] legalább egy elemzést. A szóalakokhoz az elemzés során kapott lemmát, valamint kiválasztott (egyelőre korlátozott számú) morfológiai jegyet tanítottunk be.

A kísérleteket neurális hálózatokkal végeztük. A bemeneti rétegen (70 neuron) szóalakokat helyeztünk el egy első alkalommal felhasznált szóreprezentációs technikát használva. Az aktivációk innen egy rejtett rétegbe (80 neuron) haladtak tovább tanítható, súlyozott kapcsolatokat használva, 1:N projekcióval. A rejtett rétegből hasonlóan kialakított kapcsolatok vezettek a kimeneti réteghez, ahol egyrészt 70 neuron végezte a szóalakkal asszociált lemma reprezentációját ugyanazzal a módszerrel, amivel a bemenetet kezeltük (elméletileg végtelen számú szó ábrázolását lehetővé téve), másrészt bizonyos mennyiségű, alapvető morfológiai információkat ábrázoló neuronokat is betanítottunk, az adott kísérlet függvényében. A tanítás a „visszafelé terjesztés" módszerével történt (minden bemenetre képeztük az aktuális súlyokat használva a kimeneteket, kiszámítottuk a teljes hibát, majd a hibát visszafelé terjesztve módosítottuk a súlyokat).

Minden minta (szóalak-lemma pár) legalább 650 alkalommal került betanításra. A bemeneteken és a kimeneteken [0;1] intervallumba eső valós értékek jelentek meg. A kimeneten mind a lemmát, mind a morfológiai jegyeket osztályoztuk a következő módon: a 70 valós értékből álló lemma-kimenetet a legközelebbi ismert lemma célvektornak feleltettük meg, a morfológiai jegyeket pedig 0,4 kimeneti érték alatt 0-nak (jegy hiánya), 0,4-től pedig 1-nek (jegy megléte) osztályoztuk.

A betanítást és a tesztelést a LENS neurális hálózat szimulátorban végeztük [2]. Az 1. ábrán példaként egy hálózat betanításának szimulációs eredményét mutatjuk be, amelyen alul, balra megfigyelhető a hibadiagram, a jobb alsó sarokban pedig a betaní-

tási és tesztelési minták egyenkénti vizsgálatára alkalmas „unit viewer” ablakban az első mintára (az *a* határozott névelőre) kapott aktivációs szintek (alul a bemeneti csoport, fölötte a 80 neuronos „rejtett” réteg, felettük a kimenetek).



1. ábra: LENS képernyőfotó.

Fontos kiemelni, hogy az itt bemutatott kísérleteinkben a többértelműség (az alternatív alaktani elemzések) kezelése komoly problémát okozott már a tervezés fázisától kezdve. Adott keretek közt alternatívák betanítása nem lehetséges, hiszen egy alternatíva jelenléte (azonos inputra különböző kimeneti célok) a betanítást elrontja. Természetesen a valóságban a környezet különbözősége jelenti azt az információt, ami alapján az egyértelműsítés elvégezhető. A morfológiai elemzés szokásos, véges állapotú automatákat használó változata olyan kimenetet ad, amiben az alternatívák mind megjelennek, és egy későbbi mondattani elemzés során ez vagy egyértelműsíthető, vagy további elemzések bevezetéséhez vezet (és ekkor a problémát tovább delegáljuk a szemantikai szintre). A többértelműség kezelésében azonban nem feltétlenül jelent megoldást az összes elemzés visszaadása egy későbbi egyértelműsítés reményében (ahogyan azt a lexikai szemantika vonatkozásában a SenseEval/SemEval versenyekben láthattuk). Éppen ezért a későbbiekben sem az alternatívák enumerációja, hanem a figyelembe vehető paraméterek bővítése (pl. a mondatban szereplő további szavak, morfémák figyelembevétele) és ezek alapján egyértelmű kimenet előállítása a hosszú távú célunk. Jelen rendszerünket úgy terveztük, hogy szófajonként egy elemzést tudunk kezelni; ha egy szó Hunmorph-os elemzése ennek nem felelt meg, akkor kizártuk a kísérletből. Ezen a szűrőn 42625 szóalak ment át, ami a Hunmorph által összesen elemzett 63531 alak 67%-a (ez egyben a felidézési érték, amely mellett rendszerünk Hunmorph-hoz viszonyított pontossága értendő).

