A topikalizálhatóságot befolyásoló tényezők statisztikai vizsgálata

Szécsényi Tibor – Kovács Viktória
SZTE Általános Nyelvészeti Tanszék; SZTE Nyelvtudományi Doktori Iskola
szecsenyi@hung.u-szeged.hu; viktoria.kovacs12@gmail.com

Kivonat: É. Kiss Katalin írásaiban több helyen is megemlíti, hogy az ige bővítményeinek a topikalizálhatóságát, a topikalizálással kapott mondat semlegességét befolyásolják a bővítmény egyéb tulajdonságai is: a bővítmény esete, határozottsága, szemantikai jegyei és thematikus szerepe (É. Kiss 1987; 1992; 2002). Tanulmányunkban azt vizsgáljuk meg, hogy ez az intuitív megállapítás alátámasztható-e a tényleges nyelvhasználati adatokkal. Korpuszból származó egyszerű, semleges mondatokban annotáltuk a ténylegesen topik pozícióban levő és a potenciálisan topikalizálható kifejezéseknek ezen tulajdonságait, és statisztikai elemzéssel bemutatjuk, hogy ezek a tényezők valóban összefüggést mutatnak azzal, hogy egy összetevő topik pozícióba kerül-e.

Kulcsszavak: topik; thematikus szerep; határozottság; korpusz

1. Alapszórend, semleges szórend, topikalizálhatóság

É. Kiss Katalin a *Configurationality in Hungarian*-ben azzal a feltevéssel szemben hoz példákat, hogy a magyar nyelv SVO alapszórendű lenne (É. Kiss 1987, 24–25). Ezek a mondatok az anyanyelvi beszélők ítéletére alapozva semlegesek, azonban bennük az ige előtt nem, vagy nem csak alanyi összetevőt találunk, többnyire topik pozícióban. A példák mellett a nem alanyi topikalizált elemek egy-egy releváns jellemzője került kiemelésre, sugallva azt, hogy amennyiben ezek a tényezők megtalálhatók egy egyébként topikalizálható összetevőn, akkor semleges mondatok esetében azok topik pozícióba kerülhetnek. A kiemelt tényezők a következők (É. Kiss számozását átvéve): határozott kifejezés (15), tulajdonnév (16), önálló referenciájú elem, összehasonlítva a tőle függő referenciájú alannyal (ami tipikusan birtokos kifejezés) (17), [+human] jegyű kifejezés (18), experiens thematikus szerepű elem, összehasonlítva a téma alannyal (19)–(20), zéró általános alany mellett megjelenő tárgy (21), és végül a birtokos szerkezetből kimozgatott datívuszi birtokos (22).

(15) Az igazgatót felhívta egy újságíró.

(16) Jánost megbüntette a rendőr.

(17) Az igazgatót figyelmeztette a titkárnője.
(18) Jánost elütötte a vonat.
(19) Jánosnak tetszik Éva.
(20) Jánost idegesíti a zaj.
(21) Jánost keresték.
(22) Jánosnak gyűrött volt a ruhája.

Ezekben a példákban, az alany kitüntetett pozícióját vitatandó, a topikalizált elemek mindig az alannyal lettek szembeállítva.

A *Strukturális magyar nyelvtan Mondattan* kötetében az É. Kiss által írt fejezetben a hasonlóan felsorakoztatott példák elővezetéseként É. Kiss erősebb állítást fogalmaz meg (1992, 106): „Bár az ige bármely argumentuma szabadon topikalizálható, a topik kiválasztása nem minden esetben tűnik egyformán jelöletlennek. A topik kiválasztását az alábbi tényezők látszanak befolyásolni: I. az összetevők [±humán], [±élő] volta; II. az összetevők specifikussága; III. az összetevők thematikus szerepe; IV. a szövegkörnyezet." Itt nem egyszerűen csak azt állítja, hogy az említett jegyekkel rendelkező kifejezések is lehetnek semleges mondat topikjai, hanem hogy ezek jelöletlen, vagyis semleges mondatban inkább kerülnek topik pozícióba, szemben az ilyen tulajdonsággal nem rendelkező alanyi vagy más összetevőkkel.

