Az ige helyhatározói bővítményeinek megkülönböztetése és az argumentumszerkezeti variánsok korpusz alapú szétválasztása

Szécsényi Tibor[1], Virág Nándor[2]

[1,2] SZTE Általános Nyelvészeti Tanszék
[1] szecsenyi@hung.u-szeged.hu
[2] virag.nandor9910@gmail.com

Kivonat: Tanulmányunkban Szécsényi (2019) argumentumszerkezeti modelljét vesszük alapul, melyben nincs kategorikus vonzat-szabad bővítmény megkülönböztetés, helyette a különböző argumentumtípusokat az egyes igék melletti megjelenési valószínűségével jellemezve adja meg az ige argumentumszerkezetét. Röviden ismertetjük az argumentumszerkezet valószínűségi vektor alapú modelljét és bemutatjuk, hogyan lehet a modell segítségével az igék helyhatározói bővítményei között megkülönböztetni a valódi vonzatokat és a tematikus vonzatokat a szabad bővítményektől. Ezután a több argumentumszerkezeti variánssal rendelkező igéknél illusztráljuk, hogyan lehet az ige mellett megjelenő több vonzatszerű bővítményt egy vagy több argumentumszerkezeti variánshoz rendelni.

1 Bevezetés

Az igék argumentumszerkezetének ismerete nélkülözhetetlen a természetes nyelvi szövegek szintaktikai (lásd pl. Kovács és mtsai, 2016) és szemantikai feldolgozásához (lásd pl. Gildea és Jurafsky, 2002; Palmer és mtsai, 2005). Az argumentumszerkezet vonzatlistaként való megadása nem tükrözi vissza azt a nyelvhasználói viselkedést, hogy az elvileg kötelező vonzatok sem jelennek meg mindig az ige mellett, más szabad bővítmények pedig az egyes igék mellett viszonylag gyakoriak.

Tanulmányunkban Szécsényi (2019) argumentumszerkezeti modelljét vesszük alapul, melyben nincs ilyen kategorikus vonzat-szabad bővítmény megkülönböztetés, helyette a különböző argumentumtípusokat az egyes igék melletti megjelenési valószínűségével jellemezve adja meg az ige argumentumszerkezetét. A 2. szakaszban röviden ismertetjük az argumentumszerkezet valószínűségi vektor alapú modelljét. A 3. szakaszban bemutatjuk, hogyan lehet a modell segítségével az igék helyhatározói bővítményei között megkülönböztetni a valódi vonzatokat és a tematikus vonzatokat a szabad bővítményektől. Végül a 4. szakaszban a több argumentumszerkezeti variánssal rendelkező igéknél illusztráljuk, hogyan lehet az ige mellett megjelenő több vonzatszerű bővítményt egy vagy több argumentumszerkezeti variánshoz rendelni.

A `https://github.com/szecsenyi/MSZNY2022-Szecsenyi-Virag` githubon elérhetők a kutatáshoz tartozó korpuszadatok.

2 Az igei argumentumszerkezetek vektor alapú jellemzése

Az argumentumszerkezetek jellemzése a hagyományos leíró nyelvészeti elméletekben az ige (régens) vonzatainak felsorolásával és azok elvárt tulajdonságainak megadásával történik. De születtek más jellegű modellek is, a magyarban például (Sass, 2018; 2020) duplakocka modellje lehetővé teszi a vonzatok elkülönítését az idiomatikus kifejezésektől, (Kálmán, 2006; 2016) pedig az igei bővítménytípusok több fajtáját is megengedő graduális modellt javasol.

A természetes nyelv feldolgozása során az argumentumszerkezeti leírások használata problémába ütközik, ha az csak az igei vonzatok leírását tartalmazza: a feldolgozandó szövegekben nem csak a vonzatai jelennek meg az igének, hanem más bővítmények is, továbbá az ige vonzatai is sokszor hiányoznak az ige mellől. Máskor azt nehéz eldönteni egy argumentumszerkezet meghatározásakor, hogy egy bővítmény vonzatnak számít-e vagy sem.

Tanulmányunkban a (Szécsényi, 2019) által bemutatott argumentumszerkezet-modellt használjuk, amely nem tesz különbséget a vonzatok és egyéb bővítmények között. A modell az ige mellett megjelenő bővítményeket azok egy jellemző tulajdonsága alapján csoportosítja, esetünkben ez leginkább a bővítmény fejének az esete.

Jelenleg 32 argumentumtípust különböztetünk meg: PV, CP_cnd, CP_imp, CP_ind, HKM, inf, nom, acc, dat, BAN, ON, RA, VAL, UL, BA, RÓL, HOZ, BÓL, TÓL, NÁL, VÁ, IG, ÉRT, KÉNT, KOR, SZOR, NKÉNT, ADP, ADV, FROM, IN, TO. A PV az igekötői bővítmény, CP_cnd-től inf-ig a különböző mondatbővítmények vannak, nom-tól NKÉNT-ig az esetragok, ADP és ADV a névutós és határozói kifejezések, a FROM, IN és TO pedig az irányhármasságot (is) kifejező esetragok és névutók meta-típusa. Ez utóbbi három típus nincs kiegészítő disztribúcióban az előbbiekkel, de a később tárgyalt tematikus vonzat – valódi vonzat megkülönböztetésnél kulcsszerepet játszhatnak. Az igék argumentumszerkezetét ezen argumentumtípusok megjelenési valószínűségével adjuk meg, vagyis egy 32 dimenziós valószínűségi vektorral. Egy argumentumtípushoz tartozó valószínűségi érték 1, ha az adott ige mellett annak bővítményeként mindig megjelenik az adott típusú kifejezés, 0, ha sohasem. Ez a két szélső érték szinte soha nem jelenik meg az igék argumentumszerkezeti leírásában, ha korpuszból meghatározott valószínűségi értékekkel dolgozunk, mivel a valódi nyelvhasználat során még a valódi vonzatok sem jelennek meg minden esetben egy ige mellett (pl. pro-drop, ellipszis, rövid válasz), viszont nagyon sok bővítménytípus lehet adjunktum is a mondatban. Emiatt jellemzően 0 és 1 közötti értékek figyelhetőek meg. A hagyományos vonzat-szabad bővítmény bináris megkülönböztetést egy valószínűségi küszöbérték megadásával kaphatjuk vissza: ha egy argumentumszerkezetben egy argumentumtípus valószínűségi értéke nagyobb, mint a megadott küszöb, akkor az vonzat, egyébként nem. A vonzatsági küszöb argumentumtípusonként változó lehet.

