Gondolatok a (magyar) statisztikai szintaktikai elemzőkről

Farkas Richárd

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
rfarkas@inf.u-szeged.hu

Kivonat: Jelen munkában áttekintést adunk a statisztikai szintaktikai elemzés nemzetközi állapotáról, a magyar szintaktikai elemzők fejlesztéséhez hasznos szempontok szem előtt tartásával. Négy kérdéscsoportot tárgyalunk bővebben: (i) Mi a különbség és hasonlóság különböző nyelvek szintaktikai elemzése/elemezhetősége közt? Érvelünk amellett, hogy a magyar nyelvre kidolgozott elemzési módszerek más nyelvek elemzéséhez is hasznos tanulságokkal szolgálhatnak. (ii) Tárgyaljuk, hogy a magyar nyelv statisztikai szintaktikai elemzése nem mondható nehezebbnek, mint bármely más nyelvé, de az elemzők továbbfejlesztéshez nyelvspecifikus módszerek kidolgozása szükséges. (iii) Általánosságban összevetjük továbbá a két legelterjedtebb szintaktikai reprezentációt, a konstituens- és függőségi reprezentációkat és (iv) érvelünk a belső kiértékelési metrikák kizárólagos használata ellen.

1 Bevezetés

A szintaxis a szavak szószerkezetekké és mondatokká kapcsolódásának szabályait írja le. Az egyes mondatok szintaktikai elemzése igen fontos bemenete számos számítógépes nyelvészeti alkalmazásnak, mint például információkinyerés, véleménydetekció, kivonatolás vagy gépi fordítás.

Jelen munkában rövid áttekintést adunk a statisztikai szintaktikai elemzés nemzetközi állapotáról, majd részletesen tárgyalunk négy témát, amelyek – véleményünk szerint – a közeljövő statisztikai szintaktikai kutatásait uralni fogják és a magyar szintaktikai elemzők fejlesztésének szempontjából is alapvető fontosságúak.

Statisztikai elemző alatt *adatvezérelt* (data-driven) elemzőket értünk, azaz olyan megközelítéseket, ahol a nyelvtani mintázatok egy kézzel annotált korpusz (adat) formájában állnak rendelkezésre és a cél olyan elemző készítése, amely a korpusz elemzéseit próbálja automatikusan reprodukálni. Jelen munkában kizárólag ezekre koncentrálunk és nem célunk, hogy ezeket a kézzel írt nyelvtanokkal összehasonlítsa. A tanulmány aktualitását az adja, hogy míg a nemzetközi porondon a statisztikai szintaktikai elemzők túlnyomó többségben vannak, addig magyarra csak néhány ilyen kezdeményezést ismerünk. Reményeink szerint ez a tanulmány hozzájárul statisztikai technikák kiaknázásához, magyar nyelvre történő adaptálásához vagy kidolgozásához.

2 Miért érdekes a magyar nyelv szintaktikai elemzése a nemzetközi kutatásban?

Az elmúlt két évtizedben az angol szintaktikai elemzők látványosan fejlődtek, azonban szinte minden statisztikai szintaktikai elemző módszer angol nyelvre lett kidolgozva. Ugyan az elmúlt években erősödő trend, hogy egyéb nyelvek szintaktikai elemzését is vizsgáljuk, az angolra kidolgozott módszerek adaptációja közel sem triviális, gondoljunk csak olyan nyelvekre, ahol a szavak összetett belső szerkezettel rendelkeznek vagy a szórend szabad [1].

A természetes nyelveket szintaktikai elemzés szempontjából egy ún. konfigurációs tengelyre helyezhetjük fel. A spektrum egyik végén az angol mint erősen konfigurációs nyelv található, míg a másik végén a magyar (a tagalog és warlpiri mellett), ahol a legtöbb mondatszintű szintaktikai információt a morfológia kódolja. De természetesen még a végleteken sem beszélhetünk tisztán konfigurációs vagy csak tisztán nemkonfigurációs nyelvekről, hiszen a morfológia az angolban is fontos szerepet játszik, és a magyarban is vannak szórendi megkötések.