A 2002-es *The Syntax of Hungarian*-ban, bár az összes korábban említett, topikalizálhatóságot befolyásoló tényezőt nem sorolja föl, É. Kiss meg is magyarázza, hogy miért ezek a tényezők befolyásolják a topik pozícióba kerülést: az emberi gondolkodásban az emberek kerülnek a középpontba, ezért leginkább [+humán] jeggyel rendelkező kifejezések a mondat alanyai, topikjai.[1]

É. Kiss Katalin tehát a magyar mondatszerkezetet leíró munkássága során többször is megemlít olyan grammatikai tényezőket, amelyek befolyással bírnak arra, hogy egy-egy összetevő végül topikalizálódik-e vagy sem. A topikalizálhatósági vagy topikalizálódási „hajlam" úgy értelmeződik, hogy a semleges szórendű (és semleges intonációjú) mondatokban ezek az összetevők inkább szeretnek topik pozícióban lenni. Nem világos azonban sem az, hogy mit kell „semleges szórenden" érteni, sem az, hogy hogyan értelmezhető, mérhető ez

[1] „Subjects are in fact more frequent topics than objects – but the link between subjecthood and topichood is only indirect. We tend to describe events from a human perspective, as statements about their human participants – and subjects are more often [+human] than objects are. In the case of verbs with a [–human] subject and a [+human] accusative or oblique complement, the most common permutation is that in which the accusative or oblique complement occupies the topic position." (É. Kiss 2002, 9)

a „hajlandóság". Ha a semleges szórendet úgy értjük, hogy a kontextusából kiragadott mondat esetében ez a szórendi variáns kapná a legmagasabb grammatikalitási ítéletet, akkor a topikalizálhatósági hajlamot nem lehet tényleges nyelvhasználati gyakorisági adatként értelmezni, hiszen a nyelvhasználat során mindig adott a nyelvi és nem nyelvi kontextus. Tényleges nyelvhasználat során viszont nem az említett tényezők határozzák meg, hogy egy összetevő topik pozícióban jelenik-e meg végül, hanem csakis a beszélői szándék: amiről a beszélő információt szeretne közölni, az kerül topik pozícióba, függetlenül attól, hogy az milyen morfológiai, szintaktikai és szemantikai jegyekkel rendelkezik, milyen thematikus szerepű.

Tanulmányunkban azt szeretnénk megvizsgálni, hogy egy egyszerűen definiált semleges szórendet figyelembe véve az É. Kiss által is említett grammatikai tényezők összefüggésbe hozhatóak-e a korpuszadatok által leírt nyelvhasználat során azzal, hogy a kifejezések végül topik pozícióban foglalnak-e helyet, és ha igen, milyen jellegű ez az összefüggés.

2. A topikalizálhatóságot befolyásoló tényezők

A vizsgálat során ötféle grammatikai tényezőt vettünk figyelembe: eset *(case)*, szintaktikai osztály *(np_type)*, szemantikai osztály *(sem_type)*, thematikus szerep *(theta)* és hossz *(length)*.

Ezek közül az első és egyben a legfontosabb a topikalizálható kifejezések esete, más grammatikai tényezők hatását ugyanis mindig így értelmezzük: „X hiába nem alanyesetű, ha rendelkezik Y tulajdonsággal, mégis gyakrabban topikalizáljuk". Megvizsgáltuk tehát, hogy van-e összefüggés a topikalizálható kifejezések esete és topikalizálódás között. A topikalizálható elemek esetei közül figyelembe vettük a szerkezeti eseteket: *nominative*, *accusative*, *dative*; az *instrumentalis* esetet; viszont egységesen kezeltük az inherens eseteket: *other*. Erre azért volt szükség, mert ezek az esetek nagyon kis számban voltak jelen a korpuszban, és így biztosíthattuk a statisztikai kiértékelhetőséget. Egy további kategóriát is vizsgáltunk, a névutós kifejezések esetében *pp* címkét rendeltünk a topikalizálható elemhez. A vizsgálat során így 6 esetet különböztettünk meg, azokat, amelyek a vizsgált adatok között megfelelő számban fordultak elő.

A topikalizálható kifejezések szintaktikai osztályozása során használt kategóriák: határozott kifejezés *(def)*, határozatlan kifejezés *(indef)*,[2] tulajdonnév

[2] A határozottság és a határozatlanság a vizsgálat során pusztán szintaktikai tulajdonságként szerepelt, vagyis hogy a kérdéses főnévi csoport határozott vagy határozatlan névelővel szerepelt-e.

(name), (személyes és visszaható) névmás *(pron)* és demonstratívum *(dem)*. A tulajdonneveken, névmásokon és demonstratívumokon kívüli főnévi kifejezéseket a névelőjük alapján osztályoztuk, a névutós kifejezéseket a névutóval összekapcsolt főnévi kifejezés alapján, a határozókat pedig a jelentésük határozottsága alapján (pl. *tegnap* – határozott, *egyszer* – határozatlan).

