GYULAI LÍVIA

SZTE BTK Nyelvtudományi Doktori Iskola
liviagyulai95@gmail.com
https://orcid.org/0000-0001-8855-4160

Gyulai Lívia: Az igék csoportosítása az igekötők argumentumszerkezetben okozott változása alapján
Clustering verbs based on the effect of preverbs on argument structure
Alkalmazott Nyelvtudomány, Különszám, 2023/2. szám, 157–173.
doi:http://dx.doi.org/10.18460/ANY.K.2023.2.009

Az igék csoportosítása az igekötők argumentumszerkezetben okozott változása alapján

Clustering verbs based on the effect of preverbs on argument structure

The research presented in this study examines the effect of Hungarian preverbs on argument structure, examining the preverb *el* 'away' in combination with 85 verbs using automatic syntactic methods based on corpus data. According to my hypothesis, verbs can be classified by cluster analysis contingent on the change in the verbal argument structure that the appearance of a preverb triggers. The verbs forming one class are those where the preverb causes a similar effect on certain argument types. To even out the different results of repeated K-means clusterings, a so-called "heatmap" was created. The heatmap shows the probability of two verbs getting in the same group. This method is applicable to examine the argument structure changing effects of any preverb with the help of automatic tools.

Keywords: corpus-based, preverbs, clustering, argument structure, Hungarian

1. Bevezetés

Kutatásom az igekötők argumentumszerkezetre való hatásainak vizsgálatára irányul. Jelen tanulmányban ennek egy részfeladatának bemutatását tűztem ki célul: lehet-e automatikus módon meghatározni, hogy mely igék alkotnak egy csoportot az igekötő argumentumszerkezetben okozott változása alapján. Az igekötők megjelenése a mondatban megváltoztathatja az ige argumentumszerkezetét az igekötő nélküli előfordulásokhoz képest. Az igekötőkre vonatkozó hazai szakirodalomban ennek a jelenségnek pusztán szintaktikai szempontból való vizsgálata hiánypótló.

Számos kutatás foglalkozott már az ige-bővítmény viszony vizsgálatával, többek között ilyen kutatás eredményeképpen jött létre a Mazsola (Sass, 2009) lekérdezőeszköz is. Segítségével megtudhatjuk, hogy milyen szavak jelennek meg leggyakrabban egy adott ige mellett bővítményi pozícióban, azonban jól hasznosítható az igekötő-kutatásban is, jelen dolgozatban is a Mazsola segítségével történt az adatgyűjtés.

Szécsényi (2019) az igék argumentumszerkezeti variánsainak automatikus módon való meghatározását tűzte ki célul korpuszalapú kutatásában. Szécsényi nem hagyományosan definiálja a bővítmény fogalmát, mely nem tesz különbséget vonzat és szabad bővítmény között, hanem egy skaláris megközelítést javasolt a

bővítmények „vonzatságának” méréséhez: az igék mellett megjelenő bővítményekhez egy 0 és 1 közötti értéket rendel a korpuszban való megjelenési gyakoriság alapján. Kutatásom során ezen az argumentumszerkezeti modellen értelmezem az igekötők argumentumszerkezet-változtató hatását. Kiinduló hipotézisem az, hogy az igekötő megjelenése nem befolyásolja az egyes argumentumok ige melletti előfordulási gyakoriságát.

Az igekötők számítógépes nyelvészeti vizsgálatához kapcsolódóan több kutatás is született, melyek közül most kettőt emelnék ki a következő rövid szakirodalmi áttekintésben:

Kalivoda (2017) az igekötők pontosabb gépi annotálásához kíván megoldást nyújtani, foglalkozik továbbá ige-igekötő párok azonosításával és azzal, hogy az igekötő milyen távol helyezkedik el az igétől.

Pethő et al. (2022) azzal a problémával foglalkozik, hogy a korpuszok annotálása során az igétől elváló igekötőt hagyományosan mindig külön tokenként szokták kezelni, tehát nincs jelölve, hogy mely igéhez tartozik az adott igekötő. Ezt orvosolja az emPreverb, amely egyrészt az igekötő igéhez való kapcsolására szolgáló módszer, másrészt javaslat arra, hogy hogyan lehet a korpusz annotációjában ezt megjeleníteni.

Az igekötők, valamint az igei vonzatszerkezetek korpuszalapú vizsgálatához kapcsolódóan megemlíteném továbbá Sass (2015) két nyilvánosságra hozott nyelvi erőforrását: „Az egyik a régi MNSZ tagmondatainak sekély szintaktikai elemzéssel ellátott változata, mely a Mazsola lekérdező adatbázisaként szolgál, a másik pedig az ebből az adatbázisból automatikusan származtatott igeiszerkezet-lista, melyből a Magyar Igei Szerkezetek című szótár is született” (Sass, 2015: 1). Ezen nyelvi erőforrások lehetővé teszik a kutatások új korpuszokon való kipróbálását.

Kutatásaim során eddig részben az intuícióra támaszkodva vizsgálódtam. A jelen kutatás újszerűségét az adja, hogy automatikus módszerekkel határozom meg, mely igék esetében okoz hasonló változást az igekötő megjelenése az argumentumszerkezetben. Az eddigi korpuszvizsgálataim során egy pilotkutatásban nem kompozicionális igekötős szerkezeteket vizsgáltam (Gyulai, 2019). A kutatás alapjaként kézi annotációval határoztam meg a vizsgált igék argumentumszerkezeti variánsait, majd ezt kiterjesztve az igekötők legjellemzőbb argumentumszerkezet-változtató hatásait tanulmányoztam (Gyulai, 2021), melyhez számítógépes módszereket alkalmaztam. Előbbihez az *el* igekötő hét igével való összekapcsolódásának korpuszadatait vizsgáltam. Ezt kibővítve másik kísérletem során a korpuszban megtalálható összes igekötő összes igével való összekapcsolódását igekötőnként összesítve tanulmányoztam. Ezen kutatás során az igekötők összesített hatását figyeltem meg, tehát nem az egyes igékkel való kapcsolatát. A közös tényező ezekben a kutatásokban a felhasznált korpusz volt, vagyis a Szeged Dependencia Treebank (Vincze et al., 2010). Felmerült azonban az igény, hogy más, nagyobb adathalmazon is végezzek vizsgálatot, mivel az eddig használt korpusz egyes igekötő–ige párokból kevés előfordulást tartalmaz,

így a jelen kutatás alapját a Mazsola (Sass, 2009) lekérdezőeszköz adta. Az említett korpuszok előnyeiről és hátrányairól a 2.1 fejezetben olvashatunk. Jelen vizsgálat az előbbiekben elvégzett kutatások elegye a következőkben bemutatott értelemben.

1.1. Célkitűzések, hipotézisek

A kutatás legfőbb célja egy olyan módszer kidolgozása, amely az igekötők argumentumszerkezet-változtató hatásainak szintaktikai szempontból való, automatikus módszerek segítségével elvégzett vizsgálatára irányul: a kutatás az *el* igekötő 100 igével való összekapcsolódását vizsgálja. Ahhoz, hogy kellőképpen nagy adathalmaz álljon rendelkezésre a kutatás elvégzéséhez, az a 100 ige került kiválasztásra, amely a Magyar nemzeti szövegtár 2. változata (Oravecz et al., 2014) szerint a leggyakrabban szerepel egy mondatban (egybe írva vagy az ige előtt és után maximum 2 tokennel) az *el* igekötővel. A kutatáshoz kapcsolódó hipotézis, hogy az igekötőkkel összekapcsolódó igéket egy klaszterező algoritmus segítségével osztályokba lehet sorolni az automatikusan elvégzett szintaktikai elemzésből kinyert igekötős és igekötőtlen igék argumentumszerkezeti vektorai között történő változás alapján ezáltal elkülöníthető az igekötők többféle argumentumszerkezet-változtató hatása.
A jelen tanulmányban bemutatott módszer egy általános metódus, amely bármely igekötő argumentumszerkezet-változtató hatásainak elkülönítésére alkalmas. Ebben az írásban az *el* igekötő példáján mutatom be, hogyan alkalmazható ez a módszer egy korpuszalapú szintaktikai igekötő-kutatáshoz. A klaszterezéssel kapcsolatos várakozásom, hogy az osztályozó olyan igéket fog egy csoportba sorolni, melyeknél hasonló jellegű változások történtek az igekötő megjelenésének hatására. Céljaim között szerepel továbbá az egy futtatás után kapott kimenet és a hőtérkép által kapott csoportosítás összevetése is.