A konfigurációs (vagy morfológiai gazdagsági) spektrum szempontjából az egyes nyelveket három szorosan kapcsolódó jelenség mentén érdemes vizsgálni. A *morfológiai gazdagság*ot mérhetjük azzal, hogy egy adott szónak hány morfológiai alakja lehetséges. Például egy főnévnek angolban 2, németben 8, míg magyarban több száz formája lehet. A *szinkretizmus* szintje mérhető azzal, hogy ugyanazon szóalak hány különféle morfológiai alaknak feleltethető meg. Végül a *konfigurációsság* szintje arra vonatkozik, hogy a szavak és frázisok sorrendjének mennyire erős szerepe van a szintaktikai kapcsolatok reprezentálásában. Az angol erősen konfigurációs nyelv, a szórend meghatározza az egyes főnévi csoportok nyelvtani szerepét, míg a magyarban szinte bármilyen sorrend lehet nyelvtanilag helyes [2].

A konfigurációsság és a morfológiai gazdagság közötti negatív korreláció nyilvánvaló. A gazdag morfológia jelöli a nyelvtani szerepeket, nem szükséges azokat még a szórenddel is kifejezni. Másrészről, ha a morfológia nem ad elég támpontot, akkor a konfigurációs nyelvek a szórend rögzítésével tudják az egyes nyelvtani szerepeket kifejezni. Például angolban az igét követő főnévi csoport a tárgy, így nem szükséges azt a morfológia szintjén is jelölni. A szinkretizmus egy köztes megoldásnak is tekinthető a gazdag morfológia és a szórend rögzítése között. Segítségével egy gazdagabb morfológia kevesebb – igaz, többértelmű – felszíni formával kifejezhető. A többértelműség pedig feloldható szórendi jelek alapján (amelyek a kötött szórendnél kevésbé szigorú szabályokat használnak).

A fenti gondolatmenet alapján a morfológiailag gazdag(abb) nyelvekre fejlesztett szintaktikai elemzők legnagyobb kihívása, hogy hogyan valósíthatnak meg az angol rendszereknél erősebb együttműködést a morfológiai elemzés és a szintaktikai elemzés közt. Nyitott kutatási kérdések [3], hogy

- Milyen morfológiai információkat érdemes felhasználni a szintaktikai elemzéshez?
- Hogyan érdemes a morfológiai információt reprezentálni (a szófaji kódok, frázisok, függőségi élek szintjén)?

- Hogyan hatnak egymásra a morfológiai és szintaktikai jelenségek és hogyan lehet ezek kölcsönhatását hatékonyan kiaknázni?
- Hogyan kezeljük az ismeretlen szóalakokat, amelyek nagyon gyakran csak egy ismert szó korábban nem látott morfológiai formája?

Ezeknek a kérdéseknek a vizsgálata kapcsán a magyar mint állatorvosi ló érdekes szerepet tölthet be a morfológiailag gazdag nyelvekre kidolgozott módszerek tesztelésében. Sőt érdekes tanulságokkal szolgálhatnak a konfigurációs spektrum közepén helyet foglaló nyelvek, mint például a német számára is [2].

3 Nehéz-e a magyar szintaktikai elemzés?

A szakmai közbeszédben gyakran hallunk olyan kijelentéseket, hogy egyik vagy másik nyelv szintaktikai elemzése „nehezebb" feladat, mint másiké. Ráadásul számítógépes nyelvészeti körökben ezt a statisztikai elemzők által elért pontosságnak szokták megfeleltetni. Például a magyarról a „CoNLL-2007 többnyelvű függőségi elemzés" verseny [4] óta az volt a közgondolkodás, hogy a magyar szintaktikai elemzés egy nagyon nehéz feladat, mivel a legjobb rendszerek közel 10 százalékponttal rosszabb eredményeket értek el a magyar korpuszon, mint az angolon.

Véleményünk szerint ezekből a számokból nem szabad messzemenő következtetéseket levonni. A pontosságmetrikák közvetlen összehasonlítása például azonnal megkérdőjelezhető, ha arra gondolunk, hogy egy angol mondat 20%-kal több szót tartalmaz, mint egy magyar, és ennek a többletnek (elöljárók, személyes névmás) az elemzése relatíve egyszerű.

A [5] munkában megmutattuk, hogy magyar függőségi korpuszon is elérhető az angolhoz közeli eredmény:

1. táblázat: State-of-the-art függőségi elemzés eredményei magyar és angol nyelven. „dev" és „test" két különböző kiértékelési alkorpusz. LAS (labeled attachment score): a token szülőjének és élcímkéjének is helyesnek kell lennie. ULA (unlabeled attachment score): az élcímkézés nem releváns. Az értékek zárójelben etalon szófaji kódok alkalmazása mellett.

