Egyszer „van”, hol nem „van”: A létige kezelése függőségi nyelvtanokban

Simkó Katalin Ilona[1], Vincze Veronika[1,2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
kata.simko@gmail.com
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Cikkünkben három függőségi nyelvtani elemzést hasonlítunk össze a *van* létige kezelésének szempontjából. Az elméletek előnyeinek és hátrányainak tárgyalása után bemutatjuk, milyen eredményeket ér el egy szintaktikai elemző az egyes elemzésekben. Az ULA és LAS eredmények mellett részletes kézi hibaelemzést is végeztünk az adott szerkezet hibatípusaira koncentrálva. A cikk célja megtalálni a magyar *van* létige különböző típusainak számítógépes elemzésére leginkább alkalmas elméletet, valamint hangsúlyozni a feladatnak leginkább megfelelő elméleti keret megtalálásának és a kézi hibaelemzésnek a fontosságát.

Kulcsszavak: szintaxis, létige, kopula, hibaelemzés

1. Bevezetés

A nyelvi jelenségek nagy részére nem létezik egyetlen elfogadott nyelvészeti leírás, nemcsak az egyes keretek biztosítanak különböző megoldásokat a problémára, hanem az azokon belüli elméletek között is komoly eltérések lehetnek. Jelen cikk célja egy ilyen, a szakirodalomban már sokféleképpen leírt jelenség, a *van* ige lehetséges szintaktikai kezeléseinek megvizsgálása a számítógépes nyelvészetben, dependencia-nyelvtani keretben. A *van* ige kezelésére nem csak az elméleti nyelvészet ad számos lehetőséget, a számítógépes nyelvészet is több kísérletet tett erre [1]. Az ezek közötti választás nem egyértelmű, mindegyik megközelítésnek megvan az előnye és a hátránya is.

Cikkünk célja ezeknek az elemzéseknek a több szempontú összehasonlítása. A különböző elméletek szerint annotált korpuszon tanított elemzők eredményeinek szokásos, ULA és LAS százalékokban kifejezett összehasonlítása mellett kézi hibaelemzést végeztünk a relevánsnak tartott mondatrészek figyelembevételével.

2. A magyar létige függőségi nyelvtanokban

Cikkünkben a magyar *van* ige dependencia (függőségi) nyelvtanbeli lehetséges kezeléseit vizsgáljuk. Ennek oka, hogy ebben a keretben rendelkezésre áll három különböző elmélet szerinti teljes annotációval is ugyanaz a korpusz, a Szeged Korpusz Népszava alkorpusza [2]. A három elmélet így azonos feltételek mellett volt összehasonlítható.

2.1. Létige és kopula

A magyar *van* létigének – és a létigének számos egyéb nyelvben – egzisztenciális és kopuláris használata is létezik. Egzisztenciális használatban teljes értékű igeként viselkedik, létezést fejez ki ((1) példa). Kopuláris használat esetén az ige nem önállóan alkotja a mondat predikátumát, egy névszói rész is társul hozzá ((2) példa). A magyar nyelv – nem egyedi, de – érdekes és problémás tulajdonsága, hogy kopuláris használatban az igei paradigma bizonyos helyein (harmadik személy, jelenidő, kijelentő módban) a felszíni szerkezetben nem jelenik meg az ige ((3) példa).

(1) Sanyi a szobában van.
(2) Sanyi orvos volt.
(3) Sanyi orvos.

2.2. Funkció fej

A funkció fej elemzés a mondatban a funkciószavakat tekinti fejnek. Ilyen módon minden mondat feje a ragozott ige, legyen az akár teljes értékű ige, akár kopula. Ezt az elemzést támogatja Mel'čuk [3], aki a magyarhoz hasonló nyelvek esetén (ha a kopula csak bizonyos szám, személy, mód, idő esetén nem jelenik meg a felszíni szerkezetben) azt javasolja, hogy hiányzó ige esetén egy üres igealakot szúrjunk be a szerkezetbe.