2. Módszertan

2.1. Korpusz

Jelen kutatáshoz az eddigiektől eltérően nem a Szeged Dependencia Treebanket (Vincze et al., 2010) használtam, mivel célkitűzéseim között szerepelt az is, hogy megvizsgáljam, milyen eredmények érhetőek el egy ilyen típusú kutatásban, ha a szintaktikai elemzés nem kézzel, hanem automatikusan van elvégezve. A kutatáshoz a Magyar nemzeti szövegtár 2 (Oravecz et al., 2014, a toviábbiakban MNSz.[2]) korpusz és a Mazsola (Sass, 2009) lekérdezőeszköz szerepelt a lehetőségek között. Elsődleges felmerülő problémaként említeném, hogy a korpusz nem azonosítja, hogy milyen elváló igekötővel szerepel az ige mondatban, így az igekötőtlen igék is lehetnek más elváló igekötővel kapcsolt igék. Az MNSz.[2] (Oravecz et al., 2014) további hátránya abban mutatkozott meg, hogy a kiválasztott igék lekérése nagyon hosszadalmas folyamat lett volna, mivel minden igénél három lekérést kellett volna elvégezni: 1) azok a mondatok, melyekben az ige az *el* igekötő nélkül szerepel, 2) azok a mondatok, melyekben

az ige és az igekötő egybe írva szerepel, 3) azok a mondatok, amelyeknél az igekötő elválik az igétől, tehát azt legföljebb két tokennel megelőzi vagy követi. Az adatgyűjtéshez ezzel a módszerrel tehát 100 igénél 300 lekérést kellett volna elvégezni manuálisan. A lekérdezést az MNSz.[2]-ben lehetett volna egyszerűsíteni egy CQL-kereséssel, így azonban kiegyenlítetlen lett volna az igék közti előfordulási gyakoriság, ebben az esetben utólag kellett volna random kiválasztani igénként legfeljebb 1000 igét.

A másik opció a Mazsola (Sass, 2009) lekérdezőeszköz lett, amely a Magyar nemzeti szövegtár (Váradi, 2002) adatbázisát használja. A Mazsola (Sass, 2009) nagy előnyét jelentette, hogy ahhoz, hogy mind elváló, mind igével egybeírt előfordulásokat tartalmazó mondatokat kérjünk le, nem volt szükség az MNSz.[2]-ről szóló cikkben (Oravecz et al., 2014) leírt, hosszas adatlekérdezésre, mivel a keresőfelületen beírt igekötős ige lekérdezésekor az elváló igekötőt tartalmazó mondatokat is megkaptuk. Ez jelentős mértékben megkönnyítette az adatgyűjtést. A Mazsola hátránya volt, hogy a találatoknál nem minden esetben, sőt, sok esetben nem teljes mondatokat adott találatul. Problémát jelent ezen kívül, hogy a Mazsola (Sass, 2009) önkényesen szelektált a Magyar nemzeti szövegtár (Váradi, 2002) adatai között, például nem mutatta azokat az igéket, amelyeknek főnévi igeneves vonzata is volt. A potenciális adatforrások előnyeinek és hátrányainak mérlegelése után a Mazsola adatbázisa szolgáltatta az adatokat a kutatáshoz.

A Mazsola (Sass, 2009) által nyújtott adatokban fellelhető probléma a kiküszöbölésére a találatok közül csak azok maradtak benne az általam használt korpuszban, amelyek nagybetűvel kezdődtek és pont írásjellel zárultak. Mivel az *el* igekötővel 100 leggyakrabban előforduló igét választottam ki a vizsgálathoz, így egyes igéknél felmerült a probléma, hogy az igekötő nélküli adatoknál lényegesen kevesebb előfordulása volt vagy idiomatikussága miatt csak egyféle konstrukcióban szerepelt, így kézi szűréssel a vizsgálatban szereplő igék száma 85-re csökkent: *ad, alszik, árul, bír, bírál, bocsát, dől, dönt, enged, ér, esik, felejt, fér, fogad, foglal, fogy, hagy, hangzik, határoz, helyez, helyezkedik, hisz, hoz, indít, indul, intéz, ismer, ítél, jár, játszik, jön, jut, kap, kel, képzel, kér, kerül, készít, készül, kezd, költ, köszön, követ, küld, lát, látogat, lop, marad, megy, mesél, mond, múlik, nevez, nyer, olvas, pusztít, pusztul, rabol, rendel, ront, számol, szenved, talál, tart, telik, temet, tér, terjed, tesz, tölt, tűnik, utasít, utazik, üt, választ, válik, vállal, vár, végez, vesz, veszik, veszít, visel, visz, von*.

A kimaradt igék a következők: *akar, érik, fog, huny, juttat, kell, kezdődik, könyvel, különít, lesz, szeret, távolít, terül*, *tud*, *van*.

Az igekötő nélkül szereplő igei előfordulások száma az általam használt korpuszban 58 816 db, míg ugyanezen igék az *el* igekötővel 41 893-szor szerepeltek. Ezek az adatok a következő fejezetben leírt szűrési folyamat után kapott végleges adatmennyiségek.

2.2. Adatfeldolgozás

Az adatfeldolgozás első lépése az adatok megtisztítása volt. Amint fentebb említettem, a Mazsola keresőfelületén kapott találatok között sokszor szerepeltek nem teljes mondatok, illetve duplikátumok is. Azok a mondatok maradtak a korpuszban, amik nagybetűvel kezdődtek vagy pontra végződtek, valamint törlésre kerültek a duplikátumok. Felmerülhet a kérdés, hogy miért csak a pontra végződő teljes mondatokat hagytam benne az adatbázisban. A kijelentő mondatokat alapértelmezettként kezeltem, ekkor mutatkozik meg legjobban az ige argumentumszerkezete. A kérdő mondatok vizsgálatba való bevonása feltehetőleg nem okozna nagy változást az eredményekben, a felszólító mondatok esetében azonban sok a hiányzó argumentum, ami félrevezető eredményeket okozhatna. A megmaradt mondatokból igénként maximum 2×1000 db (igekötős + igekötő nélküli) került random kiválasztásra. Az igekötős igék között a legmagasabb előfordulás 954, a legalacsonyabb 60, átlagosan 483,4. Az igekötőtlen igéknél a legnagyobb előfordulás 995, a legkevesebb 47, átlagosan 691,9. A megtisztított, vagyis a vizsgálat során figyelembe vett adatok megtalálhatók a GitHubon.[1]

A Mazsolából (Sass, 2009) lekért mondatok megtisztítása után következett a feldolgozási fázis. Ehhez elsőként a magyarlanc (Zsibrita et al., 2013) szintaktikai elemzővel elemeztem a mondatokat.

Második lépésként az adatok feldolgozása, az argumentumszerkezetek meghatározása a Szécsényi és Virág (2022) által javasolt módon történt. Az összesített nyers adatok között még mindenféle módú ige szerepelt, a vizsgálatot azonban leszűrtem a kijelentő módú igéket tartalmazó mondatokra, mivel például a felszólító módú mondatok hiányosak lehetnek.

1. táblázat. Részlet az igekötő nélküli igék adatainak nyers táblázatából.

pvform	lemma	pvv	mood	n	PV	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL
0,00	ad	0+ad	ind	784,00	0,00	0,00	0,00	0,78	0,96	0,31	0,16	0,12	0,20	0,04
0,00	alszik	0+alszik	ind	647,00	0,00	0,00	0,00	0,43	0,09	0,00	0,18	0,12	0,01	0,05
0,00	bocsát	0+bocsát	ind	493,00	0,00	0,01	0,00	0,60	0,88	0,03	0,13	0,10	0,83	0,11
0,00	bír	0+bír	ind	845,00	0,00	0,02	0,01	0,46	0,35	0,00	0,08	0,03	0,04	0,30
0,00	bírál	0+bírál	ind	745,00	0,00	0,00	0,00	0,79	0,92	0,01	0,16	0,17	0,01	0,02
0,00	dönt	0+dönt	ind	874,00	0,00	0,00	0,00	0,81	0,03	0,01	0,19	0,29	0,04	0,05
0,00	dől	0+dől	ind	239,00	0,00	0,01	0,00	0,73	0,01	0,09	0,10	0,06	0,15	0,06
0,00	enged	0+enged	ind	761,00	0,00	0,01	0,01	0,49	0,39	0,13	0,09	0,06	0,07	0,02
0,00	esik	0+esik	ind	906,00	0,00	0,01	0,00	0,85	0,02	0,07	0,18	0,19	0,20	0,06
0,00	felejt	0+felejt	ind	211,00	0,00	0,00	0,00	0,36	0,42	0,00	0,12	0,09	0,01	0,02
0,00	fogad	0+fogad	ind	948,00	0,00	0,00	0,00	0,77	0,75	0,01	0,21	0,18	0,02	0,21
0,00	foglal	0+foglal	ind	719,00	0,00	0,00	0,00	0,71	0,92	0,01	0,25	0,13	0,01	0,06
0,00	fogy	0+fogy	ind	309,00	0,00	0,00	0,00	0,81	0,07	0,00	0,18	0,10	0,08	0,03
0,00	fér	0+fér	ind	219,00	0,00	0,00	0,00	0,66	0,00	0,00	0,07	0,02	0,06	0,02
0,00	hagy	0+hagy	ind	535,00	0,00	0,01	0,00	0,54	0,79	0,02	0,19	0,13	0,11	0,03

A táblázatban szerepel az igék szótöve, módja, előfordulási száma, adott esetben a hozzákapcsolt igekötő (ilyenkor a pvv oszlop pl.: *el+ad;* a pvform oszlop *el*), a további oszlopokban pedig a különböző bővítménytípusok

[1] https://github.com/gyulailivia/AlkNyelvDok22.

előfordulási gyakorisága látható, tehát az ige skaláris argumentumszerkezeti vektora (Szécsényi, 2019), amely az ige egyedi jellemzője. A táblázatban látható, hogy a tervezett 1000 előforduláshoz képest kisebb adatok vannak, ami az adatok megtisztításának eredménye. Ahogy már említettem korábban, előfordult, hogy az igekötő nélküli előfordulások száma kisebb egy-egy igénél, mint az igekötős esetben, mivel az igék úgy lettek kiválasztva az MNSz.[2]-ben (Oravecz et al., 2014) való előfordulás alapján, hogy az *el* igekötővel vannak összekapcsolódva, erre láthatunk példát a *felejt* igénél is, ami a korpuszomban igekötő nélkül csak 211-szer szerepelt.