		ULA	LAS
Szeged Dependencia Treebank	dev	89,7 (91,1)	86,8 (89,0)
	test	90,1 (91,5)	87,2 (89,4)
CoNLL-2009 angol korpusz	dev	91,6 (92,7)	88,5 (90,0)
	test	92,6 (93,4)	90,3 (91,5)

Ez annak tudható be, hogy magyarra a morfológiai elemző [6] és egyértelműsítő igen jó hatékonysági fokkal működik, és ahogyan azt az előző fejezetben is tárgyaltuk, a magyarban a morfológia kódolja a nyelvtani szerepek jelentős részét, így a szintaktikai elemzés viszonylag egyszerű feladatnak mondható.

Véleményünk szerint a statisztikai elemzők (mind konstituens, mind dependencia) mára elérték azt a fejlettségi szintet, hogy algoritmikus, nyelvfüggetlen javításokkal

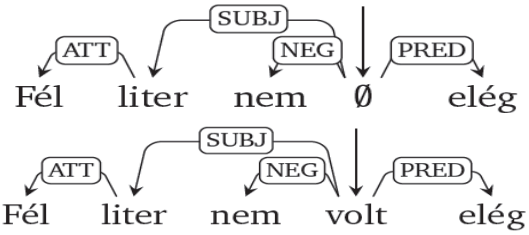
már jelentősen nem javíthatóak. Ehelyett az egyes nyelvek (és annotációs irányelvek) sajátosságait figyelembe vevő megoldások szükségesek. Az [5] munka keretében szisztematikusan elemeztük az angol és magyar függőségi elemzők hibáit és megmutattuk, hogy a hibák nagy része nyelvspecifikus.

3.1 Statisztikai elemzők a CoNLL-2007 és a Szeged Dependencia Treebankeken

Az [5] cikkben azt is tárgyaltuk, hogy a különbség a 2007-es és a 2012-es eredmények közt az annotáció különbségeinek tudható be. Egyrészt míg 2007-ben a frázisstruktúrákból automatikusan konvertált függőségi fák álltak csak rendelkezésre, 2011-re elkészült a Szeged Dependencia Treebank [7], amelyben az automatikus konverzió kimenetét manuálisan javították. A kézi javítás – elsősorban a mellérendelések újfajta kezelésének és a melléknévi frázisok belső szerkezetének köszönhetően – egy tisztább tanuló és kiértékelő adatbázist eredményezett. Másrészt maga az annotációs séma is megváltozott. Például a korábbi 49 élcímke helyett csak 29 szerepel a 2011-es korpuszban (elkerülve, hogy a nyelvtani szerepek duplán, az esetragokban és az élcímkéken is jelölve legyenek).

3.2 Függőségi elemzés virtuális csomópontokkal

Az [5] hibaelemzése szintén rávilágított arra, hogy a Szeged Dependencia Treebank virtuális csomópontjai okozzák a legtöbb problémát a statisztikai elemzőknek. A korpuszépítés során virtuális csomópontok kerültek beszúrásra abban az esetben, ha a létige (kijelentő mód jelen idő E/3. alak) nem jelenik meg a felszínen, illetve elliptikus összetételekben. A virtuális csomópontok az igét helyettesítik ezeken a helyeken. A virtuális csomópontok bevezetésének motivációja kettős. Egyrészt a függőségi elemzők számára (is) az ige a mondat központi eleme, így az ige nélküli mondatokkal nagyon nehezen boldogulnának. Másrészt a szintaktikai elemzést felhasználó célalkalmazások (például gépi fordítás) számára hasznos, ha például jelen és múlt idejű szerkezetek ugyanolyan struktúrában jelennek meg. Erre példa az alábbi két elemzési fa is:



Mivel a függőségi elemzés (fa) csomópontjai általában a mondat szavai, ezért az – angol nyelvet szem előtt tartva kidolgozott – elemzők nincsenek felkészítve arra, hogy új csomópontot legyenek képesek beszúrni az elemzés folyamán. A [8] munka keretében kidolgoztunk és összehasonlítottunk három eljárást a virtuális csomópontok automatikus beszúrására:

