Módosított morfológiai egyértelműsítés és integrált konstituenselemzés a magyarlanc 3.0-ban

Farkas Richárd[1], Szántó Zsolt[1], Vincze Veronika[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged Árpád tér 2.
e-mail:{rfarkas,szantozs,zsibrita}@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat Cikkünkben bemutatjuk a magyarlanc programcsomag [1] legújabb fejlesztéseinek számítógépes nyelvészeti szempontból érdekes tanulságait. Röviden érintjük a webes szövegek elemzésére kiterjesztett tokenizálót, majd hosszabb összehasonlítást közlünk a legmodernebb szófaji egyértelműsítők eredményeiről. Végül a szintaktikai elemzés terén vizsgáljuk különböző morfológiai kódkészletek hatását a függőségi elemzésre és bemutatjuk a magyarlancba integrált statisztikai konstituenselemzőnket. A magyarlanc programcsomag ingyenesen elérhető honlapunkon[3].

1. Bevezetés

A magyarlanc programcsomag [1] magyar nyelvű szövegek nyelvi előfeldolgozását hajtja végre a mondatra bontástól a morfológiai elemzésen át a szintaktikai elemzésig. Ebben a cikkben bemutatjuk az elemző lánc legújabb változatát. Először röviden ismertetjük a webes szövegekre (is) optimalizált tokenizálót, majd összehasonlítjuk a különféle szófaji egyértelműsítők teljesítményét. Bemutatjuk azt is, hogy az eltérő kódrendszerek miképpen befolyásolják a morfológiai és függőségi elemzők eredményességét, továbbá végül kitérünk a magyarlancba újonnan integrált konstituenselemző modulra is.

2. Robosztus tokenizáló

A magyarlanc v2.1 mondatra és tokenekre bontó első modulját újraimplementáltuk, majd robosztusabbá tettük, hogy a közösségi média sajátosságait is figyelembe vegye. A két legfontosabb ilyen kiegészítés az URL-ek felismerését célzó reguláris kifejezések beépítése, illetve egy emotikonazonosító szabályrendszer, ezek esetében két korábbi megoldásra építettünk[4,5], de kiegészítettük azokat speciális szabályokkal.

[3] rgai.inf.u-szeged.hu/magyarlanc

[4] https://gist.github.com/uogbuji/705383

[5] https://github.com/twitter/commons/blob/master/src/java/com/twitter/common/text/extractor/EmoticonExtractor.java

Egy fő elvárás a szegmentálóval kapcsolatban, hogy továbbra is a Szeged Korpusz [2] tokenhatárainak megfelelően szegmentáljon annak érdekében, hogy a korpuszon gépi tanított módszerekkel kompatibilis maradjon. A tokenizálót ezért a fejlesztés során folyamatosan lefuttattuk a Szeged Korpuszon és a Szeged Web Korpuszon [3], és annak kimenetét a gold standard tokenhatárokkal összevetettük, majd az esetlegesen szükséges változtatásokat beépítettük a rendszerbe.

3. Morfológiai egyértelműsítő rendszerek

A fejlesztés során módosítottuk a magyarlanc által alkalmazott szófaji egyértelműsítő rendszert. A módosításnak többféle motivációja is volt, egyrészről olyan licencet szerettünk volna használni, amely segítségével a magyarlanc alkalmazható ipari projektek részeként, másrészről az eddigiekben használt maximum entrópia Markov-modellre (MEMM) építő Stanford POS Tagger [4] mellett a legmodernebb szófaji egyértelműsítő rendszerek hatékonyságát is szerettük volna összevetni más szófaji egyértelműsítőkkel.

A kísérleteinkhez a Stanford POS Taggert két másik szófaji egyértelműsítővel hasonlítottuk össze. A PurePOS [5] egy morfológiai elemzővel kiegészített trigramokat használó rejtett Markov-modell (HMM) alapú elemző, míg a MarMoT [6] egy magasrendű feltételes véletlen mezőkre (CRF) építő szófaji egyértelműsítő.

A három elemző a háttérben használt matematikai modell mellett több dologban is eltér, ezek egyike a nyelvi erőforrások használata. A magyarlanc több nyelvi erőforrást is igénybe vesz a szófaji egyértelműsítés folyamatához. A meglévő szófaji címkéket először leképezi egy sokkal kisebb szófajicímke-halmazra (amelyből a szóalak ismeretében egyértelműen visszanyerhető az eredeti címke), majd az elemzés során morfológiai egyértelműsítő használatával szűri le az egyes szóalakokhoz tartozó lehetséges címkéket. Ezzel szemben a PurePos képes hatékony elemzést adni tisztán statisztikai módon, viszont a programban lehetőség van morfológiai elemző bekötésére, amivel tovább javítható a rendszer pontossága. A MarMoTot csak tisztán statisztikai módon, a tanítókorpuszon kívüli bármiféle nyelvi erőforrás használata nélkül alkalmaztuk.