Az igekötős igéket tartalmazó mondatokhoz is ilyen táblázat készült, így adva lehetőséget ugyanazon igék igekötő nélküli és igekötős előfordulásainak összehasonlítására, melynek segítségével az igekötő megjelenése által okozott változást, az igekötők argumentumszerkezet-változtató hatását lehet azonosítani.

A változás jellemzésére több módszer is felmerült, elsőként a gyakorisági vektorok különbsége. Ezzel a megoldással az látszódik, hogy mennyivel nőtt vagy csökkent egy adott bővítménytípus gyakorisága az adott igénél, azonban az arányok nem mutatkoznak meg. Például, ha egy bővítménytípus előfordulási gyakorisága 0,1-ről 0,2-re nő, az csak 0,1-es növekedést eredményez, viszont önmagához mérten ez egy nagy ugrás, mivel a duplájára nőtt, ezzel szemben, ha például egy bővítménytípus előfordulása 0,6-ról 0,7-re növekszik meg, az nem tekinthető arányaiban nagy növekedésnek. A másik lehetőség az igekötők változtató hatásának jellemzésére az egyes argumentumtípusok előfordulási gyakoriságainak hányadosa. Ebben az esetben ugyan látszódik, hogy mekkora aránybeli ugrás következett be, azonban a növekedés és csökkenés mértékét nem lehet vele mérni. Felmerült még továbbá a hányados logaritmusának használata. A logaritmus ugyan monoton átalakítást végez, mégis eltorzítja a különbségeket. Ezen szempontok figyelembevételével a különbség mellett döntöttem az igekötős és igekötő nélküli igék argumentumszerkezeti vektorai között fennálló változás mérésére. Ezeket az adatokat táblázatba (lásd 2. táblázat) rendeztem, amely celláinak színezése segít az értelmezésben is: minél sötétebb bordó egy cella, annál nagyobb mértékű növekedés figyelhető meg az adott bővítménytípusnál az *el* igekötő jelenlétében, és minél sötétebb kék egy cella, annál nagyobb mértékű a gyakoriság csökkenése, ha megjelenik az igekötő a mondatban.

3. Eredmények

3.1. Klaszterezés

A kutatás középpontjában álló módszerrel arra teszek kísérletet, hogy az eddigiek alapján előelemzett, feldolgozott adatokat a K-means klaszterezési módszerrel osztályozzam. A K-means klaszteranalízis többdimenziós vektorokat oszt K darab klaszterbe úgy, hogy minden elemet ahhoz a csoporthoz sorol, amelynek a középpontjához (mean) legközelebb van. A távolság méréséhez az euklideszi távolságot használja. A kutatás során arra használtam ezt a klaszterezési módszert,

hogy ily módon automatikusan tudjam csoportosítani azokat az igéket, amelyeknél az igekötő hasonló változást eredményez.

A K-means klaszterezés esetén megemlítendő, hogy minden futtatás alkalmával új csoportosítást ad eredményül, ha nem állítunk be egy ún. seedértéket, amely default beállításnál random lenne. Mivel ennek a feladatnak az elvégzéséhez a Python programnyelvet használtam, így volt lehetőség ennek a seedértéknek a beállítására, tehát a futtatások kimenetelei véletlenszerűek, de reprodukálhatók.

2. táblázat. Az igekötős és igekötő nélküli előfordulási gyakoriságok különbsége

lemma	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL
ad	-0,20	-0,08	0,02	0,02	-0,17	-0,01	-0,01	-0,02	-0,09	-0,03	0,01
alszik	-0,04	0,00	-0,06	-0,03	0,02	-0,02	-0,01	0,00	0,00	0,00	0,00
bocsát	-0,12	-0,03	0,03	0,00	-0,81	0,03	-0,05	-0,06	0,02	-0,02	0,21
bír	0,34	0,00	-0,06	-0,02	-0,04	-0,27	0,02	0,00	-0,01	0,00	0,00
bírál	0,03	0,01	0,07	0,03	0,02	0,04	-0,25	0,00	-0,01	0,00	0,01
dönt	0,60	-0,01	-0,07	-0,19	-0,01	0,02	-0,03	-0,01	-0,59	0,00	0,00
dől	0,00	-0,08	0,16	0,21	-0,09	-0,02	0,03	-0,24	-0,01	-0,05	-0,08
enged	0,17	-0,12	-0,03	-0,02	0,00	0,01	-0,08	-0,08	0,00	-0,02	-0,03
esik	-0,01	-0,06	0,08	-0,07	-0,19	-0,03	-0,08	-0,10	-0,25	-0,01	0,00
felejt	-0,01	0,01	-0,07	-0,07	0,02	-0,01	-0,05	0,00	0,00	0,00	0,00
fogad	0,16	0,01	-0,09	0,01	0,03	-0,13	-0,05	-0,01	0,02	0,00	0,00
foglal	0,01	0,00	0,05	0,04	0,01	0,00	-0,03	-0,08	-0,03	0,00	0,01
fogy	-0,06	0,02	-0,05	-0,04	-0,03	-0,02	-0,08	0,00	-0,02	0,00	-0,03
fér	0,00	0,00	0,29	0,24	-0,05	0,01	0,09	-0,27	0,00	-0,28	-0,01
hagy	0,00	-0,01	-0,10	-0,03	-0,09	0,03	-0,06	-0,01	0,00	0,00	0,01
hangzik	-0,01	-0,09	0,24	0,30	0,00	0,03	-0,41	0,00	0,09	0,00	0,00

Az 1. és 2. táblázat teljes adatmennyisége megtalálható a GitHubon.

Az első lépés a klaszterezéshez az igekötő nélküli és igekötős igék argumentumszerkezeti vektor értékeinek összevetése volt, amit a 2. táblázat segítségével tehetünk meg, melyben az igekötős és igekötő nélküli előfordulások különbsége látható.

Ezután a lépés után történt a klaszterezés. A program bemenetét ez a táblázat adta. A K-means klaszterezés esetén meg kell adnunk a programnak, hogy hány csoportba ossza az adatainkat. Ebben a kísérletben 2-től 20-ig terjedő csoportok kerültek meghatározásra egyes lépésküszöbbel. Erre azért volt szükség, mivel nem tudhatjuk, hogy hányféle argumentumszerkezet-változtató hatása van az igekötőnek, ezért több lehetőséget is ki kellett próbálni. A program kimenete egy olyan klaszterezés, melynél egy táblázatban vannak összesítve az összes klaszterezés (2 csoport, 3 csoport, 4 csoport stb.), ezt oszlopokban csatolja hozzá a bemeneti fájlhoz. Az adatok Excel fájlba való átültetése után kézi szűréssel elemezhetjük a klaszterező által javasolt csoportokat. A 3. táblázat a klaszterező által kapott kimenetet tartalmazza. A klaszterező megtalálható a GitHubon.[2]

A klaszterezőtől azt várjuk el, hogy a csoportokat úgy alakítsa ki, hogy azok az igék kerüljenek egybe, melyeknél megközelítőleg ugyanolyan mértékű változást okozott egy-egy bővítménytípusnál az igekötő megjelenése. A klaszterezés során

[2] https://github.com/gyulailivia/AlkNyelvDok22

minél több csoportot szeretnénk kialakítani, annál tovább bontja az algoritmus a már meglévő, nagyobb csoportokat, tehát tovább finomítja a klasztereket.

A 3. táblázatban a csoportosítás egy részletét láthatjuk. Ebben a formában az adatok még nincsenek rendezve, tehát nem azok az igék vannak egymás mellett, amelyek egy csoportba tartoznak. A következőkben a 7 klaszterbe való sorolás eredményeit ismertetem.