Az argumentumszerkezet valószínűségi vektorát korpuszból határozhatjuk meg. Egy morfológiailag és szintaktikailag annotált korpuszból vesszük azokat a tagmondatokat, amelyek az adott igét tartalmazzák, és megszámoljuk, hogy ezekben a tagmondatokban hány olyan van, amelyben az ige mellett megadott típusú bővítmény szerepel maximális összetevőként: az igét tartalmazó tagmondatok és az igét és a bővítményt tartalmazó tagmondatok aránya adja az argumentum megjelenési valószínűségének az értékét. Jelenleg kétfajta korpusz feldolgozására van elkészített feldol-

gozási láncunk: a kézzel annotált Szeged Dependecia Treebank (SZDT) (Vincze és mtsai, 2010), illetve tetszőleges (de tipikusan MNSz-ból származó) szövegek magyarlanc segítségével elemzett változata (Oravecz és mtsai, 2014; Zsibrita és mtsai, 2013). A tagmondatok és a tagmondatokat alkotó maximális összetevők meghatározása a mondatok függőségi elemzéséből származnak.

argType	freq	argType	freq	argType	freq	argType	freq
nom	0,559251	BA	0,033539	ÉRT	0,005942	CP_ind	0,186934
acc	0,326353	RÓL	0,024550	KÉNT	0,006258	HKM	0,094095
dat	0,055825	HOZ	0,023189	KOR	0,005671	inf	0,104655
BAN	0,115328	BÓL	0,022399	SZOR	0,003663	ADP	0,105482
ON	0,099435	TÓL	0,016555	NKÉNT	0,001271	ADV	0,324841
RA	0,085415	NÁL	0,011734	PV	0,117021	FROM	0,065994
VAL	0,083331	VÁ	0,008620	CP_cnd	0,012245	IN	0,261352
UL	0,062121	IG	0,010440	CP_imp	0,004107	TO	0,161969

1. táblázat. Az argumentumtípusok összesített előfordulási gyakorisága a Szeged Korpuszban.

Az 1. táblázatban látható a Szeged Korpusz magyarlanccal elemzett változatában az egyes argumentumtípusok összesített előfordulási gyakorisága, amely a kijelentő módú igét[1] tartalmazó tagmondatok számának (132 951) és ezekben a tagmondatokban előforduló argumentumtípusok előfordulási számának a hányadosa (a legkisebb előfordulási gyakorisághoz is több mint 500 előfordulási szám tartozik). Ezek az adatok általában használhatók az argumentumtípusok egyes igék melletti előfordulásának vizsgálatánál vonzatsági küszöbértéknek: ha egy ige mellett az egyik argumentumtípus előfordulási gyakorisága magasabb a táblázatban megadott értéknél, akkor tekinthetjük vonzatnak.

Az argumentumszerkezet valószínűségi vektorként való értelmezhetőségét egy kisebb argumentumtípus-halmazon mutatjuk be. Kilenc helyhatározói esetrag (BÓL, BAN, BA, RÓL, ON, RA, TÓL, NÁL, HOZ) előfordulási gyakoriságát vizsgáltuk 13 ige mellett: *ad*, *beszél*, *fél*, *hisz*, *indul*, *javasol*, *jön*, *kap*, *lát*, *nevet*, *rak*, *teremt*, *úszik*. Az esettanulmányok megmutatják, hogyan különböztethetjük meg a helyhatározói esetragok három különböző használatát.

3 Helyhatározói esetragok eloszlási mintázatai

A kutatás ezen részén azokat a ragos kifejezéseket vizsgáltuk, amelyek irányhármasság szerinti hármasokat alkotnak. A vizsgált ragos kifejezések a BÓL, BAN, BA (belső érintkezéses viszony), RÓL, ON, RA (külső érintkezéses viszony), TÓL, NÁL, HOZ (közelítő viszony) esetekben álltak. Feltételeztük, hogy az ezen ragokkal álló igei bővítményeket lehetséges osztályozni előfordulási gyakoriságuk alapján, de szem

[1] Vizsgálatunk során azért szorítkoztunk a kijelentő módú igét tartalmazó tagmondatokra, mert a főnévi igeneves szerkezetek esetében a tagmondathatárok nehezen meghatározhatók, illetve a kijelentő módú tagmondatokat tekintettük az igék argumentumszerkezetét legtisztábban megmutató adatoknak.

előtt tartottuk, hogy az így alkotott csoportok nem lesznek egyértelműen, diszkréten elkülöníthetőek. A csoportokat két tengely mentén állítottuk fel a következő módon:

Az első a vonzat-szabad bővítményi tengely, melynek egyik végpontja a teljesen szabad bővítménység, másik végpontja pedig a teljesen vonzati bővítménység. A másik tengely az adott rag kompozicionalitására vonatkozik. A kompozicionalitás elve szerint az összetett nyelvi kifejezések jelentése kiszámítható az őket alkotó kifejezések jelentéséből és kapcsolódási módjukból. Ez értelmezhető a ragos kifejezésekre is, így itt ezt az értelmezést használjuk. Tehát lehetnek a ragos kifejezések kompozicionálisak és nem kompozicionálisak. A csoportok egy-egy példával láthatóak a következő táblázatban.

