

Dömötör Andrea

**Hány VAN nincs?
A zéró létigék jellemzői és számítógépes elemzési
lehetőségei**

Doktori (PhD) értekezés

Tézisfüzet

Pázmány Péter Katolikus Egyetem
Bölcsészet- és Társadalomtudományi Kar
Nyelvtudományi Doktori Iskola
Nyelvtechnológia Műhely

Témavezető

Dr. Prószéky Gábor

egyetemi tanár, az MTA doktora

Budapest, 2025

1. Célkitűzés

A disszertáció a magyar létigés mondatok azon eseteit vizsgálja, ahol a morfológiailag jelöletlen esetben (azaz kijelentő mód, jelen idő, egyes szám 3. személyben) a létige nem, vagy nem kötelezően jelenik meg a mondatban. Ide tartozik az úgynevezett zéró kopula jelensége, illetve azok a létigét tartalmazó nem nominális mondatok, ahol a *van* igei elem elhagyható (pl. *Kulcs a lábtörlő alatt.*). Előbbi már jól ismert és leírt jelenség, utóbbiról azonban még viszonylag kevés említést találunk a szakirodalomban.

A disszertáció elsősorban számítógépes és adatvezérelt szemléletű. A fő elméleti kérdése az, hogy mely esetekben opcionális a létige a nem nominális mondatokban, és erre leginkább empirikus úton, a korpuszadatok elemzésével keres választ. A "klasszikus" zéró kopula kérdéskörében a számítógépes nyelvfeldolgozás problémáira helyezi a hangsúlyt, az ige nélküli mondatok ugyanis komoly kihívást jelentenek a gépi szövegelemzés (elsősorban a függőségi elemzés) számára.

A disszertációban összefoglalt kutatásoknak tehát kettős célja van. Az elméleti nyelvészet szempontjából a létigék típusainak és az elhagyhatóságuk szempontjainak jobb megértését tűzte ki célul. Ezen kívül a nyelvtechnológia igényeit is szem előtt tartva kitér a számítógépes elemzést segítő algoritmusok kidolgozására és a gépi tanuláshoz használható feladatspecifikus korpuszok létrehozására is.

2. Módszertan

A disszertáció mind módszereiben, mind céljaiban nagy hangsúlyt fektet a korpusznyelvészet szempontjaira. Az egyre nagyobb változatosságot mutató és egyre nagyobb méretű elérhető korpuszok megjelenése, valamint a számítógépes módszerek alkalmazása korábban elképzelhetetlen mennyiségű szöveges adat elemzését teszi lehetővé, ami segítheti a nyelvi jelenségek jobb megértését, vagy éppen új vagy eddig nem vizsgált jelenségek felfedezését. A korpuszok használata mára az elméleti nyelvészetben is általánossá vált (Teubert, 2005), a gépi tanulásra alapozott nyelvtechnológiai fejlesztésekben pedig egyenesen elengedhetetlen.

A korpusznyelvészetben a szöveges adatok felhasználásának két fő módszere van. Korpuszalapúnak (corpus-based) nevezzük azt a megközelítést, amikor valamilyen eleve meglévő hipotézist tesztelünk az adatokon, például gyakoriságok vagy disztribúciók vizsgálatával. A korpusvezérelt (corpus-driven) megközelítés esetén viszont csak minimális előzetes feltevéseink vannak, minden következtetést a korpuszatadatokból vonunk le (Tognini-Bonelli, 2001). A disszertációban ismertetett kutatásokra nagyrészt az előbbi megközelítés jellemző. Az általam végzett korpuszalapú mérések a zéró létigék előfordulásait, a létigés mondatok lexikális és disztribúciós jellemzőit, valamint a kopulás mondatok szerkezetét (szórendi változatait) vizsgálják, jellemzően meglévő elméletekre alapozva.

2.1. Felhasznált korpuszok

- **MNSz2** (Oravecz et al., 2014)

A Magyar Nemzeti Szövegtár v2.0.4. verziója 1,04 milliárd szövegszót tartalmaz 6 stílusrétegből és 5 regionális nyelvváltozatból. A szövegeket egy tokenizáló-szegmentáló, morfológiai elemző és POS-tagger modulokat tartalmazó pipeline dolgozta fel. A korpusz weben elérhető grafikus keresőfelülete a Sketch Engine keretrendszerbe tartozik. Ennek fő funkciói a szóalak, szótő és címke szerinti keresés, a gyakorisági listák készítése, valamint a kollokációkereső.