3. táblázat. A klaszterező által kapott kimenet részlete Excel táblázatban

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	2	3	4	5
ad	-0,20	-0,08	0,02	0,02	-0,17	-0,01	-0,01	-0,02	-0,09	-0,03	0,01	0,00	0	2	0	4
alszik	-0,04	0,00	-0,06	-0,03	0,02	-0,02	-0,01	0,00	0,00	0,00	0,00	0,01	0	2	0	4
bocsát	-0,12	-0,03	0,03	0,00	-0,81	0,03	-0,05	-0,06	0,02	-0,02	0,21	0,09	0	1	2	0
bír	0,34	0,00	-0,06	-0,02	-0,04	-0,27	0,02	0,00	-0,01	0,00	0,00	0,01	1	0	3	2
bírál	0,03	0,01	0,07	0,03	0,02	0,04	-0,25	0,00	-0,01	0,00	0,01	0,01	0	0	3	4
dönt	0,60	-0,01	-0,07	-0,19	-0,01	0,02	-0,03	-0,01	-0,59	0,00	0,00	0,00	1	0	1	2
dől	0,00	-0,08	0,16	0,21	-0,09	-0,02	0,03	-0,24	-0,01	-0,05	-0,08	-0,06	0	0	3	1
enged	0,17	-0,12	-0,03	-0,02	0,00	0,01	-0,08	-0,08	0,00	-0,02	-0,03	0,00	1	0	3	4
esik	-0,01	-0,06	0,08	-0,07	-0,19	-0,03	-0,08	-0,10	-0,25	-0,01	0,00	0,25	0	1	2	4
felejt	-0,01	0,01	-0,07	-0,07	0,02	-0,01	-0,05	0,00	0,00	0,00	0,00	-0,01	0	2	0	4
fogad	0,16	0,01	-0,09	0,01	0,03	-0,13	-0,05	-0,01	0,02	0,00	0,00	-0,01	1	0	3	4
foglal	0,01	0,00	0,05	0,04	0,01	0,00	-0,03	-0,08	-0,03	0,00	0,01	0,01	0	0	3	4
fogy	-0,06	0,02	-0,05	-0,04	-0,03	-0,02	-0,08	0,00	-0,02	0,00	-0,03	0,00	0	2	0	4
fér	0,00	0,00	0,29	0,24	-0,05	0,01	0,09	-0,27	0,00	-0,28	-0,01	-0,01	0	0	3	1

A klaszterezéssel kapcsolatban azt várom, hogy lesznek olyan csoportok, amelyeknél egy adott bővítménytípus esetén a csoport összes igéjénél hasonló mértékű növekedés/csökkenés figyelhető meg. A növekedést a piros cellák, míg a csökkenést a kékek jelentik, tehát egy ideális csoport esetén kék és/vagy piros oszlopokat látunk.

Az 4. táblázatban a 0 jelű csoport elemeit láthatjuk.

4. táblázat. 7-es csoportosítás, 0 jelű csoport

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
esik	-0,01	-0,06	0,08	-0,07	-0,19	-0,03	-0,08	-0,10	-0,25	-0,01	0,00	0,25	0,00	0,00	-0,02	0
marad	-0,01	-0,04	-0,02	-0,06	-0,06	0,04	0,01	0,00	0,01	0,00	-0,03	0,24	0,00	0,00	-0,04	0
vár	-0,05	-0,01	-0,07	-0,11	-0,24	-0,03	0,01	-0,04	-0,01	0,00	-0,01	0,23	-0,01	0,00	-0,03	0
üt	0,00	-0,02	0,06	0,05	-0,17	-0,01	0,04	-0,09	0,00	-0,01	-0,01	0,25	0,02	-0,02	0,00	0
választ	-0,10	-0,09	-0,08	-0,09	-0,07	0,01	-0,07	-0,02	0,01	-0,01	-0,03	0,50	-0,01	-0,19	0,00	0
von	0,08	0,00	-0,02	-0,07	-0,02	0,02	-0,05	-0,21	0,09	0,00	0,14	0,46	0,00	0,00	0,00	0
tér	0,00	0,01	0,18	0,01	-0,41	0,01	0,17	-0,03	0,01	-0,03	-0,02	0,56	0,01	0,00	-0,02	0

Az első csoportot az *esik, marad, vár, üt, választ, von, tér* igék alkotják. A 0 jelű csoportnál egyértelmű növekedést figyelhetünk meg a *-tÓl* ragos bővítménytípus esetében. Emellett csökkenés történik a *-rA* ragos bővítmények előfordulási gyakorisága esetén. Korábbi kutatásom során (Gyulai, 2021) már megfigyelhető volt, hogy amikor egy bővítménytípus előfordulási gyakorisága megnövekszik, mellette egy másik esetében csökkenés következik be. A 4. táblázatban látható igék az intuíciónak megfelelően is egy csoportot alkotnak, mivel megfigyelve őket azt a megállapítást tehetjük, hogy mindegyik esetében egy „absztrakt” *-tÓl* ragos bővítmény előfordulási gyakorisága

növekszik meg, amely az igék esetében (szemantikai értelemben) valamiféle változást jelentenek.

Az 1 jelű csoporthoz tartozó adatokat az 5. és 6. táblázat mutatja.

5. táblázat. 7-es csoportosítás, 1 jelű csoport részlete.

lemma	acc	dat	BA(	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ(	NÁ	VÁ	IG	7
jár	-0,03	-0,05	0,13	-0,03	0,00	-0,16	0,10	-0,05	0,00	0,00	0,00	-0,02	-0,02	0,00	-0,02	1
temet	-0,04	0,00	0,19	0,02	0,01	0,02	0,06	-0,18	0,00	0,00	-0,02	0,00	0,00	0,00	0,00	1
bírál	0,03	0,01	0,07	0,03	0,02	0,04	-0,25	0,00	-0,01	0,00	0,01	0,01	0,00	0,00	0,04	1
foglal	0,01	0,00	0,05	0,04	0,01	0,00	-0,03	-0,08	-0,03	0,00	0,01	0,01	-0,01	0,00	0,00	1
ismer	0,01	0,04	0,07	0,02	0,00	0,13	0,00	0,00	-0,01	0,00	-0,05	-0,01	0,00	0,00	0,00	1
készít	0,04	-0,02	0,02	0,02	0,02	-0,04	0,00	0,00	-0,13	-0,01	-0,09	0,00	0,00	0,00	0,01	1
dől	0,00	-0,08	0,16	0,21	-0,09	-0,02	0,03	-0,24	-0,01	-0,05	-0,08	-0,06	0,02	0,00	0,02	1
fér	0,00	0,00	0,29	0,24	-0,05	0,01	0,09	-0,27	0,00	-0,28	-0,01	-0,01	0,01	0,00	0,01	1
hangzik	-0,01	-0,09	0,24	0,30	0,00	0,03	-0,41	0,00	0,09	0,00	0,00	0,01	-0,01	0,00	0,00	1
helyez	0,07	-0,01	0,26	0,18	-0,27	-0,01	0,01	-0,42	0,00	0,00	0,01	0,00	0,10	0,00	0,00	1
pusztul	0,00	0,01	0,16	0,09	-0,01	-0,05	-0,07	-0,02	0,00	0,00	0,02	0,03	0,01	0,00	0,01	1
terjed	0,00	0,01	0,31	0,05	0,00	-0,05	-0,06	0,00	-0,02	0,00	0,02	-0,30	0,01	0,00	-0,41	1

Az 1 jelű csoport első felébe a *jár, temet, bírál, foglal, ismer, készít, dől, fér, hangzik, helyez, pusztul, terjed* igék tartoznak. Ebben a csoportban is elsősorban növekedést lehet megfigyelni, méghozzá a *-bAn* ragos bővítménytípusnál és helyenként az *-On* ragos bővítményeknél is. A *dől-helyez* igék esetében az *-On* bővítménytípus megnövekedése valamiféle hely- vagy időhatározókra utalhatnak. Elvétve látunk csökkenést is, de egységes tendenciát nem mutatnak a kék cellák. A csoport másik fele más változást mutat, melyet a 6. táblázatban láthatunk.

Az 1 jelű csoport másik felébe a *követ, lop, rabol, rendel, ér, bír, kezd, képzel, mesél, nyer, szenved, számol, veszít, végez, enged, fogad* igék tartoznak. Egyértelmű növekedés jelent meg a tárgyi bővítmény esetében ezen igéknél. A klaszterező ugyan egy csoportba sorolta ezeket az igéket, de láthatóan két külön hatást lehet azonosítani ezen az egy csoporton belül. Ezen igék esetében megfigyelhető, hogy igekötő nélkül valamiféle habituális értelemben használjuk őket, ilyenkor nemigen kapcsolódik hozzájuk bővítmény, azonban az *el* igekötő megjelenésével bekerül a vonzatszerkezetbe a cselekvés tárgya is (pl. *Gábor lop* vs. *Gábor **el**lopta Karcsi bácsi gyertyatartóját*)

A 7. és 8. táblázat a 2 jelű csoport adatait tartalmazza.

