Szintaktikai címkekészletek hatása az elemzés eredményességére

Simkó Katalin Ilona[1,2], Kovács Viktória[2], Vincze Veronika[1,3]

[1]Szegedi Tudományegyetem, Informatikai Intézet,
Szeged, Árpád tér 2.
simko@hung.u-szeged.hu
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék,
Szeged, Egyetem u. 2.
viki921015@hotmail.com
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Cikkünkben az univerzális dependencia szintaxis címkekészlet változtatásainak a szintaktikai elemzés közvetlen és a szintaxist felhasználó alkalmazások által elért eredmények változására gyakorolt hatását vizsgáljuk három kísérlet keretében. Megvizsgáljuk a határozói-, az alárendelő mellékmondati-, és a funkciócímkék hatását a standard kiértékelési metrikákkal elért eredményekre, a fő, tartalmascímkék helyes felismerésére, valamint egy adott alkalmazás eredményeire.
Kulcsszavak: szintaxis, dependencia, címkekészlet, kiértékelés

1. Bevezetés

A szintaktikai leírások között ma már nem csak elméletben, hanem a számítógépes gyakorlatban is egyre több alternatíva közül választhatunk. Már a magyarra is léteznek konstituens [1], dependencia [2] és LFG [3] nyelvtani számítógépes nyelvészeti leírások, treebankek, ám az egyes elméleti kereteken belül is több különböző reprezentáció érhető el.

Az egyes keretek konkrét reprezentációi között kisebb és nagyobb különbségekkel találkozhatunk: a konstituens nyelvtani keretben készült Szeged Treebank 1.0 verziójának [4] reprezentációjában csak a főnévi csoportok és a tagmondat határok kerültek annotálásra, a 2.0 reprezentációban [1] már melléknévi, határozószói és más frázisok is jelölve vannak. A dependencia nyelvtan keretében eltéréseket láthatunk például a Szeged Dependencia Treebank [2] és a magyar univerzális dependencia treebank [5] között. Címkekészletüket tekintve egyes címkék az egyik reprezentációban elkülönülnek, míg a másikban nem; valamint egyes kategóriák esetén az elemek kötése is eltérő, például a koordináció esetén.

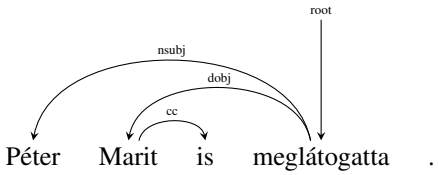
Cikkünkben különböző dependencia címkekészletekkel végzett kísérleteink eredményeit mutatjuk be. Először a standard címkézett (LAS) és címkézetlen (UAS) kiértékeléssel kapcsolatos problémákat mutatjuk be, majd a kísérleteinkhez felhasznált címkekészleteket. Végül közöljük az eredményeinket és egy NLP-s feladat kapcsán bemutatjuk azt is, hogy az eltérő címkekészletek használata szignifikánsan befolyásolja annak eredményességét.

2. Kiértékelés és címkekészletek

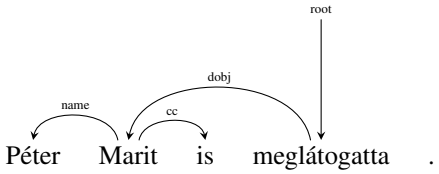
Dependencia szintaktikai elemzések közötti különbségeket általában az elemzők által elért UAS és LAS eredmények alapján állapítunk meg. Ezek a kiértékelési metrikák minden szót egyformán figyelembe vesznek: UAS eredmény esetén a megfelelő helyre kötött szavak, LAS esetén a megfelelő helyre, megfelelő címkével kötött szavak százalékos arányát viszonyíthatjuk egymáshoz.

Egy mondatban egy funkciószó téves kötése ugyanolyan hatással van az UAS és LAS eredményekre, mint egy tartalmas szóé, annak ellenére, hogy mind nyelvészeti szempontból, mind egy alkalmazás számára sokkal "nagyobb hiba" a tartalmas szó tévesztése. Az (1) ábrán a *Péter Marit is meglátogatta* mondat helyes dependencia nyelvtani szerkezetét láthatjuk az univerzális dependencia reprezentációjában. A (2) ábrán az alany hibásan a tárggyal együtt névelemként van elemezve, míg a (3) ábrán az *is* funkciószó hibája látható. Mivel a többi címke és kötés helyes, a (2) és a (3) mondatok UAS és LAS eredményei megegyeznek.