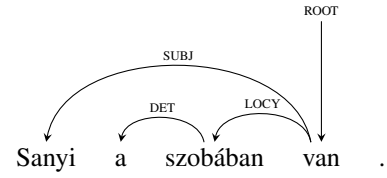
Ennek megfelelően lett létrehozva a Szeged Dependencia Treebank [4] szintaktikai annotációja. Jelen cikkben a treebank Népszava cikkekből álló részével dolgoztunk.

A treebankben minden mondat feje egy ige: ez lehet teljes értékű, jelentéses ige, jelentéssel nem bíró kopula vagy egy, a korpuszba kézzel beszúrt, üres 'VAN' fej a (3) példához hasonló mondatokban.

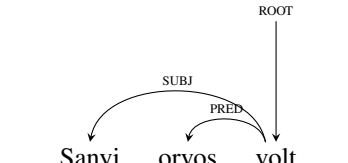
Az elemzés előnye az elemző számára, hogy az összes ige hasonlóképpen viselkedik a mondatokban és minden mondatban van ige. A módszer nagy hátránya viszont az előfeldolgozásban kézzel beszúrt VAN csomópont, ami életszerűtlenné teszi az automatikus elemzést.

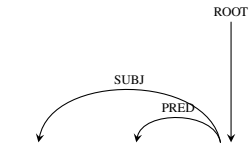
A (4) - (6) példák a (1) - (3) példamondatok elemzései ebben az elméletben.

(4)



(5)



(6)


2.3. Tartalmas fej

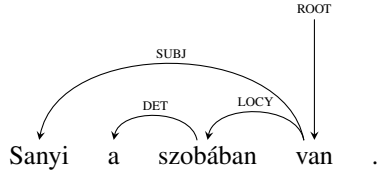
A tartalmas fej megközelítés alapja, hogy a mondatok fő elemeinek a jelentéses egységeket tekinti, a funkciószavak ezekhez kapcsolódnak. Ebben az elemzésben a *van* ige csak tartalmas igeként lesz fej, vagyis a (1) példához hasonló esetekben. A kopula *van* igét tartalmazó mondatok feje a névszói predikátum, az ige ehhez kapcsolódik.

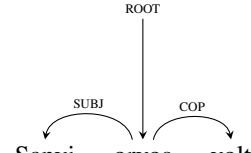
Az univerzális dependencia projektben [5] a különböző nyelvek könnyebb összehasonlíthatóságának érdekében tartalmas fej elemzést alkalmaznak. Mivel a funkciószavak rendszerei erősen különböznek az egyes nyelvekben, a funkció fej elemzésben sokkal nagyobb eltérések mutatkoznának egy-egy mondat különböző nyelvekre fordított változatának szintaktikai leírásában. Jelenlegi munkánkban a Szeged Dependencia Treebank Népszava alkorpuszának univerzális dependencia elveknek megfelelően átalakított változatát használtuk [6].

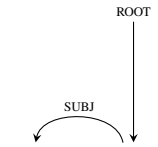
A treebankben a mondatok feje a ragozott, tartalmas ige (köztük a nem kopula *van*). Névszói predikátumot (is) tartalmazó mondatokban a fej mindig a névszó, a kopuláris létige (ha megjelenik) ehhez kapcsolódik, így nem okoz problémát a szerkezetben, ha az ige nem jelenik meg a felszínen.

Az elemzés előnye, hogy nincs szükség kézzel beszúrt igei csomópontokra, valamint elkülöníthető egymástól a *van* létige kopuláris és nem kopuláris használata. Hátránya, hogy kopuláris igéből jóval kevesebb tanítópélda található egy-egy korpuszban, mint jelentéses igéből, így nehézséget okozhat az elemzőnek megtanulni ezt a mintázatot, főleg olyan kis méretű korpusz esetén, mint amilyennel a jelenlegi kutatásban dolgoztunk.

A (1) - (3) példamondatok elemzései tartalmas fej elméletben a (7) - (9) példákban láthatóak. (Az egységes megjelenés miatt az univerzális dependenciában használt címkék helyett a cikkben a Szeged Dependencia Treebank címkéit használjuk.)