Erőforrások szempontjából bár a kiértékelés mindhárom elemző esetén gyorsnak mondható, a tanítási időben nagy eltérések vannak a rendszerek között. A leggyorsabb a PurePos, amely másodpercek alatt képes egy modellt felépíteni a teljes Szeged Korpuszból. Ez a folyamat a MarMoT esetén azonos hardver mellett pár órát, míg a Stanford POS Tagger esetén napokat vesz igénybe.

3.1. Eredmények a Szeged Korpuszon

A rendszerek doménen belüli hatékonyságának vizsgálatához a Szeged Korpuszt vettük alapul. A Szeged Korpusz mind a 6 alkorpuszát véletlenszerűen felosztottuk 80-20 arányban tanító és kiértékelő korpuszra.

Az 1. táblázat az egyes rendszerek hatékonyságát tartalmazza a Szeged Korpusz egyes doménjein tanítva és kiértékelve. Az eredmények meghatározásához a

1. táblázat. Szófaji egyértelműsítők hatékonysága a Szeged Korpusz alkorpuszain.

	sz. tech.	jog	irodalom	rövidhír	újság	iskolás
magyarlanc	94,08	97,51	95,89	95,92	94,07	96,00
PurePos	94,15	97,09	94,06	97,35	93,63	95,27
Purepos + MA	94,75	97,39	95,90	96,88	94,33	96,01
MarMoT	95,88	97,73	95,74	98,03	95,75	96,32

teljes morfológiai leírás szerinti pontosságot használtuk, azaz mind a fő szófajnak, mind a morfológiai jegyeknek egyezniük kellett. A *magyarlanc* a magyarlancban eddigiekben is használt Stanford POS Tagger eredményeit tartalmazza. A $PurePos + MA$, illetve $PurePos$ a PurePos morfológiai elemzővel kibővített, illetve a nélküli változatát jelölik.

A legjobb eredményeket az – irodalmi szövegek kivételével – minden esetben a MarMoT érte el. Az irodalmi szövegek esetén a PurePos és Stanford POS Tagger holtversenyben végzett az első helyen.

A PurePos esetén átlagosan 0,61 százalékpontot javítva hat esetből ötször szerepelt jobban a morfológiai elemzőt is használó változat. A magyarlánc és a PurePos versenyében az előbbi több esetben tudott jobban szerepelni a morfológiai elemzőt nem használó PurePos változatnál. A morfológiai elemző használata mellett viszont a PurePos három alkorpuszon jobb, kettőn pedig közel azonos eredményt el, mint a magyarlanc.

3.2. Eredmények közösségimédia-szövegeken

A vizsgálatok során cél volt az is, hogy az elemző ne csak előre megszerkesztett (regények, újságcikkek, ...) szövegeken tudjon jól működni, hanem a nyelvi szabályokat sokkal kevésbé betartó internetes közösségi médiából származó szövegeken is hatékonyan működjön. Az elemzők számára a tanítóhalmaztól eltérő domén mellett az is kihívást jelent, hogy ezek a szövegek sokkal kevésbé szerkesztettek és ellenőrzöttek, mint a Szeged Korpuszban található egyéb szövegek. A mondat szerkezetében lévő eltérések mellett a közösségi médiából származó szövegek nagy mennyiségben tartalmazhatnak helyesírási hibákat vagy olyan szóalakokat, amelyek egyáltalán nem jellemzők az irodalmi, újságírói nyelvre.

A vizsgálatainkhoz két, közösségi médiából származó tesztkorpuszt [3] használtunk, mindkét esetben a teljes Szeged Korpuszon tanítottunk. A gyakori kérdések korpusz (faq) a gyakorikerdesek.hu oldalon feltett kérdésekből és arra érkező válaszokból áll, míg a $facebook$ korpusz Facebookról származó bejegyzéseket és a hozzájuk tartozó kommenteket tartalmazza. A két korpusz szerkesztettsége erősen eltér, hiszen míg a gyakori kérdések általában előre átgondolt és megszerkesztett kérdéseket és válaszokat tartalmaz, addig a facebookról származó bejegyzések sokszor csak egy hirtelen jött gondolatot fogalmaznak meg, és az alattuk található kommentek sokkal inkább hasonlítanak valós idejű társalgásra, mint átgondolt és előre megszerkesztett szövegre.