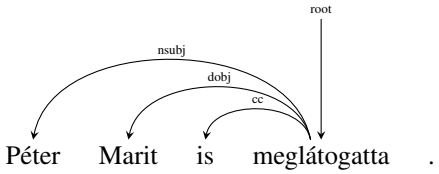
(1)



(2)



(3)



Álláspontunk szerint a szintaktikai elemzés önmagában nem végalkalmazás, hanem az előfeldolgozás része magasabb szintű alkalmazások számára, ezért nem egyforma fontosságú minden nyelvtani szerepű eleme a mondatnak. A standard UAS és LAS kiértékelések ezt nem mindig tükrözik megfelelően. Erre megoldást jelenthet a súlyozott kiértékelés, ahol a fontosabbnak ítélt címkék nagyobb súlyozással, a funkciószavak kisebb súlyozással járulnak az összesített eredményhez; a címkékre kivetített F-mérték, amelyben a számunkra fontosnak ítélt címkék által adott eredményt vehetjük figyelembe; vagy az adatbázisok átcímkézése, ahol a kevésbé releváns címkék összevonásával javítható lehet az elemző releváns címkéken elért teljesítménye. Kísérleteink kiindulópontja a magyar unvierzális dependencia treebank [5], amelyből több treebanket hoztunk létre a teszteléshez különböző címkék összevonásával. Így három típusú, öt darab új treebanket hoztunk létre: a határozószói címkék összevonásával, az alárendelő címkék összevonásával, valamint a funkciószavak címkéinek összevonásával.

Ezeket egymással és az eredeti treebankkel UAS és LAS eredményeken kívül a különböző címkékre mért F-mérték szempontjából hasonlítottunk össze egymással, valamint egy alkalmazásban felhasználva. A következőkben bemutatjuk az egyes új címkekészleteket.

2.1. Határozószók

A magyar univerzális dependencia treebank a Szeged Dependencia Treebank-ből [2] "örökölte" szemantikai információkat is tartalmazó határozói címkéit, amelyek megkülönböztetnek idő és helyhatározókat, és az irányhármasság szerint is különbséget tesznek. Így például a *tto* szintaktikai címke a "meddig" kérdésre válaszoló időhatározót jelöl, míg a *locy* címke "hol" kérdésre válaszoló helyhatározókat kapcsol a szerkezethez. Álláspontunk szerint, ezek között a címkék között dönteni már nem a szintaxis feladata, hanem szemantikai megkülönböztetés.

Két új címkekészletet hoztunk létre: az elsőben időhatározó (*advmod:time*) és helyhatározó (*advmod:loc*) kategóriákká vontuk össze az eredeti 6 címkét, a másodikban mind a hat címkét a már meglévő, általános határozói *advmod* címkével vontuk össze. Kutatási kérdésünk ebben a kísérletben, hogy ezeknek a szemantikai jellegű kategóriáknak az összevonásával nő-e a szintaktikai elemzés hatékonysága.

2.2. Alárendelés

Az univerzális dependencia projektben [6] bevezetett címkekészlet kilenc különböző címkét használ alárendelő mellékmondattípusok megkülönböztetésére. Második kísérletünkben arra voltunk kíváncsiak, hogy milyen hatással van az eredményekre a sokféle alárendelő mellékmondati címke.

Ebben az esetben egy új címkekészletet készítettünk, amelyben ezt a kilenc címkét vontuk egy kategóriába.

2.3. Funkciószavak

Legfőbb célunk a funkciószavak-tartalmas szavak megkülönböztetés vizsgálata volt. Álláspontunk szerint a szintaktikai elemzés legfontosabb célja a fő tartalmas szavak szintaktikai viszonyainak helyes felismerése, így a mondatok állítmányának, alanyának és tárgyának felismerése. Kíváncsiak voltunk, hogy a kisebb funkciócímkék összevonása hogyan változtatja meg a szintaktikai elemzők által elért eredményeket.

Ebben a kísérletben szintén két új címkekészlettel dolgoztunk: az első esetben a legtisztábban funkciócímkéket vontuk egy *funct* címke alá, a második esetben az összes funkciócímkét két új címke alá vontuk össze az erősen funkciócímke típusúakat, és a funkció- és tartalmascímkék között elhelyezhetőek elkülönítve.

Kutatási kérdésünk, hogy a szintaktikai elemzést felhasználó alkalmazások számára kevésbé fontos funkciócímkék összevonása megnöveli-e a szintaktikai elemzés hatékonyságát egészében, UAS és LAS eredményeket tekintve, valamint csak a fő, tartalmas címkék figyelembevételével.