(7)


(8)


(9)


2.4. Komplex címke

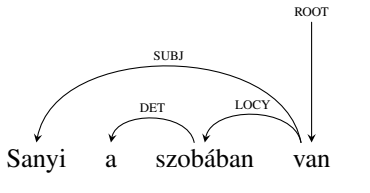
A harmadik megvizsgált elmélet a komplex címkés elemzés. Az elmélet létrehozásában a cél az üres csomópontok eltüntetése volt olyan módon, hogy az elemzésből látszódjon, honnan hiányzik az ige. A komplex címke elemzés alapjában véve egy funkció fej elemzés, attól csak a felszíni szerkezetben meg nem jelenő kopulát tartalmazó mondatok elemzésében tér el.

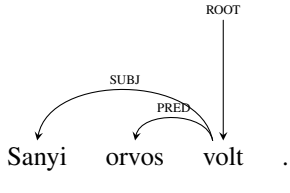
A meg nem jelenő, harmadik személy, jelenidő, kijelentő módú kopula helyett a névszói predikátum lesz a mondat feje, a hiányzó ige ennek a címkéjén és a belőle kiinduló címkéken van jelölve. A címkén szerepel ennek a hiányzó igének a címkéje, az ige és az a címke, amivel az adott szó az igéhez kapcsolódna.

A komplex címkés elemzéshez szintén a Szeged Dependencia Treebank Népszava részének egy átalakított változatát használtuk [7]. Ebből a treebankből automatikusan lettek eltávolítva a beszúrt VAN csomópontok, a hozzá kapcsolódó relációk pedig automatikusan átalakítva. A treebankben csak a meg nem jelenő kopulák elemzése tér el az eredeti, funkció fej változattól.

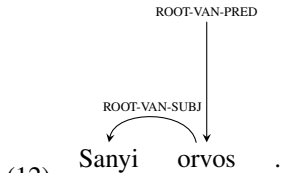
Ennek az elemzésnek szintén előnye, hogy nincs szükség VAN csomópontok beszúrására. Hátránya, hogy csak a megjelenő és a meg nem jelenő igéket különbözteti meg egymástól az elemzésben, valamint hogy potenciálisan végtelen sok címkére szükség lehet például meg nem jelenő kopulát tartalmazó tagmondatok koordinációja esetén.

Komplex címkéket tartalmazó elemzések a (10) - (12) példákban láthatóak a (1) - (3) példamondatokra, valamint a (13) példában két meg nem jelenő kopulát tartalmazó tagmondat koordinációja.

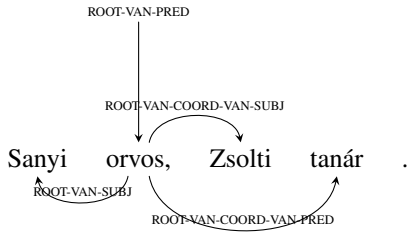
(10)


(11)


(12)



(13)



3. Hibaelemzés

A három, azonos szövegeken meglévő treebanken a Bohnet parsert [8] tanítottuk etalon morfológiai címkék használata mellett, majd a hagyományos, címkézetlen ULA és címkézett LAS értékeket számoltuk rajtuk. Ezek az eredmények a 1. táblázatban a teljes treebanken kiértékelve, illetve az *Egzisztenciális* sorokban csak a teljes értékű ige, jelentéses *van* igét tartalmazó mondatokra, a *„Megjelenő" kopula* sorokban a mondat felszíni szerkezetében megjelenő kopulát tartalmazó mondatokra, a *Virtuális kopula* sorokban a meg nem jelenő kopulát tartalmazó mondatokra számolt értékek vannak feltüntetve.

1. táblázat. ULA és LAS értékek a Funkció fej, Tartalmas fej és Komplex címkés elemzés teljesítményére.