A bemeneten megjelenő szóalakok és a kimeneten elvárt lemmák reprezentálására olyan vektorokat képzünk, amelyben az ABC minden betűjének két vektorelem felel

meg. Az egyik azt mutatja meg, hogy az adott betű a szó hányadik karakterpozícióján fordul elő *először*, a másik pedig azt, hogy az adott betű a szó (szó végétől számítva) hányadik karakterpozíción fordul elő *utoljára*. Ha egy szóban egy betű kettőnél többször szerepel, ami nem ritka jelenség, akkor az adott betű első és utolsó előfordulásának helye lesz rögzítve, a többiről nem tárolunk információt. A módszert Tóth [3] javasolta, ahol több reprezentációs eljárás is szerepel, és a módszerek előzetes tesztelését angol írott, angol fonetikusan átírt és magyar szavakon végezte el. Az ottani kísérletekből látszik, hogy a betűk *utolsó* előfordulásának jegyzése önmagában is nagyon hatásos eszköz egy szó felismerésében, de egy további adat (itt: az első előfordulások felhasználása) fokozza az eljárás pontosságát. Ezek a módszerek nem kölcsönösen egyértelmű leképezéseket valósítanak meg, de ha ez az adott felhasználáshoz szükséges, akkor is rendkívül alacsony a hiba. Mostani kísérletünkben 23 olyan szópár volt, melyek olyan szavakból álltak, amelyeknek reprezentációja azonos volt. Ez a jelenség a vizsgált 42625 szónak kevesebb mint 1 ezrelékét érintette, ezért nem tekintettük jelentős hibaforrásnak, és ezeket a szavakat is megtartottuk.

Első kísérletünkben a szófaji felismerést mértük, miközben a kimeneten a lemmát leíró egységek teljesítményét nem figyeltük. A *főnév* jegyet 82%, az *igét* 90%, a *melléknevet* 84%, a *határozószót* 96%, az *egyéb* kategóriát (*névelő*, *kötőszó*, *számnév*, stb.) 97% pontossággal jelezte a rendszer a 42625 szavas szólistán mérve.

Második kísérletünkben öt hálózatot tanítottunk be, ezek sorrendben a főneveket, igéket, mellékneveket, határozókat és végül az egyéb morfológiai kategóriákat kezelték, és *alak–lemma*, valamint *alak–morfológiai jegy* asszociációt végeztek úgy, hogy bemenetükön a szóalakok, a kimenetükön pedig a lemmák és morfológiai jegyek voltak ábrázolva. A *főnévi* hálózat esetében a figyelt jegyek (gyakoriságuk alapján kiválasztva) a többes szám, a birtokos eset és a tárgyeset, az *igei* hálózatban a többes szám, a múlt idő, az 1. és 2. személy, valamint a tárgyas ragozás voltak; a *melléknéveknél* a többes számot vizsgáltuk, a *határozószóknál* nem volt megfigyelt jegy. Az *egyéb* kategóriában (5. hálózat) a Hunmorph további főkategóriáit (*névelő*, *kötőszó*, *számnév* stb., összesen 9 db) azonosítottuk 1-1 neuronnal. Amennyiben a bemeneten megjelent szóalaknak nem volt az adott hálózatnak megfelelő kategóriájú elemzése, a kimeneten a „lemmahiány" lemma megjelenését vártuk, a lemma neuronok egyedi mintázatát figyelve (tehát szintén lemmaasszociációs feladatként); a morfológiai kimenetek ekkor inaktívak voltak. A hálózatokon mért pontosságot az 1-5. táblázatokban foglaltuk össze.

1. táblázat: A főnévi hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány" (= inkompatibilis kat.)	15528	12667	82%
helyes lemma (kivéve: „lemmahiány") (baseline: 1:8297 ≈ 0,01%)	27097	18818	69%
lemmaasszoc. összesen	42625	31486	*74%*
morfológia (27097 főnévre)			*87%-97%*

2. táblázat: Az igei hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	32393	31716	98%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:3102 ≈ 0,03%)	10232	5204	51%
lemmaasszoc. összesen	42625	36920	*87%*
morfológia (10232 igére)			*94%-97%*

3. táblázat: A melléknévi hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	32533	31830	98%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:6325 ≈ 0,02%)	10092	3675	36%
lemmaasszoc. összesen	42625	35505	*83%*
morfológia (1 jegy, 10092 melléknév)			*91%*

4. táblázat: A határozói hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	40448	40380	99%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:2079 ≈ 0,05%)	2177	233	11%
lemmaasszoc. összesen	42625	40613	*95%*

5. táblázat: Az „egyéb” hálózat pontossága a 2. kísérletben.