	Vonzat	Szabad bővítmény
Nem kompozicionális	1. Valódi vonzat pl. *bízik Péterben*	3. Egyéb szabad határozó pl. *kínjában nevet*
Kompozicionális	2. Tematikus vonzat pl. *Debrecenben/Szegeden lakik*	4. Szabad helyhatározó pl. *énekel az erdőben*

2. táblázat. A bővítmények csoportosítási lehetősége.

A csoportokat értelmezve: a valódi vonzatok csoportját jellemzi, hogy az ilyen bővítménnyel álló igék egy konkrét ragos kifejezést vonzanak, aminek kompozicionális jelentése nem jelenik meg ebben az esetben. Magas gyakorisággal állnak az igei alaptag mellett, sokszor a kívánt jelentés eléréséhez elengedhetetlen a mondatban való megjelenésük.

A tematikus vonzatok csoportjában olyan elemeket találunk, amelyeket az igei alaptag tematikus szerepszerűen vonz, azaz egy irányhármasság szerinti irányt kíván maga mellé, legyen az bármelyik a megfelelő alakokból.

Az egyéb szabad határozók olyan mondatrészek, amelyekben a rag jelentése nem kompozicionális, de csak néha jelennek meg mondatban, nem tekinthetők vonzatnak.

Az utolsó csoport, a szabad helyhatározók csoportja. Ezek esetlegesen jelennek meg, valóban helyviszonyt fejeznek ki. Már itt érdemes megjegyeznünk, hogy az egyéb szabad határozók és a szabad helyhatározók csoportja jelen eszközökkel nem elkülöníthető, így őket egy csoportnak kell kezelnünk, szabad bővítmények néven.

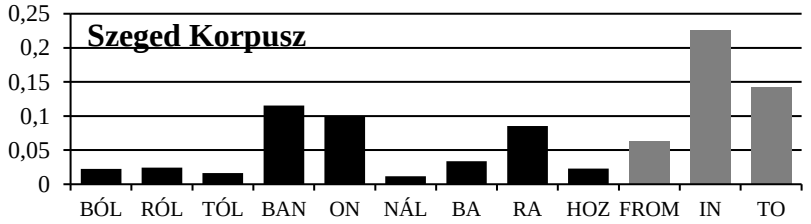
A helyhatározói esetragos bővítményeket 13 ige esetében vizsgáltuk meg: *ad*, *beszél*, *fél*, *hisz*, *indul*, *javasol*, *jön*, *kap*, *lát*, *nevet*, *rak*, *teremt*, *úszik*. Az igék kiválasztási szempontja az volt, hogy előreláthatólag legyenek köztük a megkülönböztethető három csoport mindegyikébe tartozók. További szempont volt még, hogy az igéknek ne legyen sok argumentumszerkezeti variánsa.

A korpuszépítéshez az adatokat az MNSz-ból (v2.0.5) lekért, igénként 1500 véletlenszerűen kiválasztott mondat adta. (Gyulai, 2019) alapján a vizsgált igék nem lehetnek igekötősek, mivel az megváltoztatná a vonzatszerkezeteket, ezzel torzítva az adatsort, ezért eleve olyan mondatokat kértünk le a korpuszból, amelyben a vizsgált igék közelében nem volt igekötő. A mondatok elemzését a magyarlanc függőségi elemzővel végeztük el. Ezután további kézi ellenőrzést végeztünk az elemzett mondathalmazon, részben az esetlegesen a korpuszba került elváló igekötős igék kiszűrése, részben a rosszul elemzett mondatok kitörlése végett. Rosszul elemzett mondatnak azok számítottak, amelyek eleve nem teljes mondatok voltak, vagy amelyekben a

vizsgált igét tartalmazó tagmondat határai vagy annak fő összetevői hibásan lettek meghatározva. Ilyen rosszul elemzett mondat az összes mondat 2–3 százaléka volt.

A korpusz mondatainak megszűrése után az előző szakaszban bemutatott elemzési láncot futtatva kaptuk meg a valószínűségi és gyakorisági táblázatokat, amelyeknek adataiból következtetni tudtunk.

Ahhoz, hogy egy általános eloszlást érhessünk el a megfigyeléshez, a Szeged Korpusz adatait használtuk. Megfigyelhető, hogy általánosságban a BAN, ON és *RA* ragos elemek előfordulása a leggyakoribb a vizsgált kifejezések közül, a többi argumentumtípus viszont viszonylag alacsony előfordulást mutatnak. Már itt is észrevehetjük, hogy a közelítő viszonyt kifejező TÓL, NÁL és HOZ fordulnak elő a legkevesebbszer saját terceikben. Az 1. ábrán láthatjuk az így kapott eredményeket, amelyek értékben megegyeznek az 1. táblázatbeli értékekkel. Feketével az egyes ragok előfordulási gyakoriságát láthatjuk (a legkisebb gyakoriságú NÁL is 1560 előfordulási számot takar), szürkével pedig az irányhármasság szerinti csoportok összesített eredményét.

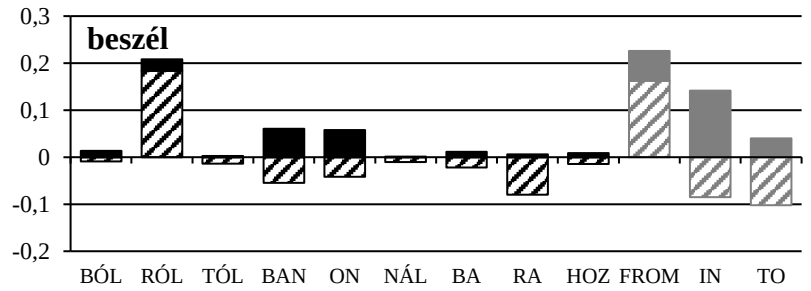


1. ábra: A Szeged Korpuszban található vizsgált ragok előfordulási gyakorisága.