A 2 jelű csoport is nagy elemszámmal rendelkezik az 1 jelű csoporthoz hasonlóan, valamint ebben a csoportban is kétfelé tudjuk osztani az igéket. A klaszter egyik felénél a tárgyi, a *-bA* és az *-On* ragos bővítménytípusoknál figyelhető meg csökkenés. Ezzel szemben a klaszter többi eleménél nem látható szisztematikus változás egyik bővítménytípus esetében sem. Ezt a 8. táblázat mutatja.

A csoport ezen felénél ugyan sok esetben láthatunk csökkenést az egyes bővítménytípusok esetében, mégis egyértelmű tendenciát nem lehet meghatározni. A *-bA, -On* és *-rA* ragos bővítménytípusok esetén majdnem minden esetben csökkenést láthatunk, azonban ezek leginkább kismértékű változások.

Mivel az igekötő nemigen okozott változást ezeknél az igéknél, így itt az igekötő perfektiváló hatását érhetjük tetten.

6. táblázat. 7-es csoportosítás, 1 jelű csoport részlete.

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
követ	0,21	0,00	0,09	-0,01	0,03	-0,03	-0,03	-0,01	0,00	0,00	0,01	-0,01	0,01	0,00	-0,02	1
lop	0,27	0,01	0,10	0,13	0,01	0,00	0,01	-0,06	0,05	0,00	0,14	-0,01	0,00	0,00	0,01	1
rabol	0,16	0,00	0,08	0,09	0,01	-0,01	-0,08	0,00	0,04	-0,02	-0,01	0,04	0,01	0,00	0,00	1
rendel	0,26	-0,03	0,14	0,15	-0,02	0,00	0,01	-0,05	0,00	-0,07	-0,03	-0,03	0,02	0,00	0,03	1
ér	0,20	-0,01	0,09	0,01	-0,01	-0,05	-0,02	-0,05	-0,01	-0,01	0,01	0,00	0,01	0,00	0,00	1
bír	0,34	0,00	-0,06	-0,02	-0,04	-0,27	0,02	0,00	-0,01	0,00	0,00	0,01	0,00	0,00	-0,03	1
kezd	0,19	0,00	-0,02	-0,04	0,02	-0,18	-0,01	-0,20	-0,03	0,00	0,00	0,00	-0,01	0,00	0,00	1
képzel	0,42	-0,11	0,10	0,05	0,00	0,06	0,02	-0,03	-0,01	0,00	0,00	0,00	0,01	0,00	0,00	1
mesél	0,30	0,08	0,04	0,01	0,01	0,02	0,02	0,00	-0,23	0,01	0,01	0,01	0,00	0,00	0,02	1
nyer	0,26	0,00	-0,01	-0,05	-0,09	-0,02	-0,04	-0,02	0,00	0,01	-0,01	0,00	-0,01	0,00	0,00	1
szenved	0,24	0,00	-0,05	0,01	0,01	0,02	0,00	0,00	0,00	0,00	0,01	-0,08	0,01	0,00	-0,01	1
számol	0,31	0,07	-0,04	0,03	-0,02	-0,44	0,03	0,00	0,00	-0,01	-0,02	0,00	-0,01	0,00	0,06	1
veszít	0,26	0,00	-0,06	-0,06	0,01	0,00	0,00	0,00	0,00	0,00	-0,37	0,01	-0,01	0,00	0,00	1
végez	0,30	0,00	-0,04	-0,26	0,01	-0,07	0,03	0,00	-0,01	0,00	0,02	-0,01	0,00	0,00	0,01	1
enged	0,17	-0,12	-0,03	-0,02	0,00	0,01	-0,08	-0,08	0,00	-0,02	-0,03	0,00	0,00	0,00	0,00	1
fogad	0,16	0,01	-0,09	0,01	0,03	-0,13	-0,05	-0,01	0,02	0,00	0,00	-0,01	0,00	-0,01	0,00	1

7. táblázat. 7-es csoportosítás, 2 jelű csoport részlete.

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
intéz	-0,50	-0,01	-0,21	-0,11	-0,01	0,12	0,01	0,00	-0,01	-0,34	-0,01	-0,01	0,00	0,00	-0,01	2
kap	-0,40	-0,01	-0,11	-0,05	-0,14	0,02	-0,02	-0,02	-0,03	-0,03	-0,04	-0,11	0,00	0,00	-0,01	2
látogat	-0,35	0,00	0,01	-0,03	0,15	0,00	-0,04	0,12	0,01	0,05	0,00	-0,01	0,00	0,00	-0,01	2
tesz	-0,20	-0,05	-0,18	-0,07	0,01	-0,01	0,02	0,04	-0,02	0,00	0,03	-0,02	-0,02	-0,30	-0,01	2
tölt	-0,24	-0,01	-0,20	-0,05	0,01	0,37	0,01	-0,05	0,00	0,00	-0,03	0,00	-0,02	0,00	0,01	2
visel	-0,29	-0,01	-0,10	-0,08	0,00	-0,01	0,09	0,00	0,00	0,00	-0,01	-0,02	0,00	0,00	-0,02	2
árul	-0,25	0,01	-0,19	-0,20	0,00	0,01	-0,04	0,00	0,20	0,00	0,00	-0,01	0,00	0,00	-0,01	2
hoz	-0,15	-0,01	-0,12	-0,09	-0,09	-0,02	-0,02	-0,12	0,01	0,00	0,05	0,01	-0,01	0,00	-0,01	2
múlik	-0,01	0,00	-0,02	-0,61	0,04	0,02	0,07	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	2
ront	-0,15	-0,01	-0,02	-0,18	-0,03	-0,02	-0,05	-0,01	-0,01	0,00	-0,01	0,00	0,01	0,00	0,00	2
talál	-0,11	-0,14	-0,28	-0,13	-0,11	0,03	-0,02	-0,02	0,01	0,00	0,01	0,00	-0,04	0,00	0,00	2
vesz	-0,18	0,00	-0,23	-0,34	-0,03	-0,03	-0,05	-0,28	0,00	-0,01	0,01	0,15	-0,01	0,00	-0,01	2
vállal	-0,12	0,02	-0,14	-0,04	0,01	-0,01	-0,04	0,00	0,00	0,00	-0,02	-0,01	0,00	0,00	0,00	2
ad	-0,20	-0,08	0,02	0,02	-0,17	-0,01	-0,01	-0,02	-0,09	-0,03	0,01	0,00	-0,01	0,00	0,00	2

A 9. táblázatban a 3 jelű csoport elemzését láthatjuk.

A 3 jelű csoport egy nagyon kirívó klaszter, mivel összesen egy elemet számlál. Ez a klaszter nem csupán a 7 felé bontás esetén áll önmagában, hanem rendre egyelemű csoportot alkot a *válik* ige. Ennek oka, hogy valóban egyedi változást okoz az *el* igekötő az ige argumentumszerkezetében. A *válik* ige igekötő nélkül szinte kizárólag a *-vÁ* ragos bővítménytípussal szerepelt a vizsgált mondatokban, azonban az igekötő megjelenésének hatására ez a bővítménytípus letiltódik és helyette a *-tÓl* ragos bővítménytípus előfordulási gyakorisága emelkedik meg. Ez az ige a többi vizsgált igéhez képest tehát valóban a periférián helyezkedik el, mivel teljesen idiomatikus: *válik vmivé* igéje érezhetően nem ugyanaz, mint az *elválik vmitől* (pl. *Mici nem a sminktől válik szebbé* vs. *Mici már másodszorra vált el Kristóftól*).

A 10. táblázat a 4 jelű csoport adatait tartalmazza.