- **Szeged Treebank** (Vincze et al., 2010)

A Szeged Treebank az első magyar nyelvű, kézzel annotált, függőségi elemzést is tartalmazó korpusz. Összesen 82000 mondatból (1,2 millió szó) áll, melyek 6 különböző szövegtípusból származnak. A korpusznak a disszertáció szempontjából legfontosabb tulajdonsága, hogy tartalmaz üres igei fejeket, többek között a zéró kopulák jelölésére.

- **OPUS OpenSubtitles** (Lison és Tiedemann, 2016)

Az OPUS OpenSubtitles egy 62 nyelven elérhető, összesen 22,1 milliárd szövegszót tartalmazó, filmfeliratokból álló párhuzamos korpusz. Ennek magyar-angol részéből készült a 3. fejezetben felhasznált lemmatizált, morfológiailag elemzett és egyértelműsített, szó szinten megfeleltetett korpusz (Novák et al., 2019), amely 644,5 millió tokennyi angol szövegből és ezek magyar fordításából áll.

3. A disszertáció felépítése és főbb tézisei

A disszertáció elsődleges vizsgálati területe a létigék elhagyhatóságának szempontjai, melyet valós, korpuszokból mért nyelvi adatokra alapoz. Ezen kívül kitér a nominális mondatok néhány szerkezeti tulajdonságára, és bemutat egy automatikus zérókopula-beillesztő eszköz kidolgozására tett kísérletet.

Az elméleti bevezetőt követő 2. fejezet a disszertációban használt módszereket foglalja össze. Ide tartoznak a korpuszalapú és korpuszvezérelt megközelítések jellemzői, a gépi tanulás, a kiértékelési metrikák, valamint a számítógépes mondatelemzés, azon belül elsősorban a függőségi elemzés. Szintén ez a fejezet mutatja be a kutatásokhoz használt korpuszokat és nyelvfeldolgozó eszközöket.

A 3. fejezet egy olyan számítógépes nyelvfeldolgozó eszköz fejlesztését írja le, amely alkalmas a zéró kopulás mondatok automatikus felismerésére, sőt képes a zéró kopulát a mondatok megfelelő helyére beilleszteni. A rendszer alapja egy párhuzamos korpusz, ahol a zéró kopulás mondatokban a célnyelvi oldalon szerepel a kopula helyét jelző címke. Ezt tanítóanyagként felhasználva egyneurális gépi fordítót tanítottunk be arra, hogy a nominális mondatokba beillessze a zéró kopula címkét. Bár az eszköz pontossága egyelőre nem éri el azt a szintet, hogy érdemes legyen korpuszpépítésre használni, a fejlesztését célzó kutatás mégis számos tanulsággal szolgált.

A 4. fejezet azokat a "zéró létigéket" vizsgálja a korpuszokban, amelyek nem a nominális mondatok körébe tartoznak. A fejezet három jelenségekört elemez részletesen. A 4.2. fejezet az *Ott az ajtó*.-típusú mondatokra fókuszál, és arra keres választ, hogy miért támogatja a helyre utaló demonstratív elemek (*itt, ott, hol*) jelenléte a létige elhagyását. A 4.3. fejezetből az derül ki, hogy nem csak a *van* létige hagyható el bizonyos mondattípusokból, hanem annak tagadó változata, a *nincs/sincs* is. Ez az alfejezet a *Semmi gond*.-típusú mondatok lexikális sajátosságait vizsgálja. Végül, a 4.4. fejezet a *lehet* modális segédige elhagyhatóságának kérdését helyezi középpontba, azaz hogy mely igék esetén lehetséges az *Innen jól látni*.-típusú segédige nélküli szerkesztésmód.

Az 5. fejezet a kopulás mondatok néhány szerkezeti (szórendi) kérdésével foglalkozik. Az 5.1. fejezet, a zérókopula-beillesztő program fejlesztésének tanulságaira alapozva, olyan eseteket mutat be, ahol a zéró és a testes kopulás mondatok szerkezete eltér egymástól, és emiatt a kopula elhagyása hibás mondatot eredményez. Az 5.2. fejezet pedig a predikatív névszók szerkezeti

helyének kérdését járja körül: mely esetekben jellemző ezekre a fókuszálás, és mikor preferált a semleges (igemódosító) szórend.

Végül, a 6. fejezet a predikatív névszók (segéd)igei tulajdonságait, illetve a kopula "tartalmas jelentésének" gyengülését vizsgálja két szórendi kérdés (az igekötő-leválás és az -e kérdő partikula helye), illetve az igékre jellemző határozói és infinitívuszi vonzatok tükrében. Utóbbiaknál a fejezet kitér a segédigévé válás útjára lépett predikatív névszókra (*kell, muszáj, szabad*) is.