Az eredmények bemutatásaképpen minden csoportból egy-egy igével illusztráljuk a csoportra jellemző tulajdonságokat.

3.1 Valódi vonzattal álló igék

A *beszél* ige prototipikus valódi vonzattal áll. Vonzata a RÓL, így azt várhattuk el az adatok elemzésénél, hogy ez magas előfordulási gyakoriságot fog mutatni, míg a többi toldalék gyakorisága az átlaghoz mérten csökkenni fog.



2. ábra: A helyhatározói ragok gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival a *beszél* ige esetében.

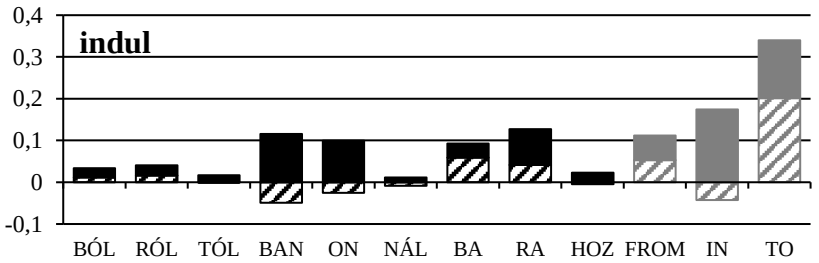
Az eredmények a 2. ábrán láthatóak. Az ábrán az oszlopok 0 fölötti része mutatja az argumentumtípus előfordulási gyakoriságát. Az oszlopok berácsozott részei mutatják a Szeged Korpusz adataihoz mért változást: a 0 fölötti rácsozás az argumentumtípus gyakoriságának a Szeged Korpuszhoz viszonyított növekedésének, a 0 alatti rácsozás pedig a csökkenésének a mértékét mutatja. Az ábrán egyértelműen látszik, hogy minden előfordulás csökkent, csak a RÓL esetében látunk növekedést, és megállapítható, hogy a vizsgált kifejezések közül ez az ige vonzata.

Az adatokat tekintve láthatjuk, hogy az általában leggyakoribb BAN, ON és RA gyakorisága csökkent, a RÓL pedig jelentősen megnövekedett, a vizsgált mondatok 20,80%-ában fordult elő a vonzatnak tekinthető ragos kifejezés. A *beszél* ige 1399 előfordulásában tehát a Szeged Korpusz 2,4%-os átlagos előfordulásához képest majdnem 10-szeres növekedést tapasztalhattunk.

3.2 Tematikus vonzattal álló igék

Az *indul* ige érdekes esetet mutat: tematikus vonzattal áll, azaz az adott irányhármasság szerinti toldalékok közül több is megemelkedett előfordulással mutatkozik. Érdekessége abban áll, hogy nem csak egy, hanem kettő irányt is vonz tematikus szerepként, a kiindulópont és a célpont jelentésű ragok gyakorisága is megnövekedett.

Az ige jelentéséből már adódik a vonzatszerkezeti különlegesség, hiszen inherensen tartalmazza azt, hogy az indulási tevékenységnek része ez a két jelentésaspektus. Az adatokra tekintve azt láthatjuk, hogy a belső és külső érintkezéses viszonyt kifejező BÓL és RÓL, illetve BA és RA jelennek meg megnövekedett számban, azonban a terceket kiegészítő TÓL és HOZ továbbra is alacsony előfordulási gyakorisági adatokat mutat, ezzel is bizonyítva, hogy a közelítő viszonyt kifejező csoport igen ritka. A két különböző tematikus vonzatot kívánó vonzatszerkezet megjelenése lehet az oka annak, hogy az előfordulási számok, bár emelkedtek, nem lettek túl magasak. Ezt azonban jelen eszközökkel nem tudtuk megállapítani, a későbbi, együtt-előfordulással kapcsolatos fejezetben viszont ezzel a jelenséggel foglalkozunk.



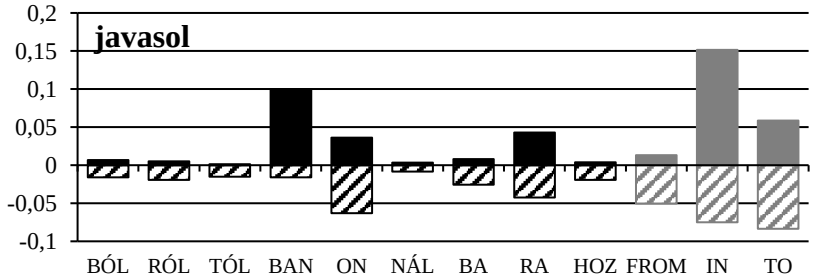
3. ábra: A helyhatározói ragok gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival az *indul* ige esetében.

A 3. ábrán láthatjuk, hogy a fentebb említett ragok előfordulása a Szeged Korpuszhoz mérten megemelkedtek. Az érdekesség az utolsó, összesített oszlopokban látható, hiszen a FROM és a TO adatsor is emelkedett. Ezeken megfigyelhető igazán a tematikus vonzati kategória sajátossága: a csoportba tartozó ragok együttesen emelik meg az előfordulások számát.

3.3 Szabad bővítménnyel álló igék

A szabad bővítményekkel álló igék csoportjára jellemző, hogy nincsenek kiugró előfordulási gyakorisággal álló adatok. Elvárható a Szeged Korpusz adataihoz hasonló eloszlási mintázat, még ha a pontos számok nem is egyeznek meg az ott mért arányokkal. Azt figyeltük meg, hogy az előfordulások gyakorisága általában csökken. Ennek az lehet az oka, hogy a Szeged Korpusz adatai között szerepeltek olyan igék is, amelyek a vizsgált ragos kifejezéseket vonzzák, így azok megemelik az előfordulási mutatókat.

