Statisztikai konstituenselemzés magyar nyelvre

Szántó Zsolt, Farkas Richárd

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

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

Kulcsszavak: konstituenselemzés, morfológiai kódkészlet

1. Bevezetés

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

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

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

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

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

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

2. Korpusz, kiértékelési metrikák

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

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

3. Kiterjesztett lexikai modellek

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Morfológiai kódok automatikus összevonása

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

4.1. Eljárás morfológiai jellemzők értékeinek összevonására

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

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

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

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

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

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

4.2. Baseline preterminális halmazok létrehozása

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

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

4.3. Eredmények különböző preterminális halmazokkal

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

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

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

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

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

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

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

5. Konklúzió

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

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

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

Köszönetnyilvánítás

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

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

Hivatkozások

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

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