Összegezve, a disszertáció fő tézisei a következők:

1. Egy angol-magyar párhuzamos korpusz segítségével létrehoztam egy olyan magyar nyelvű szövegtörzset, amelyben speciális címkék jelzik a zéró kopulák helyét. Ezt tanítói anyagként felhasználva munkatársaimmal betanítottunk egy gépi fordítót arra, hogy tetszőleges mondatba beillesse a(z esetleges) zéró kopulát. Egy ilyen eszköz hasznos lehet a számítógépes mondatelemzés számára, a függőségi elemzési modellek ugyanis sok esetben (pl. a Szeged Treebank esetén is) alkalmaznak üres igei fejezetet a kopula nélküli nominális mondatok elemzésére. Az általunk fejlesztett eszköz a tanítói törzsszel megegyező forrásból származó szövegeken közel 90%-os pontosságot ért el. Más szövegtípusokon tesztelve azonban ennél jóval alacsonyabb eredményt kaptunk. Ennek az az oka, hogy a tanítói törzs szövegei, a speciális módszertan miatt, csak egyetlen műfajra (filmfeliratok) korlátozódnak. A zérókopula-beillesztő eszköz általánosan alkalmazhatóvá tételéhez szükség van a tanítói anyag más szövegtípusokkal való kibővítésére.
2. Korpuszmérésekkel kimutattam, hogy a zéró létige a magyarban nem csak a hagyományos értelemben vett névszói állítmányt érintő jelenség, hanem különböző feltételek mellett más típusú létigék is opcionálisan elhagyhatók jelen idő harmadik személyben. Ezek jelentős része állandósult kifejezés (pl. *rendben, vége, tele, stb.*) vagy műfajspecifikus jelenség (pl. irodalmi leírások). Néhány opcionálisan elhagyható létigét tartalmazó mondatípus azonban produktívnak mondható, ilyenek a helyre utaló demonstratív elemet (*itt/ott/hol*) vagy tagadó kifejezést (*semmi, sehol*) tartalmazó mondatok.

3. Az MNSz2-ből mért adataim alátámasztják azt a hipotézist, miszerint az igemódosító *itt/ott* gyengíti a módosított ige deskriptív tartalmát, így az valójában a létige (valamivel specifikusabb) szinonímájává válik. É. Kiss (2004) Ha a módosított ige maga a létige (*van*), akkor az emiatt szemantikailag redundánssá válik, így a funkciója inkább kopulaszerű inflexióhordozóhoz lesz hasonló. Ezért ha a mondat jelen idő, egyes szám 3. személyben áll (azaz nincs jelölt inflexió), a létige zéró hangalakú is lehet. Ez elméleti magyarázattal szolgál arra a jelenségre, miszerint opcionálisan a lokatív és egzisztenciális létigék is elhagyhatók a mondatokból, ha valamilyen, lehetőleg minél általánosabb jelentésű helyre utaló névmás (*itt/ott/hol*) előzi meg őket.
4. A tagadó formájú (*nincs, sincs*) létigék csak abban az esetben hagyhatók el a mondatból, ha az tartalmaz valamilyen egyéb tagadó értelmű kifejezést (*semmi, sehol, se*). Korpuszadatok alapján kimutattam, hogy bizonyos *semmi* + főnév kifejezések esetén különösen gyakori az ige nélküli szerkesztésmód, például: *semmi baj, semmi gond, semmi szükség (valamire), semmi ok (valamire), semmi kétség (valami felől), semmi köze (valamihez), semmi értelme (valaminek), semmi baja (valaminek/valakinek), semmi akadálya (valaminek)*.
5. Bizonyos speciális esetekben nem csak a létige, hanem annak hatóképzős alakjával azonos alakú segédige (*lehet*) is elhagyható a mondatokból. Korpuszgyakorisági vizsgálatok alapján megállapítottam, hogy a modalitást magába sűrítő, segédige nélküli infinitívuszos szerkesztésmód az *experiens*-alanyú, többargumentumú (elsősorban tranzitív) igékre jellemző leginkább, közülük is leggyakrabban az észlelést-érzékelést kifejező igékre.
6. Korpuszadatok alapján megmutattam, hogy a hagyományos "lexikális ige" – "inflexióhordozó funkcióelem" megkülönböztetés nem minden esetben magyarázza egyértelműen a létige elhagyhatóságát. Bizonyos esetekben ugyanaz a lexikális elem előfordul testes létigével és anélkül is. Ennek oka lehet a szófaji többértelműség (*kész, tele*), illetve a mennyiségre (*sok, kevés*) vagy állapotra (*rendben*) utalás. A "környezeti helyzetre" utaló névszók (*hideg, meleg, sötét*) esetén akkor szükséges a létige jelenléte, ha a mondatban nincs (explicit) alany, amire a jelzett tulajdonság vonatkozik. A *kész* esetén jelentésváltozás figyelhető meg a létigés és az ige nélküli szerkesztésmód között: előbbi általában napszakra, utóbbi pedig elkésett szándéokra vonatkozik.