A *javasol* tipikus ige ebben a kategóriában. Az összes vizsgált argumentumtípusnál csökkenést látunk, csak a BAN éri el nagyjából azt az arányt, ami a Szeged Korpuszban megfigyelhető. A 4. ábrán láthatjuk, hogy az összes oszlopnál kisebb-nagyobb lefelé irányuló sötét oszlop látszik, tehát ez egy jó példa a helyhatározóragos vonzat nélküli igékre.



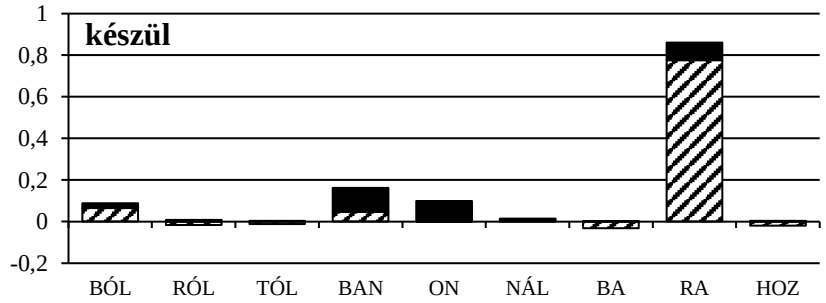
4. ábra: A helyhatározói ragok gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival a *javasol* ige esetében.

3.4 Összesített eredmény

A végrehajtott esettanulmányok alapján messzemenő következtetések nem levonhatóak, de tendenciákat megfigyelhetünk, amelyek bizonyítani látszanak a hipotézisünket, miszerint az általunk felállított csoportok az előfordulási gyakoriságok figyelembevételével megállapíthatóak. Több ige és nagyobb esetszám vizsgálata megmutathatja a pontosabb csoporthatárokat és a felismerhető mintázatokat, amelyeket akár egy automatikus elemző is megtalálhat.

4 Argumentumszerkezeti variánsok elkülönítése

A tematikus helyhatározói vonzatoknál megfigyelhettük, hogy több, az irányultság tekintetében hasonló argumentumtípus megjelenési gyakorisága is megnőtt a Szeged Korpuszban megfigyeltekhez képest. Ekkor úgy elemeztük az adott igét, hogy van neki vonzata, de nem egy specifikus argumentumtípust vonz, hanem egy meghatározott tematikus szerep betöltésére alkalmas bővítményt. Más igéknél is megfigyelhetünk hasonló jelenséget, vagyis hogy nem egyetlen argumentumtípus gyakorisága nő meg, hanem többé is. Az 5. ábrán például a *készül* ige argumentumszerkezetét és a Szeged Korpuszhoz viszonyított változást láthatjuk:

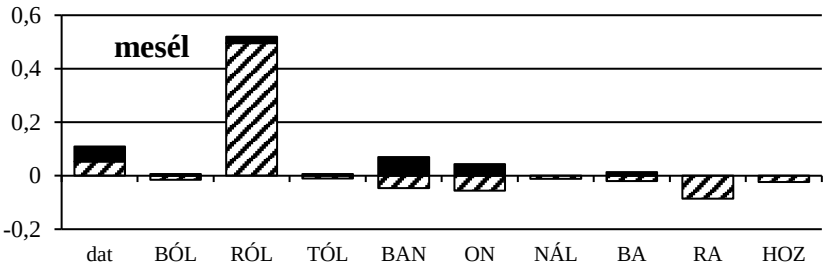


5. ábra: A helyhatározói ragok gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival a *készül* ige esetében.

Látható, hogy bár a BAN megjelenési gyakorisága is megnő kis mértékben, de igazán jelentős növekedést a BÓL és a RA argumentumtípusok esetében figyelhető meg. Ez a két argumentumtípus azonban tematikusan nem sorolható egy csoportba. Itt valójában ugyanannak az igének két különböző jelentése okozza a több argumentumtípusnál is megfigyelhető gyakoribbá válást: pl. *A cipő bőrBŐL készül*, de *A család születésnapRA készül*. A két argumentumtípus a *készül* ige mellett kiegészítő disztribúcióban figyelhető meg, vagy a BÓL, vagy a RA jelenik meg egy tagmondatban. A tematikus vonzatok esetében szintén kiegészítő disztribúciót találunk, de ott az ige jelentése nem változik az argumentumtípussal: *A vonat DebrecenBE/SzegedRE indul*.

Ha egy ige esetében egynél több argumentumtípus megjelenési gyakorisága is megnövekedik, de az argumentumtípusok kiegészítő disztribúcióban állnak egymással, akkor azt mondjuk, hogy az ige két argumentumszerkezeti variánssal rendelkezik. Ha az argumentumok ugyanabba a tematikus csoportba tartoznak, akkor az ige argumentumszerkezeti variánsainak a jelentése megegyezik, egyébként jellemzően különböző jelentésűek. A különböző jelentésű argumentumszerkezeti variánsok (vagy egyszerűen: különböző variánsok) különböző argumentumszerkezeti vektorral jellemezhetőek. A tematikus vonzattal rendelkező variánsokat tekinthetjük egy variánsnak, amelyben a tematikus vonzatba tartozó argumentumtípusok szabad váltakozást mutatnak, és ezen argumentumtípusok egymáshoz viszonyított gyakorisági eloszlása egy nagyobb mintázathoz illeszkedik.