Az 1. táblázatban az új címkekészletek láthatóak.

3. Eredmények

A kísérletekben a magyar univerzális dependencia treebank címkéit a fent említett módokon összevontuk, így az eredeti mellett öt teszt treebankkel kísérleteztünk: TIME-

EREDETI	FUNCT1	FUNCT2	SUB	MODE	TIME-PLACE
acl	acl	**funct2**	**cl**	acl	acl
advcl	advcl	**funct2**	**cl**	advcl	advcl
advmod	**funct1**	**funct1**	advmod	advmod	advmod
advmod:locy	**funct1**	**funct1**	advmod:locy	**advmod**	**advmod:loc**
advmod:mode	**funct1**	**funct1**	advmod:mode	advmod	advmod:mode
advmod:obl	**funct1**	**funct1**	advmod:obl	advmod:obl	advmod:obl
advmod:que	**funct1**	**funct1**	advmod:que	advmod:que	advmod:que
advmod:tfrom	**funct1**	**funct1**	advmod:tfrom	**advmod**	**advmod:time**
advmod:tlocy	**funct1**	**funct1**	advmod:tlocy	**advmod**	**advmod:time**
advmod:to	**funct1**	**funct1**	advmod:to	**advmod**	**advmod:loc**
advmod:tto	**funct1**	**funct1**	advmod:tto	**advmod**	**advmod:time**
amod:att	**funct1**	**funct1**	amod:att	amod:att	amod:att
amod:attlvc	**funct1**	**funct1**	amod:attlvc	amod:attlvc	amod:attlvc
amod:mode	**funct1**	**funct1**	amod:mode	amod:mode	amod:mode
amod:obl	**funct1**	**funct1**	amod:obl	amod:obl	amod:obl
appos	**funct1**	**funct1**	appos	appos	appos
aux	**funct1**	**funct2**	aux	aux	aux
case	**funct1**	**funct1**	case	case	case
cc	**funct1**	**funct1**	cc	cc	cc
ccomp	ccomp	**funct2**	**cl**	ccomp	ccomp
ccomp:dobj	ccomp:obj	**funct2**	**cl**	ccomp:dobj	ccomp:dobj
ccomp:obl	ccomp:obl	**funct2**	**cl**	ccomp:obl	ccomp:obl
ccomp:pred	ccomp:pred	**funct2**	**cl**	ccomp:pred	ccomp:pred
compound	**funct1**	**funct1**	compound	compound	compound
compound:preverb	**funct1**	**funct1**	compound:preverb	compound:preverb	compound:preverb
conj	**funct1**	**funct1**	conj	conj	conj
cop	**funct1**	**funct1**	cop	cop	cop
csubj	csubj	**funct2**	**cl**	csubj	csubj
det	**funct1**	**funct1**	det	det	det
dislocated	**funct1**	**funct1**	dislocated	dislocated	dislocated
dobj	dobj	dobj	dobj	dobj	dobj
dobj:lvc	dobj:lvc	dobj:lvc	dobj:lvc	dobj:lvc	dobj:lvc
goeswith	**funct1**	**funct1**	goeswith	goeswith	goeswith
iobj	iobj	iobj	iobj	iobj	iobj
list	**funct1**	**funct1**	list	list	list
mark	**funct1**	**funct1**	mark	mark	mark
name	**funct1**	**funct1**	name	name	name
neg	**funct1**	**funct2**	neg	neg	neg
nmod	nmod	nmod	nmod	nmod	nmod
nmod:att	nmod:att	nmod:att	nmod:att	nmod:att	nmod:att
nmod:obl	nmod:obl	nmod:obl	nmod:obl	nmod:obl	nmod:obl
nmod:obllvc	nmod:obllvc	nmod:obllvc	nmod:obllvc	nmod:obllvc	nmod:obllvc
nsubj	nsubj	nsubj	nsubj	nsubj	nsubj
nummod	**funct1**	**funct1**	nummod	nummod	nummod
parataxis	**funct1**	**funct2**	**cl**	parataxis	parataxis
punct	**funct1**	**funct1**	punct	punct	punct
remnant	**funct1**	**funct1**	remnant	remnant	remnant
root	root	root	root	root	root
xcomp	**funct1**	**funct2**	**cl**	xcomp	xcomp

1. táblázat. A létrehozott címkekészletek. Az EREDETI-től eltérőek félkövérrel kiemelve.