2. táblázat. Szófaji egyértelműsítők hatékonysága közösségi médiából származó szövegeken.

	facebook	faq
magyarlanc	67,17	84,46
PurePos	67,86	86,08
PurePos + MA	70,40	86,61
MarMoT	67,76	87,49

3. táblázat. Szófaji egyértelműsítés és lemmatizáció együttes hatékonysága a közösségi médiából származó szövegeken.

	facebook	faq
magyarlanc	65,00	82,37
PurePos	66,22	85,49
PurePos + MA	66,51	83,37
MarMoT	63,59	84,61

Az egyes rendszerek eredményeit a 2. táblázat tartalmazza. Minden esetben az egész Szeged Korpuszt használtuk tanításhoz és az egyes közösségi média korpuszokon értékeltünk ki. Ezúttal az elemzők sorrendje mindkét korpuszon azonos. A facebook esetén a PurePos teljesített a legjobban, a morfológiai elemzős változat 2,64 százalékponttal ér el jobb eredményt, mint a MarMoT. A gyakori kérdéseken viszont 1 százalékpont alatti különbséggel, de a MarMoT jobban teljesített. A magyarlanc mindkét esetben alulmaradt, ennek az indoka, hogy a rendszer nagyban támaszkodik a morfológiai elemző kimenetére. A morfológiai elemző viszont helyesírási hibák, lemaradt ékezetek esetén sokszor nem tud lehetséges elemzéseket meghatározni, az ilyen esetekben az adott szót mindig *X* (ismeretlen szó) címkével látja el a rendszer.

3.3. Lemmatizáció

A szófaji egyértelműsítés mellett fontos kérdés volt az egyes szóalakokra a megfelelő szótövek meghatározása. A Stanford POS Tagger külön szótövesítésre nem képes. A magyarlanc eddigiekben arra az állításra építve tudta meghatározni a szótöveket, hogy a magyarban a szóalak és a morfológiai címke ismeretében a szótő egyértelműen meghatározható. Így a szótő megadásához egy adott szóalakra a morfológiai elemző által adott lehetséges elemzéseket használtuk.

Ezzel szemben mind a PurePos, mind a MarMoT (Lemming [7]) tartalmaz beépített statisztikai lemmatizálót. A PurePos a lehetséges lemmákat képes szóvégződések alapján statisztikai módon, vagy ha rendelkezésre áll morfológiai elemző, akkor az alapján meghatározni.

A 3. táblázat tartalmazza a szótövesítés eredményeit. Az egyes értékek a teljes morfológiai címke és a szótő együttes eltalálásának a pontosságai. Amennyiben

a címkéket is nézzük, a PurePos mindkét esetben jobban teljesített a MarMoT-nál. Viszont meglepő módon a gyakori kérdéseken jobb eredményeket ért el a morfológiai elemzőt nem használó PurePos, mint az azt használó modell.

4. Morfológiai kódrendszer

Morfológiai címkekészletben áttértünk az ún. univerzális morfológia kategória-rendszerére [8]. Az univerzális morfológia célja – az Univerzális Dependencia Projekt keretében –, hogy egy olyan univerzális, azaz nyelvfüggetlen morfológiai kódkészletet hozzon létre, mely számítógépes nyelvészeti oldalról elősegíti a morfológiai elemzők és szófaji egyértelműsítők fejlesztését, továbbá elméleti nyelvészeti oldalról megkönnyítik az egyes nyelvek kontrasztív morfológiai vizsgálatát. A projekt további célkitűzése, hogy ezen elméleti reprezentációt szorosan követő korpuszokat és treebankeket hozzon létre. Jelenleg 33 nyelvre áll rendelkezésre univerzális dependencia és/vagy morfológiai annotáció, melyek egyike a magyar.

A Szeged Korpusz 2.5-ben használatos morfológiai kódokat automatikusan alakítottuk át az univerzális morfológiai kódkészletre. Ezt a folyamatot részletesebben [8] tárgyalja. Jelen munkánkban azt vizsgáljuk, hogy a két kódrendszer közti eltérések mennyiben befolyásolják a morfológiai és szintaktikai elemzés hatékonyságát. Ennek érdekében a Szeged Korpusz 2.5 kódkészlete és az univerzális morfológia közötti különbségeket empirikus kísérletekkel támasztjuk alá.