A 6. ábrán látható a *mesél* ige argumentumszerkezeti vektora és annak változása (a helyhatározói esetragokon kívül a datívuszi argumentumtípussal kiegészülve):



6. ábra: A datívuszi és a helyhatározói ragok gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival a *mesél* ige esetében.

Itt szintén két argumentumtípus megjelenési gyakoriságának a megnövekedését láthatjuk, a datívuszi és a RÓL argumentumtípusét. A *készül* igéhez hasonlóan a két megnövekedett gyakoriságú argumentumtípus itt sem tartozik egy tematikus osztályba. Azonban attól eltérően a datívuszi és a RÓL argumentumtípus nem a *mesél* két különböző variánsánál jelentkezik vonzatként, hanem ugyanannál: *Péter a kirándulásRÓL mesélt a nagymamájáNAK*.

Ha az igék argumentumszerkezetét és argumentumszerkezeti variánsait korpusz alapján szeretnénk meghatározni, akkor csak azt vizsgálhatjuk, hogy az adott igét tartalmazó (tag)mondatban argumentumtípusba tartozó bővítmények jelennek meg, az ige különböző variánsait nem tudjuk elkülöníteni. Ekkor két kérdés adódik:

1. Hogyan lehet (lehet-e?) meghatározni, hogy egy igének több variánsa is létezik?
2. Hogyan lehet (lehet-e?) meghatározni, hogy ha egy igének több variánsa is van, akkor a különböző variánsok milyen argumentumszerkezeti valószínűségi vektorral jellemezhetőek?

Ez utóbbi kérdéshez tartozik az a feladat is, hogy meghatározzuk az ige argumentumszerkezeti variánsainak a megjelenési valószínűségét is.

Jelen tanulmányban az első kérdést próbáljuk megvilágítani. Vizsgálatunkban olyan igéket választottunk, amelyeknek vagy több argumentumszerkezeti variánsa van, és van olyan argumentumtípus, amelyiknek a megjelenését túlnyomórészt csak az egyik variánsnál várjuk (csak az egyik variánsnak vonzata), egy másik argumentumtípust pedig csak a másik variánsnál, vagy pedig olyan egyváltozatos igét, amelynek egynél több vonzata van.

Alaphipotézisünk az, hogy abban az esetben, ha egy variánsnak két vonzata is van: A és B, akkor a két vonzat megjelenésének a valószínűsége független egymástól: ha a két vonzat megjelenési valószínűsége az ige mellett P(A) és P(B), akkor annak a valószínűsége, hogy az ige mellett mindkét vonzat megjelenik, P(A)·P(B). Ha viszont a két argumentumtípus az ige más-más variánsánál vonzatok, a két argumentumtípus együttes megjelenésének a valószínűsége P(A)·P(B)-nél jóval kisebb. Ennek tulajdonképpen nullának kellene lenni, de mivel annál a variánsnál, amelynek az egyik argumentumtípus a vonzata, a másik argumentumtípus is megjelenhet szabad bővítményként, de annak sokkal kisebb a valószínűsége – ezért ebben az esetben is előfordulhatnak együtt, csak sokkal kisebb valószínűséggel.

Vizsgálatunkhoz négy igét választottunk ki. A *bevon* igének két variánsa van, amelyeknek az alanyon és a tárgyon kívül vonzata lehet a BA vagy a VAL: *Péter bevonta a barátját fóliáVAL*, ill. *Péter bevonta a barátját a beszélgetésBE*. A már említett *készül* ige két variánsa esetén a BÓL és a RA argumentumtípust vizsgáltuk, a *mesél* egyvariánsos igénél pedig a RÓL és a datívuszi argumentumtípust. A *hív* ige esetében két argumentumszerkezeti variáns vizsgáltunk, az egyiknél csak a tárgyi vonzat kötelező (az alanyon kívül): *Péter hívta a barátjáT (reggelizni)*; a másik variánsnál pedig a tárgyon kívül datívuszi vonzat is van: *Péter öcskösNEK hívta a barátjáT*. Célunk az, hogy a korpuszadatok alapján megmutassuk, hogy a *bevon* és a *készül* igének két variánsa van, a *mesél* igének pedig csak egy, illetve lehetőség szerint kimutatni a *hív* ige két variánsát is.

A korpuszvizsgálat során először lekértünk a Magyar Nemzeti Szövegtárból a Mazsola (Sass 2009) segítségével igénként ezer mondatot. A mondatok közül kiszűrtük azokat, amelyek duplikátumok vagy töredékmondatok voltak, és azokat, amelyekben a keresett ige nem kijelentő módban állt. Ezután kézi annotálással bejelöltük, hogy az

igék melyik argumentumszerkezeti variánsa szerepel a mondatban, és csak azokat mondatokat hagytuk meg, amelyekben a vizsgálni kívánt kettő vagy egy variáns volt. Végül a 2. szakaszban említett feldolgozási lánccal kigyűjtöttük tagmondatonként az adott igékre vonatkozó ige-argumentumtípus előfordulási értékeket, amelyekből a vizsgálni kívánt argumentumtípusok előfordulási számát és együtt előfordulási számát, és ezek gyakoriságát is megkaptuk. Ezek a táblázatok igénként és argumentumszerkezeti variánsonként is rendelkezésre álltak:

bevon	mind	$bevon_1$	$bevon_2$
ige	289	251	38
acc	258	222	36
VAL	41	5	36
BA	219	218	1
acc&VAL	38	4	34
acc&BA	195	194	1
VAL&BA	6	5	1

mesél	mind
ige	302
acc	77
RÓL	157
dat	33
acc&RÓL	32
acc&dat	9
RÓL&dat	7

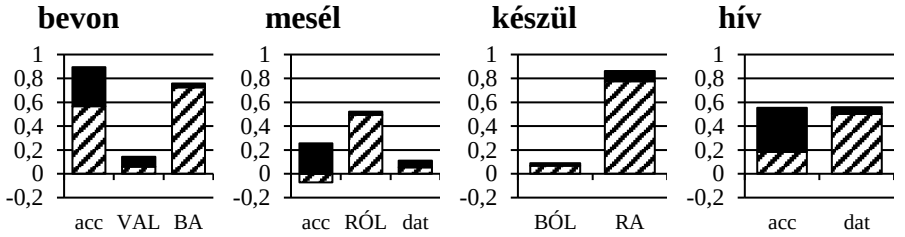
készül	mind	$készül_1$	$készül_2$
ige	475	46	429
BÓL	42	36	3
RA	409	0	409
BÓL&RA	2	0	2

hív	mind	$hív_1$	$hív_2$
ige	681	287	394
acc	377	224	153
dat	380	4	376
acc&dat	145	3	142