	Cél (db)	Elért (db)	Pontosság
„lemmahiány” (= inkompatibilis kat.)	41554	41554	100%
helyes lemma (kivéve: „lemmahiány”) (baseline: 1:678 ≈ 0,1%)	1071	8	1%
lemmaasszoc. összesen	42625	41562	*98%*
morfológia (1071 szóalakra)			*80%-99%*

A hálózatok a nem kompatibilis kategóriát, „lemmahiány” lemmát visszaadva, 82-100% pontossággal jelezték. Helyes kategóriájú alak esetén a legközelebbi lemmát 1-69% közötti pontossággal adták vissza. A gyakoribb kategóriák esetén a (létező szavakra utaló) lemmaasszociáció pontossága magasabb volt, lásd a főnévi hálózat adatait. Az adatokból az is látható, hogy a baseline értéket (ami az adott hálózat lemma kimenetén várt összes *különböző* lemmareprezentáció mennyiségével fordítottan arányos) mindegyik hálózat esetében sikerült jelentősen meghaladni. A *határozószó* és *egyéb* kategóriák nagyon kevés alakkal voltak képviselve, az elért alacsony pontosság ennek is köszönhető, ilyenkor azonban a morfológiai inkompatibilist jelző „lemmahiány” állapot visszaadása igen pontos volt. A figyelt morfológiai jegyeket (pl. többes szám, birtokos eset, tárgyeset stb.) meglehetősen jó eredménnyel jelezték a hálózatok, adott jegytől függően tartalmi szavaknál 87-97%, funkciószavaknál 80-

99% pontossággal. További kísérletekben a jegyek köre bővíthető, a skálázhatóság egyelőre nem ismert.

Utolsó kísérletünkben a mintákat véletlenszerűen, ¾ részben tanító és ¼ részben tesztelő adathalmazra osztottuk. A főnévi hálózatot a tanító mintákkal betanítottuk, majd a tesztmintákkal (melyeket a hálózat nem ismert) kiértékeltük. A főnévi elemzések esetén a lemma kimenet 71%, az inkompatibilis kategória („lemmahiány") jelzése pedig 80% pontossággal zajlott, összességében a lemmaasszociáció 74%-ban volt sikeres. A három megfigyelt főnévi morfológiai jegyet 86-96% pontossággal becsülte a rendszer, jegytől függően. Ezeket az adatokat az 1. táblázat főnévi oszlopával összevetve láthatjuk, hogy a hálózat általánosító képessége mind a lemmaasszociáció, mind a morfológiai jegyek tekintetében igen jó (a tesztadatokon mért teljesítmény semmiben sem marad el a tanítón mért pontosságtól), tehát kijelenthetjük, hogy nem a konkrét alakokat, hanem a szabályszerűségeket tanulta meg a hálózat.

Köszönetnyilvánítás

A publikáció elkészítését részben az OTKA (K 72983), részben a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatta az Új Magyarország Fejlesztési Terven keresztül az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával, továbbá támogatta a TÁMOP-4.2.2/B-10/1-2010-0024 projekt az Európai Unió és az Európai Szociális Alap társfinanszírozásával.

Bibliográfia

1. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Kilgarriff, A., Baroni M. (eds.): Proceedings of the 2nd International Workshop on Web as Corpus (2006)
2. Rohde, D. L. T.: LENS: The light, efficient network simulator. Technical Report CMU-CS-99-164. Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA (1999)
3. Tóth, Á.: Perspectives on the Lexicon. Akadémiai Kiadó, Budapest (2008)
4. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software (2005)

Fonológiai jegyek felügyelet nélküli tanulása fonemikus korpuszból

Vásárhelyi Dániel

Eötvös Loránd Tudományegyetem, BTK, Elméleti Nyelvészet Program,
e-mail:vad@budling.hu

Kivonat A modern fonológiai ábrázolás központi eleme a szegmentumok megkülönböztető fonológiai *jegyek*re történő felbontása, ami lehetővé teszi a fonológiai szabályok tömörebb és plauzibilisebb megfogalmazását. Az utóbbi időben többen próbáltak érvelni ezeknek a jegyeknek és kombinatorikus viszonyaiknak, a *jegygeometriának* a veleszületett volta mellett, miközben mások a fonológiának a lexikonból való elsajátíthatósága mellett törnek lándzsát.
Az ismertetendő kutatás célja, hogy a konvex kombinatorikus geometriák algoritmikus jellemzésének legfrissebb eredményeit felhasználva egy memóriaalapú felügyelet nélküli algoritmust adjon a jegygeometria megtanulására, ezzel letéve a garast a lexikalista álláspont mellett általánosságban a nyelvi elemek és speciálisan a fonológiai szegmentumok belső struktúrájának elsajátításában.