Annak érdekében, hogy bizonyos nehezebb nyelvtani jelenségeket külön is megvizsgálhassunk, kézzel összeállítottunk egy mondathalmazt, melyet mindkét kódkészletnek megfelelően beannotáltunk, majd a teljes Szeged Korpusz anyagán tanítva a MarMoT szófaji egyértelműsítőt, automatikusan leelemeztettük a mondatokat. A számszerű eredmények szerint az univerzális morfológián tanítva jobb teljesítményt nyújtott az elemző (92,31%-os pontosság), szemben a 2.5-ös kódkészlettel (91,45%), a különbség azonban nem jelentős. Kíváncsiak voltunk azonban arra is, hogy a két kódrendszer esetében mik a nehézséget jelentő nyelvi jelenségek, így megvizsgáltuk a morfológiai egyértelműsítő rendszer tipikus tévesztéseit.

A gyakorító és műveltető igék elemzése mindkét kódrendszernek kisebb nehézségeket okozott, illetve bizonyos homonim alakok tévesztése is előfordult mindkét kódrendszer esetében (pl. *hozzátok* igei és névmási elemzése). Az univerzális morfológia ugyanakkor helyesen elemzi a kötőszavakat, ellenben a 2.5-ös kódrendszer tévesztéseivel. Itt azonban meg kell említenünk azt a tényt, hogy az univerzális morfológia mindösszesen a kötőszavak alá- vagy mellérendelő jellegét jelöli a morfológiai jegyek között, míg a 2.5-ös kódrendszer azt is jelöli, hogy tagmondatokat vagy frázisokat köt-e össze az adott kötőszó. Természetesen ez utóbbi megkülönböztetés inkább szintaktikai, semmint morfológiai természetű, így egy további érvet szolgáltat az univerzális morfológia használata mellett, hiszen ilyen jellegű megkülönböztetésekre nincs szükség a morfológia szintjén.

A kétfajta kódrendszer hasznosságát megvizsgáltuk aszerint is, hogy mennyire nyújtanak hasznos kimenetet a függőségi elemzéshez. Ehhez a Szeged Treebank Népszava alkorpuszának univerzális dependenciára annotált verzióját hasz-

4. táblázat. Szófaji egyértelműsítés és függőségi elemzés hatékonysága 2.5-ös és univerzális morfológiai kódkészlet mellett.

	LAS	ULA	POS
2.5 kódkészlet	77,41	81,81	91,54
univerzális morfológia	76,23	81,01	92,34

náltuk, melynek 80%-án tanítottuk a magyarlancba beépített Bohnet parsert, és a maradék 20%-án pedig kiértékeltük a rendszert. A tanítás során predikált szófaji elemzést használtunk mind a 2.5-ös kódrendszer, mind az univerzális morfológia esetében. A számszerű eredmények szerint függőségi elemzésre nézve jobb teljesítményt érünk el a 2.5-ös kódkészleten (l. 4. táblázat). A teljes morfológiai kódokat tekintve az univerzális morfológia jobb eredményeket ér el, ami arra enged következtetni, hogy az könnyebben gépi tanulható.

A szintaktikai elemzéseket részletesebben is megvizsgáltuk, így fény derült arra, hogy a fontosabb nyelvtani szerepek (pl. alany, predikátum) azonosításában közel hasonló teljesítményt nyújt a két rendszer. A 2.5-ös morfológia főleg a névutós szerkezetek és az igekötők azonosításában múlta felül az univerzális morfológiát. Az univerzális morfológia előnyei közvetlenül a minőség- és mennyiségjelzők azonosításában mutatkoznak meg, illetve hatékonyabban képes kezelni az alárendelő mellékmondatok több fajtáját is.

A magyarlanc jelenlegi verziójába a nemzetközi trendeknek megfelelően az univerzális morfológiai kódrendszert integráltuk. Jövőbeli terveink között szerepel, hogy a fenti tapasztalatok alapján a szintaktikai elemzés hatékonyságát segítendő a 2.5-ös morfológia egyes jegyeit nyelvfüggő kiegészítésként felvesszük az univerzális morfológiai kódrendszerbe.