8. táblázat. 7-es csoportosítás, 2 jelű csoport részlete

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
alszik	-0,04	0,00	-0,06	-0,03	0,02	-0,02	-0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,00	-0,03	2
felejt	-0,01	0,01	-0,07	-0,07	0,02	-0,01	-0,05	0,00	0,00	0,00	0,00	-0,01	0,01	0,00	-0,01	2
fogy	-0,06	0,02	-0,05	-0,04	-0,03	-0,02	-0,08	0,00	-0,02	0,00	-0,03	0,00	-0,02	0,00	0,01	2
hagy	0,00	-0,01	-0,10	-0,03	-0,09	0,03	-0,06	-0,01	0,00	0,00	0,01	0,00	-0,02	0,00	0,00	2
indul	0,00	-0,06	-0,13	-0,10	-0,05	-0,05	0,01	-0,01	-0,01	0,02	-0,01	-0,02	0,00	0,00	-0,01	2
indít	-0,02	-0,02	0,01	-0,01	-0,11	-0,02	-0,01	-0,01	-0,02	0,00	0,00	-0,01	0,00	0,00	0,00	2
jut	-0,01	-0,13	-0,10	0,01	-0,13	0,01	0,00	-0,07	-0,02	0,07	-0,06	0,03	-0,02	0,00	0,23	2
jön	0,01	-0,01	-0,04	-0,03	0,03	-0,01	0,00	-0,10	-0,03	0,00	-0,03	0,00	-0,01	0,00	0,00	2
kér	-0,02	-0,02	-0,07	0,01	-0,05	-0,01	-0,03	-0,02	-0,02	-0,07	-0,01	-0,10	0,00	0,00	-0,01	2
készül	0,00	-0,01	-0,01	-0,03	-0,26	0,01	0,01	-0,02	-0,05	-0,01	-0,04	-0,01	0,00	0,00	0,02	2
küld	-0,06	0,11	0,03	0,04	-0,07	-0,02	0,00	-0,09	-0,02	0,01	-0,02	0,00	0,00	0,00	0,01	2
lát	0,09	-0,06	-0,13	-0,04	-0,03	0,33	-0,02	0,00	-0,02	-0,01	0,00	0,01	0,00	0,00	0,04	2
megy	-0,01	-0,02	-0,07	-0,04	-0,07	-0,02	-0,06	-0,08	0,00	0,02	0,01	0,02	0,00	0,00	0,00	2
mond	-0,08	0,06	0,02	0,00	0,03	-0,02	0,02	0,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00	2
telik	-0,01	-0,04	-0,05	-0,03	-0,15	-0,03	-0,10	-0,05	-0,01	0,00	-0,07	0,02	0,00	0,00	0,03	2
utazik	-0,04	0,00	-0,11	-0,15	-0,05	-0,02	-0,03	-0,17	0,02	0,02	0,01	-0,01	0,00	0,00	-0,02	2
visz	-0,01	-0,02	-0,03	-0,02	-0,18	-0,04	0,02	-0,17	0,02	0,00	0,06	0,00	0,00	0,00	-0,01	2
hisz	0,16	-0,11	-0,23	-0,01	0,01	0,00	0,00	0,00	0,01	0,00	0,02	0,00	0,00	0,00	0,00	2
játszik	0,11	0,01	-0,29	-0,13	0,01	0,04	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,00	2
olvas	0,05	0,00	-0,19	-0,03	0,01	-0,05	0,00	0,00	-0,14	0,00	-0,02	0,00	-0,01	0,00	0,00	2

9. táblázat. 7-es csoportosítás, 3 jelű csoport

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
válik	-0,01	-0,02	-0,05	-0,08	-0,04	0,02	0,02	0,00	-0,01	0,00	-0,03	0,20	0,03	-0,90	-0,01	3

10. táblázat. 7-es csoportosítás, 4 jelű csoport

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
bocsát	-0,12	-0,03	0,03	0,00	-0,81	0,03	-0,05	-0,06	0,02	-0,02	0,21	0,09	0,00	0,00	0,01	4
helyezkec	0,01	0,01	0,19	0,24	-0,68	0,00	-0,04	-0,11	-0,01	0,00	0,01	0,01	0,05	0,00	-0,01	4
kel	-0,07	-0,01	0,08	0,10	-0,45	0,00	0,01	-0,01	0,00	-0,01	0,05	0,01	0,00	0,00	0,02	4
költ	-0,02	0,01	0,02	0,02	-0,48	-0,01	0,00	0,00	0,00	0,00	0,02	0,00	0,01	0,00	0,07	4
utasít	0,20	0,01	0,01	0,07	-0,45	-0,01	0,06	-0,01	0,00	-0,04	0,00	0,00	0,00	0,00	0,00	4
ítél	0,03	-0,34	0,04	0,01	-0,53	0,00	0,13	0,00	0,01	0,00	-0,01	0,00	0,00	0,00	-0,01	4

A 4 jelű csoport teljes mértékben megfelel a hipotézisnek, az ideális klaszter elképzelésének. A *-rA* ragos bővítménytípus esetén nagymértékű csökkenést láthatunk az igekötő megjelenésének hatására a következő igék esetében: *bocsát, helyezkedik, kel, költ, utasít, ítél*. Kisebb-nagyobb mértékű növekedés mutatkozik ezen túl a *-bA* és az *-On* ragos bővítmények esetén. A vizsgált igék ezen csoportjánál azt a megfigyelést tehetjük, hogy az igekötős ige jelentésének kevés köze van az igekötőtlen ige jelentéséhez (pl. *kel – elkel, helyezkedik – elhelyezkedik*).

A 11. táblázatban az 5 jelű csoportot láthatjuk.

Az 5 jelű csoport esetében szintén a csökkenés az, ami elsősorban szembetűnő a táblázat alapján, méghozzá a datívuszi bővítménynél. Láthatunk továbbá kisebb-nagyobb mértékű csökkenést az akkuzatívuszi bővítmény oszlopában is, valamint

az igekötő egyedi hatásaként említhető a *nevez* esetében a *-rÓl,* a *köszön* esetében a *-tÓl* és a *tart* esetében az *-ig* ragos bővítménytípusok esetében.

11. táblázat. 7-es csoportosítás, 5 jelű csoport

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
köszön	-0,55	-0,32	0,08	0,07	0,00	0,09	-0,03	0,00	0,00	0,00	-0,01	0,31	0,01	0,00	0,01	5
tart	-0,48	-0,37	-0,16	-0,17	-0,04	-0,06	0,02	-0,01	-0,05	0,00	0,03	-0,03	-0,01	0,00	0,48	5
nevez	-0,08	-0,41	-0,03	0,02	0,01	-0,01	-0,02	-0,01	0,29	0,00	-0,01	0,00	0,00	0,00	-0,01	5
tűnik	0,00	-0,68	0,05	0,02	0,00	0,02	0,07	-0,01	0,07	-0,01	0,08	0,00	0,00	0,00	0,00	5
veszik	-0,11	-0,50	-0,01	-0,04	-0,03	0,02	-0,06	-0,21	0,01	0,00	0,01	0,00	-0,01	0,00	0,00	5

A 12. táblázatban a 6 jelű csoport adatait láthatjuk.

12. táblázat. 7-es csoportosítás, 6 jelű csoport

lemma	acc	dat	BA	ON	RA	VA	UL	BA	RÓ	HO	BÓ	TÓ	NÁ	VÁ	IG	7
pusztít	0,55	0,00	-0,25	-0,19	0,00	-0,02	-0,02	0,01	0,00	0,00	0,03	0,00	0,00	0,00	0,00	6
kerül	0,72	-0,02	-0,16	-0,13	-0,35	-0,04	-0,02	-0,40	0,00	-0,02	0,00	-0,01	-0,01	0,00	-0,01	6
dönt	0,60	-0,01	-0,07	-0,19	-0,01	0,02	-0,03	-0,01	-0,59	0,00	0,00	0,00	0,00	0,00	-0,01	6
határoz	0,76	-0,01	0,01	-0,03	0,12	-0,07	-0,07	0,01	-0,70	0,00	0,01	0,00	0,00	0,00	0,00	6

A 6 jelű csoport négy elemet számlál: *pusztít, kerül, dönt, határoz*. Ezek esetében egyértelmű és nagymértékű növekedést figyelhetünk meg a tárgyi bővítmény előfordulási gyakoriságának mértékében. Csökkenést tendenciózusan nem kifejezetten azonosíthatunk, a *-bA* és *-On* ragos bővítményeknél láthatunk több esetben csökkenést. A *-rÓl* ragos bővítményeknél a *dönt* és *határoz* igéknél láthatunk nagyobb csökkenést. Ez olyannyira megkülönbözteti ezt a két igét a klaszterben szereplő másik kettőtől, hogy a 8 csoportba sorolás esetén ők már külön klasztert alkotnak.

A K-means előnyei közé tartozik, hogy nagy adattömböket gyorsan tudunk vele feldolgozni, azonban nem hagyhatjuk figyelmen kívül az esetlegességet, ugyanis – ahogy már korábban említettem – a K-means minden futtatáskor különböző kimenetet produkál (a seedérték beállítása nélkül). Ennek ellensúlyozására többféle megoldás is felmerült, például sok futtatás eredményeinek átlagolása, illetve egy ún. „hőtérkép” készítése is, melyet a következőkben ismertetek.

3.2 Hőtérkép

Amint fentebb olvasható, a hőtérkép (melyre az angol szakirodalom a *heatmap* szót használja) azt a problémát hivatott kiküszöbölni ebben a kutatásban, hogy a K-means klaszterező minden futtatás során más kimenetet ad. Ez a feldolgozási lépés azt szemlélteti, hogy két ige milyen gyakran kerül azonos klaszterbe.