Kulcsszavak: korpusznyelvészet, természetesnyelv-feldolgozás, jegygeometria, felügyelet nélküli tanulás

1. Bevezetés

A fonológiai jegyek eloszlásalapú elsajátítása különösen érdekes lehet annak fényében, hogy bizonyos jelenségek, mint például a szonoritásprojekció memóriaalapú magyarázatához a fonémák szubszegmentális ábrázolására van szükség (lásd [2]). Amennyiben a szegmentálás szintén elvégezhető kizárólag a fonológiai input alapján, akkor nincs szükség veleszületett specifikus fonológiai tudás feltételezésére.

2. Jegygeometria

A fonémák, az őket megvalósító fónok artikulációs és akusztikus tulajdonságai alapján, számos jeggyel jellemezhetők, ezek közül az egyes nyelvek választják ki, melyek kontrasztívak, azaz megkülönböztető szerepűek és melyek redundánsak.

Dresher a [3]-ban a kontraszt fonológiai szerepét vizsgálva arra a következtetésre jutott, hogy a nyelvészek a kontrasztivitásnak két egymással inkompatibilis meghatározása között ingadoztak. A teljesen specifikált minimálpárokon alapuló

és a jegyeken hierarchikus struktúrát feltételező megközelítések közül az elsőről meggyőző módon mutatja ki annak tarthatatlanságát.

A nyelvi elemek hierarchikus jegyekkel való ábrázolása az utóbbi időben a nyelvészet más területein is széles körben elterjedt.

3. Antimatroidok

A hierarchikus kapcsolatok egyik legáltalánosabb modellje a *konvex kombinatorikus geometria* vagy a vele ekvivalens *antimatroid*, egy olyan halmazrendszer, amely az alaphalmaz elemeinek egyesével való hozzáadásával (vagy elvételével) megkapható halmazokból áll.

Belátható, hogy a megkülönböztető jegyek rendszere egy antimatroidokból álló rendszert alkot, amelyben a fonémák és azok természetes osztályai mind konvex halmazok.

4. Algoritmusok

A vizsgált korpusz különféle a szerzők által interneten szabadon hozzáférhető magyar nyelvű szövegek saját algoritmussal történő fonetizálásával készült. A korpusz méretének további növelése nem okozott lényeges változást a kutatás eredményeiben.

A korpuszból először annak trigram modelljét állítottuk elő, majd azt követően minden fonémához hozzárendeltük a $\_p_1p_2$, $p_1\_p_2$, $p_1p_2\_$ alakú környezetek egy elmosódott (fuzzy) halmazát olyan módon, hogy egy adott p fonémára a $\_p_1p_2$, $p_1\_p_2$, $p_1p_2\_$ környezetekhez rendre a pp_1p_2, p_1pp_2, p_1p_2p trigramok relatív gyakoriságát rendeltük. Rögzített 0 és 1 közötti küszöbértékre az ennél nagyobb relatív gyakoriságú környezetek halmazt alkotnak és definiálható a fonémák halmazán egy Ψ operátor olyan módon, hogy fonémák tetszőleges U halmazához hozzárendeljük azokat a fonémákat, amelyek környezethalmaza tartalmazza mindazon környezeteket, amelyeket U minden elemének környezethalmaza tartalmaz.

Amennyiben Ψ *izotón*, amelyet az a feltételezés, hogy a fonémák és a környezetek konvexek biztosít, a [4]-ben ismertetett Ψ-algoritmus egy antimatroid rendszert definiál, ami tézisünk szerint éppen a magyar fonológia jegygeometriájával azonos.

5. Eredmények

A kutatás jelenlegi szakaszában a paraméterek beállítása és a kapott antimatroid vizsgálata folyik, ami a teljes halmazrendszer mérete miatt nem egyszerű feladat, ezért a teljes halmazrendszer helyett annak kisebb fonémahalmazokra való megszorítását értékeltük.

Meglehetősen nagy (>0,01) küszöbértékekre a leggyakoribb fonémákra (e, a, t, n, k, l, o) megszorított rendszer meggyőzően egyezik egy lehetséges jegygeometriával, például az $\{e\}$, $\{e, a\}$, $\{e, a, o\}$ konvex halmazok megfeleltethetők egy $voc > back > round$ jegyhierarchiának.

Hivatkozások

1. Ball, Keith.: An Elementary Introduction to Modern Convex Geometry, Flavors of Geometry, MSRI Publications Volume 31, Cambridge, Massachusetts, (1997)
2. Daland, Robert, et al..: Explaining sonority projection effects, Phonology 28, Cambridge University Press, 197–234, (2011)
3. Dresher, B. Elan: The contrastive hierarchy in phonology, Toronto Working Papers in Linguistics, Vol 20, Toronto, 47–62, (2003)
4. Kempner, Yulia, et al.: Correspondance between two antimatroid algorithmic characterizations, The Electronic Journal of Combinatorics (www.combinatorics.org), Vol 10, RR44, (2003)

Szerzői index, névmutató