- előfeldolgozó: a nyers mondatba szúrunk be virtuális tokeneket, majd a sztenderd elemzőt alkalmazzuk. Ahhoz, hogy eldöntsük, hova érdemes beszúrni virtuális tokent, azonosítjuk a tagmondat-hierarchiát és lényegében igét nem tartalmazó tagmondatokat keresünk
- tranzakcióalapú elemző: felvettünk egy új átmenetet, ami képes új csomópontokat beszúrni
- virtuális csomópontok kódolása élcímkéken: itt a fát átalakítjuk úgy, hogy a virtuális csomópontok gyerekeinek a szülője a virtuális pont szülője lesz, élcímkéje pedig a két él címkéje összefűzve. Ezen a fán sztenderd elemzőket taníthatunk, majd azok kimenete alapján, az összetett címkék helyére virtuális csomópontot tudunk beszúrni.

A kísérleteket a Szeged Dependencia Treebanken és a német Tiger Treebank dependenciaverzióján végeztük el. Azt a következtetést vontuk le, hogy az előfeldolgozós módszer alulmarad a másik két módszerrel szemben, de az nem egyértelmű, hogy a kibővített tranzakcióalapú vagy az éleken kódolós módszer minden esetben jobb lenne a másiknál. A [8] eredményei azt is megmutatták, hogy a lokális jelenségek – mint a létige ki nem fejeződése – jó hatékonysággal megoldható problémák, míg azoknál az eseteknél (például ellipszis), ahol távoli függőségek azonosítása szükséges, a statisztikai módszerek igen alacsony pontosságot tudtak elérni.

Habár virtuális csomópontok beszúrása az angolban is szükséges lenne, ezek az esetek annyira ritkák, hogy nem foglalkoznak velük az elemzők. A virtuális csomópontok kérdése tehát egy jó példa arra, hogy (i) az angolcentrikus elemzőket nem lehet egyszerűen adaptálni egyéb nyelvekre, illetve (ii) hogy a magyar korpusz alapján kidolgozott megoldások hasznosak lehetnek egyéb nyelvek statisztikai elemzőinek kidolgozásához.

4 Függőségi vagy konstituensalapú elemző?

A létező számos szintaktikai reprezentáció közül a statisztikai módszerek túlnyomó többsége konstituens- vagy függőségi reprezentáción alapul. A kettő közül is az elmúlt 6-7 évben a függőségi elemzők lettek a divatos(abb)ak, annak ellenére, hogy semmi sem bizonyítja, hogy a függőségi elemzők jobbak vagy hasznosabbak lennének, mint a konstituenselemzők, mint ahogy azt sem, hogy kevésbé jók vagy hasznosak.

A függőségi reprezentáció előnyeként azt szokták felhozni, hogy abban a nemprojektív élek (nem folytonos konstituensek), illetve a nyelvtani szerepek egyszerűen ábrázolhatóak. Azonban ez nem vonja maga után, hogy az elemzők képesek is lennének ezt kielégítő pontossággal automatikusan reprodukálni. Például a legtöbb függőségi elemző első lépésben egy projektív elemzést generál, majd egy különálló második lépésben utófeldolgozva a fát kap nem projektív elemzéseket. Hasonló utófeldolgozási eljárás alkalmazható lenne konstituensfákon is [2]. Ráadásul vannak olyan nyelvi jelenségek is, amelyeknek viszont a konstituensreprezentáció a természetesebb módja. Ilyenek a mellérendelések, a tagmondatok hierarchikus viszonya és a frázishatárok.

Fontos megjegyeznünk, hogy ugyanakkor mindezen nyelvi jelenségek reprezentálhatóak mindkét megközelítésben [9].

A függőségi elemzők tényleges nagy előnye a sebességük. A szabadon elérhető függőségielemző-implementációk nagyságrendileg húszszor gyorsabbak, mint a konstituenselemzők. Ha az elemzők aszimptotikus időkomplexitását nézzük, mindkét reprezentációhoz léteznek lineáris idejű inkrementális elemzők. A gyakorlatban azonban a nyelvtan mérete miatt a keresési tér sokszorosa a konstituenselemzőknél a függőségi elemzőkéhez képest.