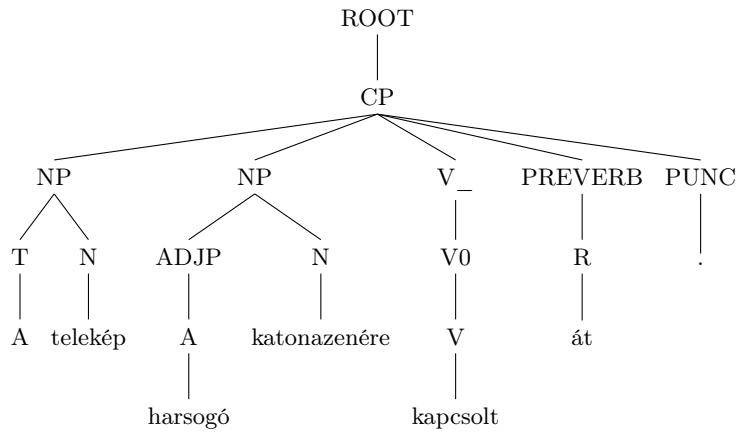
5. Konstituenselemzés

A szintaxis célja a mondatban rejlő nyelvtani kapcsolatok leírása. Az ilyen kapcsolatok megadására több eltérő reprezentáció is létezik. A számítógépes nyelvészetben a két legelterjedtebb reprezentáció a konstituensnyelvtanok és a függőségi nyelvtanok.

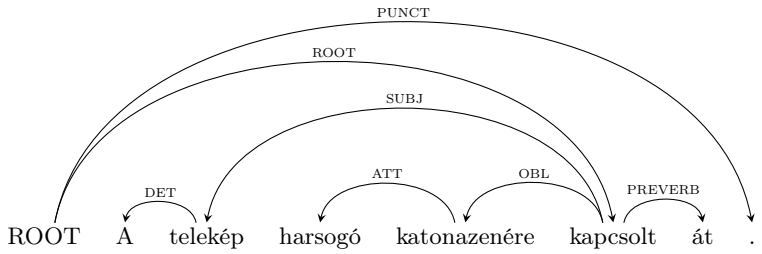
Az 1. ábra egy konstituensfát tartalmaz, amin jól látható, hogy a mondat úgynevezett konstituensekre van bontva, ezek a más szóval frázisoknak hívott egységek csoportba foglalják a szavakat (pl: NP – főnévi csoport). A fában a szavak a leveleken helyezkednek, a szófajok az úgynevezett preterminális rétegen a szavak felett, és e felett találhatók az egyes frázisok.

Ezzel szemben a függőségi fák (2. ábra) esetén a fa minden pontja egy szó és az élek a szavak közötti kapcsolatokat írják le.

A magyarlanc a korábbiakban már képes volt függőségi elemzések meghatározására, amihez a Bohnet parser [9] nevű nyelvfüggetlen függőségi elemzőt használta, a Szeged Dependencia Treebanken betanítva. Az új verzióban egy konstituenselemző modullal bővítettük a magyarlancot.



1. ábra. Konstituensfa.



2. ábra. Függőségi fa.

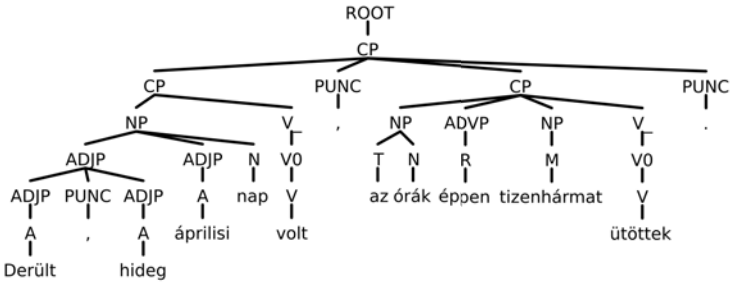
A magyar nyelv szintaktikai elemzése során probléma, hogy a meglévő rendszerek az angol nyelv igényeit figyelembe véve készültek. A magyar és az angol nyelv több szempontból nézve is alapjaiban különbözik. Az angol esetén a szintaktikai információk általában a szórendben tárolódnak, ezzel szemben a magyar nyelven a szintaxis a szavak szintjén jelenik meg toldalékok formájában. A konstituenselemzésre leggyakrabban alkalmazott valószínűségi környezetfüggetlen nyelvtanokra építő elemzők hatékonyságát nagyban rontja a toldalékolás következtében bekövetkező magas szóalakszám.

A morfológiailag gazdag nyelvek, köztük a magyar szintaktikai elemzésére hozták létre a Statistical Parsing of Morphologically Rich Languages workshop sorozatot. A magyarlancba készített rendszer a workshop keretében megrendezett SPMRL 2014 Shared Task [10] első helyezést elért rendszere [11] által bemutatott technikákra épül. Az elemző alapja a valószínűségi környezetfüggetlen nyelvtanokat alkalmazó Berkeley Parser [12].