PLACE (idő- és helyhatározói címkék két címkére összevonása), MODE (idő- és helyhatározói címkék összevonása *mode* címkével), SUB (alárendelő mellékmondati címkék összevonása), FUNCT1 (egyértelmű funkciócímkék összevonása egy kategóriába), FUNCT2 (összes nem tartalmascímke összevonása két kategóriába). A treebankeken tízszeres keresztvalidációval a Bohnet parsert [7] tanítottuk etalon morfológiai címkék használata mellett, a kiértékeléshez UAS, LAS és F-mértéket használtunk.

3.1. UAS, LAS és F-mérték globálisan

Az egyes treebankeken elért LAS, UAS és F-mértékek a 2. táblázatban láthatóak.

	LAS	UAS	F-mérték
EREDETI	81,857	84,357	0,924967
TIME-PLACE	81,317	84,364	0,915494
MODE	81,866	84,055	0,935438
SUB	81,236	84,153	0,914999
FUNCT1	81,766	84,176	0,922665
FUNCT2	82,054	84,05	0,938319

2. táblázat. Különböző címkekészletű treebankeken elért eredmények.

Az EREDETI LAS-hoz képest szignifikáns különbséget csak a FUNCT esetekben és a SUB-nál értünk el (McNemar-teszt, $p < 0{,}05$), az idő- és helyhatározói változtatások által hozott különbségek nem szignifikánsak, ezt az magyarázhatja, hogy ezek a szemantikai címkék nincsenek nagy hatással a szintaxisra. Ám ezekből az eredmények ilyen módon való kiértékeléséből csak azt a (előre is nyilvánvaló) következtetést vonhatjuk le, hogy a címkék eltalálása (mikro F-mérték) jobb kevesebb címke esetén, míg a szavak megfelelő helyre kötése (UAS) legjobban az EREDETI, vagyis a legnagyobb címkekészlettel működik globálisan az összes címkét egyformán figyelembevéve. Fő célunk viszont a különböző címkekészletek fő, tartalmas relációkra való hatásának megvizsgálása volt.

3.2. F-mérték a fő címkékre

A 3. táblázatban az egyes címkekészletekkel elért F-mértékek láthatóak a fő, tartalmas címkékre: *root*, a mondat feje; *nsubj*, a tagmondat alanya; *dobj*, a tárgy; *iobj*, részeshatározó, és *nmod:obl*, egyéb esetű, kötelező főnévi bővítmény.

	root	nsubj	dobj	iobj	nmod:obl	TOTAL
EREDETI	0,867	0,873	0,950	**0,496**	0,923	0,888
TIME-PLACE	*0,858*	0,874	*0,948*	0,443	*0,920*	*0,885*
MODE	0,867	0,874	0,951	0,436	0,924	0,888
SUB	0,867	**0,878**	0,949	0,472	**0,929**	**0,890**
FUNCT1	0,863	0,873	**0,952**	0,44	0,923	0,889
FUNCT2	**0,872**	*0,872*	0,949	*0,409*	0,924	0,888

3. táblázat. Fő címkéken elért F-mérték különböző címkekészleteken. Oszloponként a legmagasabb eredmény félkövérrel, a legalacsonyabb dőlttel.

Az adatokból látható, hogy az EREDETI címkekészletnél az iobj címkén kívül minden esetben jobb eredményeket ér el valamelyik új változat. A részeshatározó a többi címkéhez képest nagyon ritka címke, ami magyarázza eltérő viselkedését. Összességében legalacsonyabb eredményeket a TIME-PLACE címkekészlettel értünk el, ami a legalacsonyabb F-mértéket éri el összességében és három fő címkénél is ez hozza a legkisebb értéket. Legjobbnak a SUB címkekészlet tűnik a fő címkéken történt kiértékelésnél: két címkén és összességében is a legmagasabb F-mértékeket éri el.

3.3. Összevont címkék eredményei

A harmadik elemzésünkben az összevont kategóriák által elért eredményt vizsgáltuk összevonás előtt és után, így például az EREDETI címkekészlet esetén az alárendelő mellékmondatok címkéinek összesített (mikro) F-mértékét a SUB címkekészletben az ezeket összevonó címke F-mértékével. A 4. táblázat az új címkekészletek összevont címkéinek és az EREDETI címkekészlet megfelelő címkéinek összesített F-mértékben mért eredményét mutatja. Az EREDETI és SUB, valamint az EREDETI és FUNCT2 összehasonlításokban szignifikánsan jobb az eredmény az összevont címkék esetén (McNemar-teszt, $p < 0{,}05$).