A sebességnek azonban ára van. A konstituenselemzők pontosabbak, mint a függőségi elemzők. Erre épül például az „uptraining" eljárás [10] is, ami a lassú konstituenselemzők kimenetéből a gyors, de gyengébb függőségi elemzőnek tanítópéldákat generál. Többen megmutatták (például [10]), hogy ha a konstituenselemző kimenetét átkonvertáljuk függőségi fákká, jobb eredményt kapunk, mint a legjobb függőségi elemzők. Nyitott kérdés azonban, hogy mi ennek az oka:

- Empirikus eredmények csak angolra és kínaira lettek publikálva. A konstituenselemzők fölénye csak a konfigurációs nyelvek jellegzetessége?
- Angolra a függőségi elemzések a konstituensfák automatikus konverziójából születnek. A konverzió zajos vagy információvesztéssel jár?
- A konstituensreprezentáció algoritmikusan jobban tanulható?
- A függőségi elemzők még csak a tinédzser éveiket élik és néhány év múlva pontosságban is utolérik a konstituenselemzőket?

Az utolsó ponthoz kapcsolódóan megjegyezzük, hogy az összehasonlításhoz használt konstituenselemzők[1] 6-7 évvel ezelőttiek, azóta a konstituenselemzők is rengeteget fejlődtek (habár ezek a fejlesztések nem érhetőek el szabadon letölthető kód formájában). Például a [11] munkában mi is bemutattunk egy újszerű módszert, az erdőalapú rangsoroló elemzőnket, amely angol és német nyelvre is 5% hibacsökkenést eredményezett az eddigi legjobb elemzőkhöz képest.

A konstituenselemzés és függőségi elemzés radikálisan különböző módszerek alkalmazását követeli meg. Kidolgoztunk egy hibrid elemzőt is, amely a két megközelítés különbségeit aknázza ki [12]. A módszer jellemzőket nyer ki az konstituenselemzés kimenetéből, amelyeket felhasznál a függőségi elemzés folyamán (és vice versa). Az eljárás meglepően sokat javít a legjobb függőségi elemzőkön,13% hibacsökkenés a függőségi elemzőkhöz és 6% hibacsökkenés a konstituensből konvertált elemzésekhez képest.

5 Mikor jó egy elemző?

Napjainkig a statisztikai elemzőket szinte kizárólag egy – a tanító adatbázishoz lehető legjobban hasonló – kiértékelő adatbázison értékelték/értékelik ki valamilyen metrika alkalmazásával. Ezzel szemben, ha a gyakorlatban szeretnénk szintaktikai elemzést végezni, akkor (i) a célszövegek valószínűleg számos jellemzőjükben eltérnek a tanító

[1] A Brown parser (2005) és a Berkeley Parser (2006)

adatbáziseitól és (ii) a szintaktikai elemzés célja, hogy valamilyen magasabb szintű feladathoz hasznos bemenetet szolgáltasson, míg az alkalmazott mesterséges kiértékelési metrikák nem képesek a *hasznosság*ot mérni. Valós életbeli alkalmazhatóság szempontjából egy elemző akkor „jó”, ha robosztus (azaz különböző típusú vagy különböző forrásból érkező szövegeken is jól működik) és hasznos bemenetet szolgáltat a végalkalmazásoknak.

Véleményünk szerint a fenti két probléma jelentős figyelmet fog kapni a jövőben. Az (i) problémára a doménadaptációs technikák adhatnak megoldást. Például a [13] munkában bemutattuk webes szövegek elemzésére automatikusan adaptált szintaktikai elemzőinket. A (ii) problémával kapcsolatosan jelenleg is folytatunk kísérleteket. Célunk olyan technikák megtalálása, amivel – a szokásos metrikák helyett – egy célfeladatra – jelen esetben a gépi fordítás átrendezési feladatára – tudjuk optimalizálni a szintaktikai elemzőt. Egy ilyen egyszerű technika a *célzott öntanulás* [14]. Itt egy szintaktikai elemző egy mondathoz tartozó 100 legjobb elemzését kiértékeljük a célfeladathoz való hasznosság szerint (konkrét példánknál az elemzési fák alapján átrendezzük a forrásmondat szavait, majd az átrendezés jóságát számszerűsítjük egy párhuzamos korpusz automatikus szóösszerendelései alapján), majd a leghasznosabb elemzést mint tanítópéldát felhasználva újratanítjuk a szintaktikai elemzőt. Azt kapjuk eredményül, hogy míg a belső metrikák szerint az elemzések rosszabbak, a célfeladat számára azok mégis hasznosabbak.