A hőtérkép elkészítéséhez 100 alkalommal lett lefuttatva a K-means klaszterező úgy, hogy 5 klaszterbe sorolja az igéket. A reprodukálhatóság érdekében itt is be lettek állítva a seedértékek a futtatásoknál, a 100 futtatásnál 100 különböző seedérték. Ezután a program létrehozott egy olyan táblázatot, amelyben összeadja, hogy a korpuszban lévő egyes igék hányszor kerültek egy

klaszterbe egy adott másik igével. Így a táblázatban balról jobbra átlósan a cellákban mindig a 100 érték szerepel, mivel az igék önmagukkal értelemszerűen mindig egy klaszterbe kerültek. Az 5 klaszterbe való sorolás hőtérképének részletét a 4. táblázat mutatja (a teljes táblázat elérhető a GitHubon).[3]

13. táblázat. A heatmap részlete.

level_0	pusz	kerü	hatá	dönt	vége	bír	szám	mes	képz	vesz	nyer	foga	enge	szen	lop	köve	ér	rend	rabo
pusztít	100	96	78	78	65	65	62	63	60	58	56	23	21	26	20	20	20	18	16
kerül	96	100	81	81	61	61	58	59	56	54	52	19	17	22	16	16	16	14	12
határoz	78	81	100	100	43	43	40	41	38	36	34	2	1	5	1	1	1	1	0
dönt	78	81	100	100	43	43	40	41	38	36	34	2	1	5	1	1	1	1	0
végez	65	61	43	43	100	96	93	96	92	92	89	56	54	59	52	52	52	50	48
bír	65	61	43	43	96	100	97	96	94	90	91	58	56	61	54	54	54	52	50
számol	62	58	40	40	93	97	100	97	97	93	94	61	59	64	57	57	57	55	53
mesél	63	59	41	41	96	96	97	100	96	94	93	60	58	63	56	56	56	54	52
képzel	60	56	38	38	92	94	97	96	100	94	95	62	60	65	60	60	60	58	56
veszít	58	54	36	36	92	90	93	94	94	100	97	64	62	67	60	60	60	58	56
nyer	56	52	34	34	89	91	94	93	95	97	100	67	64	69	62	62	62	60	58
fogad	23	19	2	2	56	58	61	60	62	64	67	100	97	96	92	92	92	90	88
enged	21	17	1	1	54	56	59	58	60	62	64	97	100	95	93	93	93	91	91
szenved	26	22	5	5	59	61	64	63	65	67	69	96	95	100	93	93	93	91	89
lop	20	16	1	1	52	54	57	56	60	60	62	92	93	93	100	100	100	98	96
követ	20	16	1	1	52	54	57	56	60	60	62	92	93	93	100	100	100	98	96
ér	20	16	1	1	52	54	57	56	60	60	62	92	93	93	100	100	100	98	96
rendel	18	14	1	1	50	52	55	54	58	58	60	90	91	91	98	98	98	100	98
rabol	16	12	0	0	48	50	53	52	56	56	58	88	91	89	96	96	96	98	100

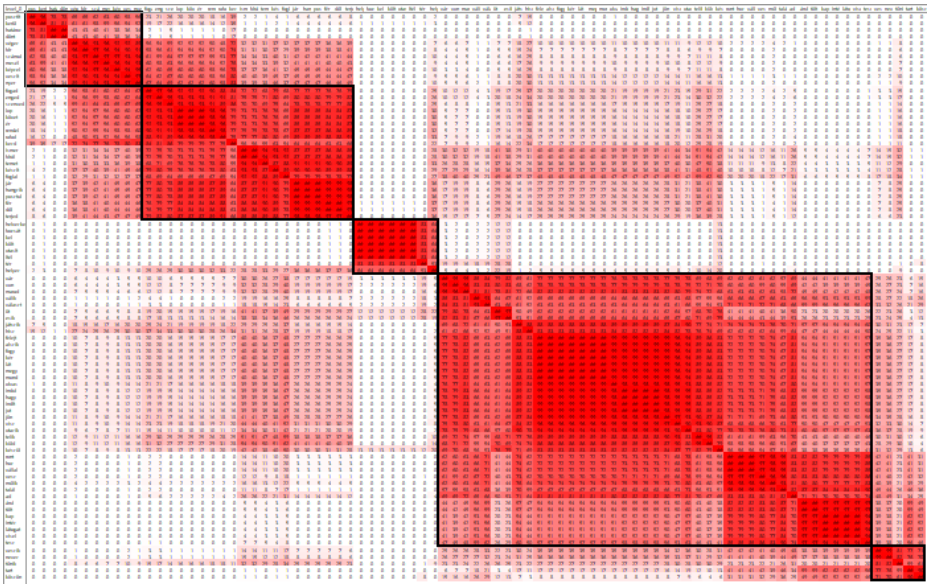
A táblázat rendezése a kutatás jelen állapotában egy dendrogram elkészítése alapján történt. A dendrogram egy fa diagram, amely egy hierarchikus klaszterezési módszer eredményeit szemlélteti: azt mutatja meg, hogy az egyes klaszterek hogyan épülnek fel úgy, hogy U-alakú kapcsolatot rajzolunk a vizsgálati minta részhalmazai közé. Az U-kapcsolás tetejénél egyesülnek a részhalmazok. Az U-link két szára jelzi, hogy mely klasztereket egyesítették. Az U-link két lábának hossza az egyesített klaszterek közötti távolságot jelenti, de a jelenlegi kutatás szempontjából a dendrogram jelentősége az igék rendezésében rejlik. A hierarchikus klaszterezés bemeneteként az a hőtérkép szolgált, amelyet a 13. és 14. táblázatban is láthatunk. A dendrogram úgy rendezi sorba az igéket, hogy azok kerülnek egymás mellé, amelyek egy klaszterbe tartoznak a hierarchikus klaszterezés alapján, így a hőtérképen is használhatjuk ezt az igesorrendezést, így kiküszöbölve a hőtérkép kézi rendezését. A dendrogram elkészítéséhez az „átlag” (angol szakirodalomban *average*) módszert alkalmaztam.

A táblázatról tehát azt tudjuk leolvasni, hogy a 100 futtatás után hányszor került egy klaszterbe az adott igepár, minél nagyobb ez a szám (maximum 100), annál valószínűbb, hogy az adott igepár esetében ugyanazt a változást okozza az *el* igekötő megjelenése. A táblázat értelmezésében a színezés is segít: minél sötétebb piros egy cella, annál többször került egy csoportba az igepár. Érdemes

[3] https://github.com/gyulailivia/AlkNyelvDok22

megfigyelni azt is, mi rajzolódik ki a hőtérképen, ha az egészet egyben nézzük: a teljes klaszterezést az 14. táblázat mutatja.

14. táblázat. A heatmap teljes változata



A hőtérképről leolvasható, hogy 5, különálló tömb látszik kirajzolódni rajta, melyeket fekete keretezés vesz körül. A táblázatban balról jobbra fentről lefelé átlósan mindig a 100 érték szerepel, mivel értelemszerűen minden ige önmagával mind a 100 futtatás alkalmával egy csoportba került. Az így kapott csoportok a K-means klaszterezések átlagaként fogható fel, tehát leolvasható róla, hogy a kapott igecsoportok milyen bizonyossággal kerülnek egy klaszterbe, következésképpen milyen bizonyossággal mutatkozik meg az igekötő argumentumszerkezet-változtató hatása.

A nagyobb tömbökön belül megfigyelhetünk több kisebb, sötétebb „foltot”. Ez azt jelenti, hogy az ötös klaszterezés nem feltétlenül elég, ha több klasztert kértünk volna, akkor ezek a sötétebb foltok külön osztályt alkotnának, de a jelen fejezetben az öt csoportba osztás adatait fogom elemezni, valamint az 1.1 fejezetben megfogalmazott célnak megfelelően összehasonlítom, hogy a hőtérkép kimenete (igecsoportok) mennyire tükrözik a K-means klaszterezés egy alkalommal való futtatása után kapott kimenetet.

Az első klasztert a következő igék alkotják: *pusztít, kerül, határoz, dönt, végez, bír, számol, mesél, képzel, veszít, nyer*. A klaszterben a legalacsonyabb adat a 34, a legnagyobb 100. Az első négy ige láthatóan szorosan kapcsolódik egymáshoz, ez az eredmény összhangban van a korábban bemutatott 7-es klaszterezés 6 jelű csoportjával. A klaszterben szereplő többi ige tőlük kissé elkülönülve jelenik meg, több klaszterszám meghatározása esetén valószínűleg ők egy külön osztályt alkotnának. Ezek az igék szintén egy csoportba tartoztak már az egyszeri klaszterezés során is, de ezzel a módszerrel bebizonyosodott, hogy valóban

ugyanaz az *el* igekötő ugyanazon argumentumszerkezet-változtató hatása jelenik meg ezen igék esetében.

A második klaszterbe a *fogad, enged, szenved, lop, követ, ér, rendel, rabol, kezd, ismer, bírál, temet, készít, foglal, jár, hangzik, pusztul, fér, dől, terjed* igék tartoznak. Itt a legkevesebb együtt előfordulás 54, míg a legnagyobb 100. a *fogad*-tól a *rabol*-ig tartó igék szorosan összekapcsolódnak a hőtérkép alapján, méghozzá az egyszeri klaszterezés során ők is a 6 jelű csoport tagjaiként szerepelnek, tehát itt is bebizonyosodott, hogy valóban ugyanazon hatás érvényesül ezeknél az igéknél. Érdekes azonban, hogy az előző csoport második felével miért nem kerültek most egy osztályba. Ha azonban tovább finomítjuk az osztályozást (pl. 8 klaszterre, melynek bemutatása jelen tanulmány esetében nem cél), akkor láthatjuk, hogy már egy osztályt fognak alkotni. A klaszter többi igéje szintén egy csoportba tartozott már az előzőekben bemutatott klaszterezés során is (1 jelű csoport).