3. táblázat. A *bevon*, *mesél*, *készül* és *hív* igék és vonzataik előfordulási száma, illetve a vonzatok együtt-előfordulási adatai a vizsgált korpuszban.

A 3. táblázatok első oszlopai a vizsgált igékre vonatkozó előfordulási számokat tartalmazzák, majd ugyanezt az ige különböző variánsai esetén. Az első sorban a megfigyelt igék (tagmondatok) számát láthatjuk, alatta az egyes vizsgált argumentumtípusok előfordulási számait, majd az argumentumtípusok páronkénti együttes előfordulási számait. Az előfordulási gyakoriságokat az argumentumtípusok (vagy párok) előfordulási számainak és az ige(variáns) előfordulási számainak hányadosaként kapjuk.

A négy ige mellett megjelenő argumentumtípusok valószínűségi értékei láthatók a 7. ábrán, illetve azok változása a kontrollként használt Szeged Korpuszhoz viszonyítva.



7. ábra: A *bevon*, *mesél*, *készül* és *hív* igék vonzatainak gyakorisági adatainak összehasonlítása a Szeged Korpusz adataival.

Az ábrákból, illetve a mögöttük levő gyakorisági adatokból a korábban elmondottaknak megfelelően nem következtethetünk arra, hogy az ábrázolt argumentumtípusok ugyanazon variáns vonzatai-e, vagy különbözőeké.

A korpuszadatokból azonban látszik, hogy a kérdéses argumentumtípusok milyen gyakorisággal fordulnak elő egyszerre az ige környezetében. Ha ugyanazon variánshoz tartoznak, akkor a hipotézisünk szerint az együttes előfordulási gyakoriság az egyes gyakoriságok szorzatához hasonló mértékű, ha viszont külön variánshoz tartoznak, akkor az együttes előfordulási gyakoriságuk jelentősen kisebb ennél. Ezeket az várt és a ténylegesen megfigyelt együtt-előfordulási gyakoriságokat mutatja a 4. táblázat.

bevon		*mesél*		*készül*		*hív*	
p(VAL)	0,142	p(RÓL)	0,520	p(BÓL)	0,088	p(acc)	0,554
p(BA)	0,758	p(dat)	0,109	p(RA)	0,861	p(dat)	0,558
p(VAL)*p(BA)	0,108	p(RÓL)*p(dat)	0,057	p(BÓL)*p(RA)	0,076	p(acc)*p(dat)	0,309
p(VAL&BA)	0,021	p(RÓL&dat)	0,023	p(BÓL&RA)	0,004	p(acc&dat)	0,213

4. táblázat. A *bevon*, *mesél*, *készül* és *hív* igék vonzatainak előfordulási gyakorisága, a gyakoriságok szorzata és a vonzatok együtt-előfordulási gyakoriságai.

Mindegyik vizsgált igénél az látszik, hogy a két argumentumtípus együtt-előfordulási gyakorisága kisebb a két argumentum előfordulási gyakoriságának a szorzatánál, azonban míg a *mesél* és a *hív* igék esetében a gyakoriságok szorzata csak kb. 1,5–2-szerese az együtt-előfordulási gyakoriságnak, a *bevon* és a *készül* igék esetében ez 5–20-szoros. A *bevon* és a *készül* igénél ez a várakozásainknak megfelelő, hiszen ott a két bővítmény különböző variánshoz tartozik. Magyarázatra szorul azonban, hogy az egyvariánsos *mesél* esetében miért kevesebb az együtt-előfordulási gyakoriság a vártnál, illetve hogy a kétvariánsos *hív* igénél miért csak ilyen kis mértékű csökkenés figyelhető meg.

Az emberi nyelvhasználat során nem egymástól független információkat közlünk a diskurzus folyamán, hanem egymásra épülő, egymást követő, egymással összefüggő információkat. Az összefüggést a diskurzus témájaként lehet azonosítani: ez sokszor egy személy vagy objektum, amiről új információkat közlünk, vagy egy esemény, amelynek új aspektusait adjuk meg.

A témaként szereplő személy/objektum általában az új információt kifejező ige alanyaként vagy tárgyaként jelenik meg, jellemzően a mondat topik pozíciójában, és gyakran el is hagyjuk, mivel a diskurzusuniverzumban könnyen elérhető, könnyen felidézhető. Ezzel szemben az új információ részét képező igei bővítmények sokkal többször jelennek meg az ige mellett a megnyilatkozásokban. A *hív* ige mindkét variánsánál van tárgyi vonzat, de az első variánsnál (*hívtam Pétert*) a tárgy általában az új információ része, míg a második variáns esetében (*Pétert öcskösnek hívtam*) a tárgy általában az ismert szereplők egyike. A datívuszi vonzatos *hív* ige mellett ennek megfelelően sokkal kisebb gyakorisággal jelenik meg a tárgyi bővítmény, mint az első variánsnál, mint az a 3. táblázatban is látható. A *hív* ige esetében a tárgynak a különböző variánsokban megfigyelhető nagyon eltérő gyakorisági értéke okozza, hogy összességében csak kicsivel kevesebb a tárgy és a datívusz együtt-előfordulási gyakorisága a két bővítmény gyakoriságának a szorzatánál, vagyis a várt nagyobb különbség elmaradásának pragmatikai-szemantikai okai vannak.