	Funkció fej	Tartalmas fej	Komplex
Egzisztenciális - ULA	86,18	80,48	86,84
LAS	91,04	77,21	82,46
„Megjelenő" kopula - ULA	82,8	75,05	83,62
LAS	77,31	71,67	77,82
Virtuális kopula - ULA	84,42	78,39	77,5
LAS	79,17	75,15	69,59
Teljes anyag - ULA	85,75	84,41	84,76
LAS	81,24	81,2	79,89

Az eredmények alapján azt állapíthatjuk meg, hogy legjobban a funkció fej elemzés teljesít, azt szorosan követi a komplex címkés elemzés, legrosszabb pedig a tartalmas fej elemzés. Azonban ezek az eredmények nem tükrözik megfelelően a vizsgálni kívánt problémát: a teljes mondatokra számolt ULA és LAS számok nem fejezik ki megfelelően a *van* létigével kapcsolatos szintaktikai elemzési problémákat, mivel a mondat egyéb

részeiben ejtett elemzési hibák ugyanúgy befolyásolják az eredményt, mint a jelenleg vizsgálni kívánt, problémás szerkezet hibái.

Ebből kifolyólag kézi hibaelemzést végeztünk a *van*-nal kapcsolatos problémákra koncentrálva. A kézi hibaelemzés során 50-50 mondatot néztünk háromszor három kategóriában: Egzisztenciális, „Megjelenő" kopulás és Virtuális kopulás mondatokon mindhárom elemzésben. A hibaelemzés során négyféle hibakategóriát vettünk figyelembe: nem megfelelő az alany címkéje és/vagy kötése vagy egy egyéb szó kapott alanyi címkét; nem megfelelő a névszói predikátum címkéje és/vagy kötése vagy egyéb szó kapott predikátum címkét; az alany és a névszói predikátum alanyesetű NP-k egymás címkéit kapják meg az elemzésben; nem megfelelő szó lesz a *van* igét tartalmazó CP feje. Azokat a mondatokat, amelyekben ezek közül a hibák közül egyik sem jelent meg, helyesnek tekintettük a jelenlegi kutatás tekintetében, attól függetlenül, hogy volt-e egyéb hiba a mondat elemzésében. A helyes mondatok százalékos aránya a 2. táblázatban láthatóak, az alsó sorban a három kategóriára vonatkoztatott átlagos eredmény található.

2. táblázat. Helyes mondatok százaléka a kézi hibaelemzés alapján a Funkció fej, Tartalmas fej és Komplex címkés elemzésekben.

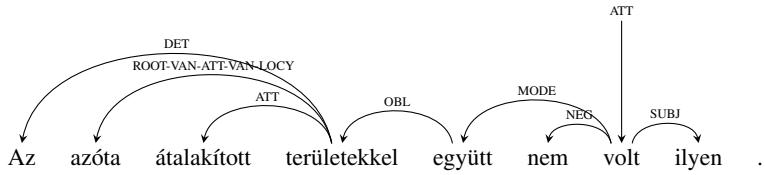
	Funkció fej	Tartalmas fej	Komplex
Egzisztenciális	78	80	80
„Megjelenő" kopula	62	42	52
Virtuális kopula	70	68	30
Összesen	70	63	54

Összességében megállapítható, hogy mindhárom elemző alanyi hibákat ejt legnagyobb számban. Ezek az alanyesetű főnévi csoportok a SUBJ helyett gyakran kapják meg más főnévi módosítók címkéit, illetve a névszói predikátum PRED címkéjét. Az alany és a névszói predikátum megkülönböztetése nehéz feladat: mindkét kifejezés alanyesetű főnévi csoport, és míg első vagy második személyű alany esetén az igével való egyeztetés miatt könnyebben elkülöníthetőek, harmadik személyű NP-nél (főként ha például mindkettő határozott névelős kifejezés) még az anyanyelvi beszélő ember számára sem egyértelmű a helyzet, mint a (14), (15) példák esetén.

(14) A kedvenc rajzfilmem a Mulán.
(15) A Mulán a kedvenc rajzfilmem.