SUB		FUNCT1		FUNCT2	
EREDETI	SUB	EREDETI	FUNCT1	EREDETI	FUNCT2
0,625	0,814	0,941	0,974	0,944	0,973
				0,708	0,817

4. táblázat. Összevont címkék és megfelelő eredeti címkék F-mértékei.

A finom nyelvészeti megkülönböztetéseken alapuló címkék közötti választás nem egyszerű az elemző számára, így az összevont címkéken szignifikánsan jobb eredményt képes elérni. Álláspontunk szerint ezek a megkülönböztetések legtöbb esetben az alkalmazások szempontjából sem relevánsak, ezért összevonásuk nem okoz problémát, főként ha emellett a tartalmas címkéken elért eredmények is jobbak.

4. Az eltérő címkék hatása az enyhe kognitív zavar felismerésére

Az eltérő címkekészletek gyakorlati hatását megvizsgálandó, egy magasabb rendű nyelvtechnológiai feladatban is kísérleteket végeztünk. Munkacsoportunk korábban létrehozott egy gépi tanuló rendszert, mely a páciensek beszédátirataiból kinyert nyelvi jellemzők alapján osztályozza a kísérleti személyeket aszerint, hogy enyhe kognitív zavarban (EKZ) szenvednek-e vagy sem [8]. A rendszerben használt egyik fontos jellemző a tartalmas és funkciószavak aránya volt a páciens megnyilatkozásában.

Jelen kutatásunk eredményeinek tükrében meg tudtuk vizsgálni, hogy vajon a funkciószavak reprezentációja befolyásolja-e az EKZ felismerésének hatékonyságát. Ennek érdekében újratanítottuk a Bohnet parsert az eredeti reprezentációt tartalmazó treebanken, illetve a FUNCT2 reprezentációt tartalmazó treebanken, majd a kapott modelleket lefuttattuk a páciensek beszédátiratain. Az így kapott kétféle függőségi elemzésből nyertük ki aztán a tartalmas szavak, illetve a funkciószavak arányát, ugyanakkor mást nem változtattunk az eredetileg is használt jellemzőkön.

A kétféle reprezentáció alapján nyert jellemzőtért felhasználva végeztük el kísérleteinket, a Weka [9] szoftver döntési fa (C4.5) algoritmusával [10], követve [8] módszereit. Az eredmények szerint az eredeti reprezentációval 57,14%-os pontosságot, míg a FUNCT2 reprezentációval 69,05%-os pontosságot sikerült elérni, vagyis a módosított reprezentáció szignifikáns hatással bír az eredmények javulására (McNemar-teszt, p = 0,0245). Az EKZ automatikus felismerésében elért kísérleti eredményeink tehát alátámasztják, hogy a megfelelő szintaktikai reprezentáció megválasztása fontos szereppel bírhat a végalkalmazások eredményességére.

5. Összegzés

Cikkünkben különböző dependencia nyelvtani címkekészletekkel végzett kísérleteinket és azok eredményeit mutattuk be. Álláspontunk szerint, mind nyelvészeti, mind NLP-s alkalmazások szempontjából fontosabb a tartalmas címkék helyes felismerése egy szintaktikai elemzésnél, mint a funkciócímkéké. Eredményeink alapján, bizonyos címkecsoportok összevonása javíthatja számunkra fontosabb címkék helyes felismerését, sőt bemutattuk, hogy a reprezentáció módosításával egy végalkalmazás eredményét is szignifikánsan javíthatjuk. Az alkalmazásunk számára megfelelően kiválasztott szintaktikai reprezentáció erősen befolyásolja az alkalmazással elérhető eredményeket.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged TreeBank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005. Lecture Notes in Computer Science, Berlin / Heidelberg, Springer (2005) 123–132
2. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
3. Simkó, K.I., Vincze, V., Farkas, R.: Többszintű szintaktikai reprezentáció kialakítása a Szeged FC Treebankben. In Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia. (2014) 67–73
4. Csendes, D., Hatvani, C., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In Alexin, Z., Csendes, D., eds.: Magyar Számítógépes Nyelvészeti Konferencia. (2003) 238–245
5. Vincze, V., Farkas, R., Simkó, K.I., Szántó, Zs., Varga, V.: Univerzális dependencia és morfológia magyar nyelvre. In: XII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2015) 322–329
6. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
7. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97
8. Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., Kálmán, J.: Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, Association for Computational Linguistics (2016) 181–187
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
10. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)