A *mesél* ige esetében viszont a RÓL és a datívuszi bővítmény ugyanannak az eseménynek két különböző aspektusát fejezi ki, és általában mindkettő az új információ részének tekinthető. Azonban az emberi nyelvhasználók nem mindig törekednek az események teljes leírására, hanem csak a legfontosabb aspektusokat közlik, így bár mindkét argumentumtípus vonzata az egyetlen variánsnak, az együttes előfordulásuk alacsonyabb lehet a vártnál.

Összegezve tehát elmondhatjuk, hogy a különböző argumentumtípusok egy ige melletti együtt-előfordulási gyakoriságának vártnál alacsonyabb értékéből következtethetünk arra, hogy azok különböző argumentumszerkezeti variánsokban jelennek meg, de a várthoz közelítő érték nem feltétlenül vezet ahhoz, hogy egy variánshoz tartozónak tekintsük őket.

5 Összegzés

Tanulmányunkban Szécsényi (2019) valószínűségi vektor alapú argumentumszerkezeti modelljét alapul véve vizsgáltuk meg az igék helyhatározói bővítményeinek eloszlási mintázatait.

A Szeged Korpusz összes igéjének valószínűségi vektorához viszonyítva az egyes igék argumentumszerkezeti vektorát, három mintázatot különítettünk el. Azokban az esetekben, amelyekben csak az egyik argumentumtípus megjelenési gyakorisága növekedett, az argumentumtípust az ige valódi vonzatának tekintettük – ebben az esetben az ige és a helyhatározó ragos bővítmény (az esetrag szempontjából) nem kompozicionális nem kompozicionáls összetétel. Azokban az esetekben, amelyekben több argumentumtípus gyakorisága is megnőtt, de az argumentumtípusok az irányhármasság szempontjából egy csoportba tartoznak, a megnövekedett gyakoriságú argumentumtípusokat egyetlen tematikus vonzat különböző megjelenési formáinak vettük. Végül azokban az esetekben, amikor egyetlen argumentumtípus gyakorisága sem növekedett a Szeged Korpuszban megfigyeltekhez viszonyítva, azt mondtuk, hogy ezeknek az igéknek nincsenek helyhatározói esetragos vonzataik, az ilyen bővítmény mindig szabad bővítmény.

Előfordul olyan eset is, amikor egy ige mellett egynél több (tematikusan nem összefüggő) argumentumtípus előfordulási gyakorisága is megnő a Szeged Korpuszhoz képest. Ekkor a két argumentumtípus lehet ugyanannak az igének két egymástól független vonzata, vagy egy ige két különböző (jelentésű) változatának az egyedi vonzata: ez utóbbi esetben az ige két argumentumszerkezeti variánsát feltételezzük. Ha a két argumentumtípus együttes előfordulásának a gyakorisága lényegesen kisebb, mint az argumentumtípusok előfordulási gyakoriságainak a szorzata, akkor az igének két argumentumszerkezeti variánsa van.

A tanulmányban egy-egy igét kiválasztva esettanulmányokat mutattunk be, amelyek feldolgozásánál egy rögzített feldolgozási láncot használtunk, azt adatokat pedig a MNSz-ból szereztük. Az eredmények jobb általánosíthatósága és további összefüggések megfogalmazhatósága érdekében szükséges a vizsgált igék számának nagyságrendekkel történő emelése.

Hivatkozások

Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. In: Computational Linguistics 28/3. pp. 245–288. (2002) doi:10.1162/089120102760275983

Gyulai, L.: Nem kompozicionális igekötős igék argumentumszerkezetének korpuszalapú vizsgálata. In: Ludányi, Zs., Gráczi, T. E. (szerk.) Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia. pp. 44–58. MTA Nyelvtudományi Intézet, Budapest (2019) doi:10.18135/Alknyelvdok.2019.13.4

Kálmán, L.: Miért nem vonzanak a régensek? In: Kálmán, L. (szerk.) KB 120. A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósi András tiszteletére. pp. 229–246. MTA Nyelvtudományi Intézet, Tinta Könyvkiadó, Budapest (2006)

Kálmán, L.: Bővítménykeretek mint konstrukciók. In: Kas, B. (szerk.) „Szavad ne feledd" Tanulmányok Bánréti Zoltán tiszteletére. pp. 61–72. MTA Nyelvtudományi Intézet, Budapest (2016)

Kovács, V., Simkó, K., Szécsényi, T.: Szabályalapú szintaktikai elemző szintaktikai szabályok nélkül. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). pp. 251–259. Szegedi Tudományegyetem, Szeged (2016)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1719–1723. European Language Resources Association (ELRA), Reykjavik (2014)

Palmer, M., Gildea, D., Kingsbury, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. In: Computational Linguistics 31/1. pp. 71–106. (2005) doi:10.1162/0891201053630264

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

Sass, B.: Az igei szerkezetek algebrai struktúrája, avagy a duplakocka modell. In: Argumentum 14. pp. 12–44. (2018)

Sass, B.: A duplakocka modell és az igei szerkezeteket kinyerő „ugrik és marad" módszer nyelvfüggetlensége, valamint néhány megjegyzés az UD annotáció univerzalitásáról. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 399–407. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2020)

Szécsényi, T.: Argumentumszerkezet-variánsok korpusz alapú meghatározása. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 315–329. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2019)

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. pp. 1855–1862. European Language Resources Association, Valletta, Málta (2010)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP 2013. pp. 763–771 (2013)