A kézi hibaelemzésből kiderült, hogy a *van* ige elemzésével kapcsolatos hibatípusok a komplex címkés elemzés teljesít legrosszabbul: az Egzisztenciális mondattípus kivételével minden kategóriában ez az elemzés teljesít legrosszabban, valamint az alanyi, a predikátumi címkéket is ez elemzi legtöbbször hibásan. A (13) példában bemutatotthoz hasonló, összetett komplex címkéket (a belső logikájuk megértésének hiányában és az egyes ilyen típusú címkékre vonatkozó kevés tanítópélda miatt) nem tudja helyesen elemezni. Emellett a komplex címkés elemzésben előfordul, hogy a komplex címkék hibákat okoznak olyan helyeken, ahol nem kellene megjelenniük, mint a (1) ábrán, ahol a

TFROM (időhatározói) címke helyett egy komplex címke jelenik meg. Az is a komplex címkés elemzés ellen szól, hogy az elemző tanításának futási ideje több, mint duplája a másik két lehetőségnek, mivel míg a funkció fej 26, a tartalmas fej elemzés pedig 50 különböző lehetséges címkét tartalmaz, addig a komplex címkés elemzés a felhasznált treebankben 200-at, potenciálisan pedig végtelen sok címke tartozhat hozzá.



1. ábra. A komplex címkés szó a kézi annotációban TFROM címkével az *átalakított* szóhoz kötve.

A funkció fej és tartalmas fej elemzések a „Megjelenő” kopula kategóriában értek el a legalacsonyabb értékeket, valószínűleg azért, mert ebben a kategóriában a legnehezebb megkülönböztetni egymástól a teljes értékű ige *van*-t a kopulától. A két elemzés a különböző hibatípusokban is hasonló mennyiségű hibát produkált. Az eredmények értelmezésénél viszont érdemes figyelembe venni, hogy a funkció fej elemzés megfelelő működéséhez egy előfeldolgozó lépés is szükséges, amelyben a hiányzó VAN csomópontokat az egyes mondatokba illesztjük, míg a tartalmas fej elemzés nem igényel ilyet.

Jelen cikk kopulás szerkezetekre vonatkozó eredményeinek figyelembevételével a tartalmas fej alapú, univerzális dependencia nyelvtani elemzés tűnik a legmegfelelőbbnek a magyar számítógépes dependenciaelemzésre.

4. Összegzés

Cikkünkben megmutattuk, hogy a tartalmas fej típusú dependencia-nyelvtani elemzéssel előfeldolgozó lépés (üres VAN fejek beszúrása) és a címkék számának jelentős megnövelése nélkül érhetünk el versenyképes eredményeket a magyar *van* létigés szerkezetek elemzésében. Jelen cikk eredményei alapján a magyar nyelv dependenciaszintaxis elemzésére a tartalmas fej típusú, univerzális dependencia elemzést látjuk a legjobban alkalmazhatónak.

További terveink között szerepel jelen kísérlet megismétlése nagyobb tanítókorpuszokkal, hasonló kísérletek elvégzése más szerkezetek alapos megvizsgálására is, valamint a három elemzés tesztelése különböző alkalmazásokban való felhasználhatóság szempontjából: megvizsgálni, hogy milyen hatással van a három elemzés egyes feladatokra, amelyek erősen függnek a szintaktikai elemzéstől (például információkinyerés, véleménykinyerés, gépi fordítás).

Célunk volt emellett azt is megmutatni, hogy egyrészt a szintaktikai kereten kívül magának az alkalmazott, konkrét elméletnek is hatása van az elemzésre, és ez a számítógépes nyelvészetben számszerűen is kimutatható. Másrészt, bár nagyon hasznos és

viszonylag egyszerűen kinyerhető eredményeket ad a hagyományos ULA és LAS kiértékelése egy-egy elemzőnek főként nagyméretű treebankek esetén, a kézi hibaelemzés és nyelvészeti alapú vizsgálat sokkal informatívabb, mélyebb összefüggésekre világíthat rá.

Hivatkozások

1. Simkó, K.I.: Magyar kopulás szerkezetek az elméleti és a számítógépes szintaxisban. Master's thesis, Szegedi Tudományegyetem (2015)
2. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
3. Polguère, A., Mel'čuk, I.A., eds.: Dependency in Linguistic Description. Studies in language companion series. Amsterdam Philadelphia, Pa. J. Benjamins (2009)
4. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
5. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
6. Vincze, V., Farkas, R., Simkó, K.I., Szántó, Zs., Varga, V.: Univerzális dependencia és morfológia magyar nyelvre. In: XII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2016) 322–329
7. Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING 2012: Posters, Mumbai, India, The COLING 2012 Organizing Committee (2012) 1081–1090
8. 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