A szóalakok nagy számának kezelésére a tanítóhalmazon nem, vagy csak ritkán látott szóalakokat lecseréljük a szófaji egyértelműsítés során megkapott fő szófaji kódra. A Berkeley Parser tanítása során kis mértékben szerepe van a véletlennek is, ennek a véletlennek a kiküszöbölésére 8 különböző modellt tanítottunk (eltérő random seed mellett), és predikáláskor a különböző modellek által egy mondatra adott valószínűségek szorzatát vesszük, így kiátlagolva a véletlen szerepét.

A valószínűségi környezetfüggetlen nyelvtanok hatékonyságának javítására úgynevezett újrarangsoroló rendszereket szoktak alkalmazni. Az alapgondolat az, hogy míg a környezetfüggetlen nyelvtannak az összes lehetséges elemzés közül kell választania, addig az általa választott legjobb k elemzésből egy lassú diszkriminatív gazdag jellemzőkészlettel rendelkező elemzővel kiválasztjuk a legjobbat. A magyarlancba is készítettünk egy újrarangsoroló rendszert, amit a területen általánosnak számító jellemzőkészletek [13,14] mellett morfológiai alapú jellemzőkkel bővítettünk [15]. Az így kapott rendszer a legaktuálisabbnak számít magyar nyelvű szövegek konstituenselemzésében.

A függőségi elemzéshez hasonlóan a magyarlanc képes vizuálisan is megjeleníteni a konstituenselemző által elkészített fákat. A megjelenítéshez a ParseTreeApplication[6] nevű fa vizualizációs szoftvert használtuk fel, a 3. ábra a magyarlanc egy példa kimenetét tartalmazza.



3. ábra. Konstituensfa a magyarlánc kimenetében.

6. Összegzés

Cikkünkben bemutattuk a magyarlanc programcsomag legújabb, 3.0 verzióját. Ennek része a webes szövegek elemzésre kiterjesztett tokenizáló, továbbá beépítettük a PurePOS morfológiai egyértelműsítőt és integráltunk egy konstituenselemző rendszert, amit morfológiailag gazdag nyelvek elemzésére dolgoztunk ki.

[6] https://github.com/ktrnka/ParseTreeApplication

A magyarlanc programcsomag ingyenesen elérhető: http://rgai.inf.u-szeged.hu/magyarlanc.

Köszönetnyilvánítás

Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Hivatkozások

1. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013) 763–771
2. Alexin, Z., Gyimóthy, T., Hatvani, C., Tihanyi, L., Csirik, J., Bibok, K., Prószéky, G.: Manually annotated Hungarian corpus. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2, Association for Computational Linguistics (2003) 53–56
3. Vincze, V., Varga, V., Papp, P.A., Simkó, K.I., Zsibrita, J., Farkas, R.: Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, Szegedi Tudományegyetem (2015) 122–132
4. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 173–180
5. Orosz, Gy., Novak, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013) 539–545
6. Müller, T., Schmid, H., Schütze, H.: Efficient Higher-Order CRFs for Morphological Tagging. In: Proceedings of EMNLP. (2013)
7. Müller, T., Cotterell, R., Fraser, A., Schütze, H.: Joint Lemmatization and Morphological Tagging with Lemming. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics (2015) 2268–2274
8. Vincze, V., Farkas, R., Simkó, K.I., Szántó, Z., Varga, V.: Univerzális morfológia és dependencia magyar nyelvre. In: XII. Magyar Számítógépes Nyelvészeti Konferencia. (2016) 322–329
9. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. (2010) 89–97
10. Seddah, D., Kübler, S., Tsarfaty, R.: Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. (2014) 103–109
11. Björkelund, A., Özlem Çetinoğlu, Faleńska, A., Farkas, R., Müller, T., Seeker, W., Szántó, Zs.: Introducing the IMS-Wroclaw-Szeged-CIS entry at the SPMRL 2014

Shared Task: Reranking and Morpho-syntax meet Unlabeled Data. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. (2014) 97–102
12. 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 the 44th annual meeting of the Association for Computational Linguistics. (2006) 433–440
13. Collins, M.: Discriminative Reranking for Natural Language Parsing. In: Proceedings of the Seventeenth International Conference on Machine Learning. ICML '00 (2000) 175–182
14. 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
15. Szántó, Zs., Farkas, R.: Special techniques for constituent parsing of morphologically rich languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden (2014) 135–144