A 3. bekeretezett tömb elemei a következők: *helyezkedik, bocsát, kel, költ, utasít, ítél, tér, helyez*. Ezek az igék egy igen jól elkülönülő csoportot alkotnak, ahol a legkisebb együttes előfordulási szám az 53, a legnagyobb a 100, melyből igen sok van ebben a klaszterben. A *tér* és *helyez* igék kivételével ezek az elemek szerepelnek az egyszeri futtatás során kapott 4 jelű klaszterben is.

A 4. klaszter egy igen nagy elemszámú csoport, tagjai a következők: *vár, von, marad, válik, választ, üt, esik, játszik, hisz, felejt, alszik, fogy, kér, lát, megy, mond, olvas, indul, hagy, indít, jut, jön, visz, utazik, telik, küld, készül, ront, hoz, vállal, vesz, múlik, talál, ad, árul, tölt, kap, intéz, látogat, visel, tesz*. A csoportban szereplő igepárok legalacsonyabb együttes előfordulása 20, a legnagyobb 100. A *felejt, alszik, fogy, kér, lát, megy, mond, olvas, indul, hagy, indít, jut, jön, visz, utazik* igék mindegyike az egyszeri futtatássorán a 2 jelű csoportba tartoztak, ők alkotnak egy egységesebb „foltot” a hőtérképen a 4. klaszterben, azonban az ő argumentumszerkezeti vektoraikat megvizsgálva már látszódott a 8. táblázatban is, hogy ez egy igen vegyes csoport, ahol nem állapítható meg tendenciózus változás az igekötő megjelenésének hatására. A másik egybefüggő csoport ezen az osztályon belül a *ront, hoz, vállal, vesz, múlik, talál, ad, árul, tölt, kap, intéz, látogat, visel, tesz* igék. A korábbi klaszterezésnél ők is a 2 jelű csoportba tartoztak, azonban az ő esetükben (7. táblázat) van hasonlóság az argumentumszerkezeti vektorok között.

Végül az 5. klaszterbe a *veszik, nevez, tűnik, tart, köszön* igék tartoznak. Ők a 11. táblázatban szereplő 5 jelű csoportot is alkotják. A hőtérképen a legalacsonyabb együttes előfordulás 59, míg a legnagyobb 100.

Fontos kihangsúlyozni, hogy a hőtérkép fentebbi leírása az öt klaszterre való 100 futtatás eredménye, amennyiben nagyobb klaszterszámot választunk, ezek a csoportok még finomabb differenciákat mutatnának, azonban a leírás célja az volt, hogy bemutassam, hogy lehet értelmezni a hőtérkép által nyújtott adatokat. Az, hogy valójában melyik csoportosítás volna a legideálisabb, az a

kutatás jelen állásában még nem került meghatározásra, azonban a bemutatott módszer láthatóan alkalmas az igék csoportosítására.

4. Következtetések

A tanulmányban bemutatott kutatás középpontjában egy olyan módszer bemutatása állt, amellyel automatikus módon lehet meghatározni, hogy mely igék alkotnak egy csoportot az igekötő argumentumszerkezetben okozott változása alapján. Az elvégzett kutatás fő célja ennek a módszernek a kidolgozása volt, amelyet az *el* igekötő 88 igével való összekapcsolódásának vizsgálatával mutattam be. A kutatást korpuszadatok alapján végeztem el, melyeket a Mazsola (Sass, 2009) lekérdezőeszköz segítségével kértem le. Ezután az adatok megtisztítása következett, majd azok feldolgozása automatikus módszerekkel. Az igekötők argumentumszerkezet-változtató hatásainak vizsgálatához a lekért igekötő nélküli és igekötős igék argumentumszerkezeti vektorok értékeinek összehasonlítását használtam alapul. Az argumentumszerkezetben megjelenő bővítmények előfordulási gyakoriságának különbségét használtam az összehasonlításhoz. Az igekötők argumentumszerkezet-változtató hatásainak megfigyeléséhez a K-means klaszterezési módszert alkalmaztam, mivel ezzel a módszerrel gyorsan lehet nagy adatmennyiséget feldolgozni. A klaszterezés során kapott eredményekből szemrevételezéssel lehetett a klaszterező által javasolt csoportokat megvizsgálni. A szemrevételezés során megállapítható volt, hogy valóban talált olyan csoportokat az algoritmus, amelyek elemeinél hasonló változást okozott az igekötő megjelenése. A K-means klaszterező esetlegességének kiküszöbölése végett sor került egy ún. „hőtérkép” elkészítésére, amely során 100 alkalommal lett lefuttatva 5 klaszterre. Az így kapott eredményeket egy hőtérképen lehetett összesíteni, amely azt mutatta meg, hogy a 100 klaszterezés során az egyes igepárok hányszor kerültek ugyanabba a csoportba. Ennek köszönhetően az is megvizsgálható lett, hogy mennyire valószínű, hogy egyes igék valóban egy klaszterbe tartoznak. Ezek alapján elmondható, hogy a kutatáshoz kapcsolódó hipotézis beigazolódott, mely szerint a klaszterező segítségével csoportosíthatóak azok az igék, amelyeknél az igekötő hasonló mértékű változást okozott a korpuszadatokból kinyert argumentumszerkezeti vektorok értékeinek összehasonlítása alapján, ugyanis valóban talált a klaszterező már egy futtatás alkalmával is olyan csoportokat, ahol hasonló változást okozott az igekötő megjelenése.

A kutatás eredményei összhangban vannak az eddig elvégzett, intuitív módszereket is alkalmazó vizsgálataim eredményeivel. A bemutatott módszer fontos elméleti hozadékkal szolgál a hazai igekötő-szakirodalom számára, mivel a klaszterezés kimeneteként kapott igecsoportok esetén szemantikai hasonlóságot is fel lehet fedezni az igék között, valamint a kutatás hatékonyan ötvözi a korpusznyelvészet és számítógépes nyelvészet területeit.

Irodalom

Gyulai Lívia (2019). Nem kompozicionális igekötős igék argumentumszerkezetének korpuszalapú vizsgálata. In Ludányi Zsófia & Gráczi Tekla Etelka (szerk.), *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019*. (44–58.) Budapest: MTA Nyelvtudományi Intézet.

Gyulai Lívia (2021). Az igekötők legjellemzőbb argumentumszerkezet-változtató hatásainak korpuszalapú vizsgálata. In Gráczi Tekla Etelka & Ludányi Zsófia (szerk.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2021.* (176–201). Budapest: Nyelvtudományi Kutatóközpont.

Kalivoda Ágnes (2017). Az igekötők gépi annotálásának problémái. In Ludányi Zsófia (szerk.), *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: Alkalmazott Nyelvészeti Doktorandusz Konferencia.* (100–108). Budapest: MTA Nyelvtudományi Intézet.

Oravecz Csaba, Váradi Tamás & Sass Bálint (2014). The Hungarian Gigaword Corpus. In *LREC 2014 Proceedings*. (1719–1723). Reykjavik, Izland: ELRA.

Pethő Gergely, Sass Bálint, Kalivoda Ágnes, Simon László & Lipp Veronika (2022). Igekötő-kapcsolás. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *MSZNY 2022, XVIII. Magyar Számítógépes Nyelvészeti Konferencia.* (77–93). Szeged: TTIK, Informatikai Intézet.

Sass Bálint (2009). „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi Tamás (szerk.), *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból*. (117–129.) Budapest: MTA Nyelvtudományi Intézet.

Sass Bálint (2015). 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet. In Tanács Attila, Varga Viktor & Vincze, Veronika (szerk.), *MSZNY 2015, XI. Magyar Számítógépes Nyelvészeti Konferencia.* (303–308). Szeged: JATEPress,

Szécsényi Tibor (2019). Argumentumszerkezet-variánsok korpusz alapú meghatározása. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XV. Magyar Számítógépes Nyelvészeti Konferencia.* (315–329). Szeged, SZTE TTIK Informatikai Intézet.

Szécsényi Tibor & Virág Nándor (2022). Az ige helyhatározói bővítményeinek megkülönböztetése és az argumentumszerkezeti variánsok korpusz alapú szétválasztása. In Berend Gábor, Gosztolya Gábor & Vincze Veronika (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. (535–647). Szeged: SZTE TTIK Informatikai Intézet.

Váradi Tamás (2002). The Hungarian National Corpus. In *Proceedings of the 3rd LREC Conference.* (385–389). Las Palmas, Spanyolország.

Vincze Veronika, Szauter Dóra, Almási Attila, Móra György, Alexin Zoltán & Csirik János (2010). Hungarian Dependency Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation.* Valletta, Malta: European Language Resources Association.

Zsibrita János, Vincze Veronika & Farkas Richárd (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP 2013.* (763–771). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.