7. A kopulás mondatok szerkezetével kapcsolatban kérdőíves és korpuszalapú módszerrel igazoltam, hogy az általában igemódosító pozícióban álló névszói predikátumok akkor fókuszálhatók, ha a jelentésük többértékűként határozható meg, azaz önmagukban is képesek alternatív jelentésű predikátumokat aktiválni. Jellemzően ilyenek a színnevek, számneves kifejezések, illetve a származást, közösséghez tartozást vagy foglalkozást jelölő névszói predikátumok. Ezzel szemben a bináris jelentésű melléknevek (amelyekhez nem tartoznak alternatív értékek az ellentétükön kívül) szinte soha nem kerülnek fókuszpozícióba. Ez megerősíti Roth (1985) és Krifka (2008) elméletét, miszerint a fókuszálás a releváns alternatívák jelenlétét jelöli.
8. A *szabad* igeragozott alakjainak korpuszvezérelt vizsgálatával demonstráltam a kopula igei inflexióként való grammatikalizálódásának (kezdeti) folyamatát. A *szabad* három különböző igeragozott formában található meg a sztenderd köznyelvben (*szabadna*, *szabadott*, *szabadjon*), melyek közül a feltételes mód esetén a korpuszadatok alapján már jóval preferáltabbnak bizonyult az igeragos alak a kopulásnál.

4. Releváns publikációk

Dömötör Andrea, Yang Zijian Győző és Novák Attila. Much Ado About Nothing – Identification of Zero Copulas in Hungarian Using an NMT Model. In: Calzolari N. et al. szerk. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille. European Language Resources Association, 2020, 4802–4810.

Dömötör Andrea, Yang Zijian Győző és Novák Attila. Nesze semmi, fogd meg jól! Zéró kopulák automatikus felismerése neurális gépi fordítással. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2020, 385–398.

Dömötör Andrea. Syntax is clearer on the other side - Using parallel corpus to extract monolingual data. In: Candito M. et al. szerk. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, Paris. Association for Computational Linguistics, 2019, 118–125.

Dömötör Andrea. Fókusz vagy igemódosító? In: Scheibl György szerk. *Lingdok 17.: Nyelvész-doktoranduszok dolgozatai*, Szeged. Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, 2018, 147–158.

Dömötör Andrea. Nem mind VP, ami állít - A névszói állítmány azonosítása számítógépes elemzőben. In: Ludányi Zsófia, Krepsz Valéria és Grácz Tekla Etelka szerk. *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2018: XII. Alkalmazott Nyelvészeti Doktorandusz-konferencia*, Budapest. MTA Nyelvtudományi Intézet, 2018, 3–10.

Dömötör Andrea. Hány VAN nincs? A létige – zéró váltakozás korpuszvezérelt vizsgálata. In: Ludányi Zsófia szerk. *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, Budapest. MTA Nyelvtudományi Intézet, 2017, 28–38.

Hivatkozások

É. Kiss Katalin. Egy igekötőelmélet vázlata. *Magyar nyelv*, 2004, 100(1):15–43.

Krifka Manfred. Basic notions of information structure. *Acta Linguistica Hungarica*, 2008, 55 (3–4):243–276.

Lison Pierre és Tiedemann Jörg. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Chair) Nicoletta Calzolari (Conference, Choukri Khalid, Declerck Thierry, Goggi Sara, Grobelnik Marko, Maegaard Bente, Mariani Joseph, Mazo Helene, Moreno Asuncion, Odijk Jan és Piperidis Stelios szerk. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA), may, 2016.

Novák Attila, Laki László János és Novák Borbála. Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból. In: XV. *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, Szeged. Szeged University, 2019, 63–71.

Oravecz Csaba, Váradi Tamás és Sass Bálint. The Hungarian Gigaword Corpus. In: Calzolari Nicoletta és et al. szerk. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Izland. European Language Resources Association (ELRA), 2014, 1719–1723.

Rooth Mats. *Association with focus*. Doktori disszertáció, University of Massachusetts, Amherst, 1985.

Teubert Wolfgang. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 2005, 10(1):1–13.

Tognini-Bonelli Elena. *Corpus Linguistics at Work*. John Benjamins, 2001.

Vincze Veronika, Szauter Dóra, Almási Attila, Móra György, Alexin Zoltán és Csirik János. Hungarian Dependency Treebank. In: *Proceedings of LREC 2010*, Valletta, Malta. ELRA, May, 2010.