6 Konklúzió

Jelen munkában tárgyaltuk a statisztikai szintaktikai elemzés fontosabb nyitott kutatási kérdéseit, a magyar szintaktikai elemzők fejlesztéséhez hasznos szempontok szem előtt tartásával.

Bemutattuk a tipológia ún. konfigurációs tengelyét, amelynek egyik végén az erősen konfigurációs angol, míg másik végen a szabad szórendű magyar található. Érveltünk amellett, hogy a magyar nyelvre kidolgozott elemzési módszerek, a gazdag morfológia miatt más – a konfigurációs spektrum közepére elhelyezhető – nyelvek elemzéséhez is hasznos tanulságokkal szolgálhatnak.

Bemutattuk azt, hogy a magyar nyelv statisztikai szintaktikai elemzése nem mondható nehezebbnek, mint bármely más nyelvé, de az elemzők továbbfejlesztéshez nyelvspecifikus, illetve annotációs irányelvekre specifikus problémák megoldása szükséges. Mivel a Szeged Dependencia Treebank statisztikai elemzése kapcsán azt láttuk, hogy a virtuális csomópontok kezelése egy igen gyakori hibaforrás, ezért kidolgoztunk három különböző módszert a virtuális csomópontok automatikus beszúrására.

Tárgyaltuk továbbá a két legelterjedtebb szintaktikai reprezentációt, a konstituens- és függőségi reprezentáció előnyeit és hátrányait. Nem törtünk lándzsát egyik megközelítés mellett sem, célunk az volt, hogy rávilágítsunk: ezidáig senki sem bizonyította, hogy egyik módszer előnyösebb lenne, mint a másik. Bemutattuk továbbá hibrid szintaktikai elemzőnket, amely a két módszer különbözőségeit aknázza ki.

Végül röviden érveltünk a belső kiértékelési metrikák ellen, hiszen a valós életben a tanító adatbázis szövegeitől eltérő szövegeket kell elemeznünk és a végcélunk nem egy jó elemzés elkészítése, hanem olyan elemzések produkálása, amelyek hasznos bemenetül szolgálnak egy célalkalmazás számára.

Köszönetnyilvánítás

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

Hivatkozások

1. Tsarfaty, R., Seddah, D., Kübler, S., Nivre, J.: Parsing Morphologically Rich Languages: Introduction to the Special Issue. Computational Linguistics (megjelenés előtt)
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 (megjelenés előtt)
3. Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., Tounsi, L.: Statistical parsing of morphologically rich languages (spmrl): What, how and whither. In: Proceedings of the NAACL Workshop on Statistical Parsing of Morphologically Rich Languages (2010)
4. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007 (2007)
5. Farkas, R., Vincze, V., Schmid, H.: Dependency Parsing of Hungarian: Baseline Results and Challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012) (2012)
6. Trón, V., Halácsy, P., Rebrus, P., Rung, A,, Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation (2006)
7. 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)
8. Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven Dependency Parsing With Empty Heads. In: Proceedings of the 24th International Conference on Computational Linguistics (2012)
9. Rambow, O.: The Simple Truth about Dependency and Phrase Structure Representations: An Opinion Piece. In: Proceedings of HLT-NAACL (2010)
10. Petrov, S., Chang, P.-C., Ringgaard, M., Alshawi, H.: Uptraining for Accurate Deterministic Question Parsing. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP) (2010)
11. Farkas, R., Schmid, H.: Forest Reranking through Subtree Ranking. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2012) (2012)

12. Farkas, R., Bohnet, B.: Stacking of Dependency and Phrase Structure Parsers. In: Proceedings of the 24th International Conference on Computational Linguistics (2012)
13. Bohnet, B., Farkas, R., Çetinoğlu, Ö.: SANCL 2012 Shared Task: The IMS System. In: Description Notes of the 2012 Shared Task on Parsing the Web (2012)
14. Katz-Brown, J., Petrov, S., McDonald, R., Och, F., Talbot, D., Ichikawa, H., Seno, M., Kazawa, H.: Training a Parser for Machine Translation Reordering. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP) (2011)