A szemantikai osztályozásnál három kategóriát használtunk: élő *(live)*, élettelen dolog *(thing)* és egyéb vagy absztrakt *(other)*. Az élő osztályt tovább lehetett volna bontani emberre, állatra és növényre, de a vizsgált korpuszban az utóbbi két kategóriára nagyon kevés példa volt. Az épületek, helyek élettelennek lettek besorolva, az intézmények, időpontok és egyéb absztrakt kifejezések egyébnek.

A negyedik megvizsgálandó tényező a topikalizálható kifejezéseknél a thematikus szerep. Egy kifejezés thematikus szerepe nehezen meghatározható, jól használható teszt talán csak az ágens thematikus szerepre létezik (l. Kenesei 2000; 2001). Ezért nem is használtuk az összes theta-szerepet, hanem csak négyet: *agent*, *patient*, *experient* és *theme*, ahol az utolsó valójában azokat a szerepeket fedi le, amelyek nem férnek be az első háromba. Ezeken kívül figyelembe vettünk még két további kategóriát is a gyakran előforduló mondathatározókra: *time* és *place*. Ezeket a címkéket azok a kifejezések kaphatták, amelyek nem közvetlenül az ige argumentumai, hanem szabad bővítményként jelentek meg a mondatban: a *Szabó Magda Debrecenben született* mondatban *theme*, a *Debrecenben kiolvastam az Abigélt* mondatban pedig *place* címkét kap a *Debrecenben*.

Az utolsó figyelembe vett tényező a topikalizálható kifejezések hossza, vagyis az őt alkotó szavak száma volt. Ez a tulajdonság nem szerepelt az É. Kiss által felsoroltak között, mivel azonban a topikalizált kifejezések szerepe az, hogy a diskurzusuniverzumban szereplő, vagyis egy korábban már valószínűleg bevezetett referenst ismét a középpontba állítson, érdemes megvizsgálni azt is, hogy a visszautalást elősegítő jegyek befolyásolják-e a topikalizálhatóságot. Ariel elérhetőségi elmélete szerint (Ariel 1990) a visszautalást, illetve a referens felidézhetőségét több tényező befolyásolja, például a kifejezés rigidsége és a kifejezés hossza.[3] A rigidséget a főnévi kifejezések szintaktikai osztályozásával már figyelembe vettük, az elérhetőségi skála egyik végén a nevek állnak, a másikon a névmások (l. még Kovács 2019), a másik tényezőt hivatott képviselni ez a tulajdonság. A topikalizálható kifejezéseket hosszuk szerint öt csoportba soroltuk, egytől négy szó hosszig, illetve az ennél hosszabbak külön osztályba kerültek *(more)*.

[3] További tényező még a kifejezés informativitása, de ezt az általunk használt formális eszközökkel nehéz ellenőrizni, ezért csak az említett két tényezőt vettük figyelembe.

A topikalizálhatóságot befolyásoló tényezőket és azok osztályait az 1. táblázat összegzi.

1. táblázat: A topikalizálhatóságot befolyásoló tényezők és azok osztályai

Tényezők	Osztályok	Osztályok száma
case	nominative, accusative, dative, instrumental, pp, other	6
np_type	def, indef, name, pron, dem	5
sem_type	live, thing, other	3
theta	agent, patient, experient, theme, time, place	6
length	1, 2, 3, 4, more	5

Összesen így elméletileg 6 × 5 × 3 × 6 × 5 = 2700 különböző osztálykombinációt lehet elkülöníteni, amik lehetnek topikalizálva vagy nem topikalizálva, a tényleges adatok között azonban ennél jóval kevesebb kombináció figyelhető meg: nehéz elképzelni például olyan nyelvi kifejezést, ami alanyesetű élő helyhatározó lenne.

3. A korpusz bemutatása és feldolgozása

A topikalizálhatóságot befolyásoló tényezők hatásainak vizsgálatához a tényleges nyelvhasználatot tükröző nyelvi adatokra van szükség. A vizsgálatunk során a Szeged Dependency Treebank (Vincze et al. 2010) adataiból válogattunk olyan mondatokat, amelyekről elmondható egyrészt az, hogy a hétköznapi nyelvhasználatot tükrözik, másrészt pedig az, hogy valamilyen értelemben semlegesnek tekinthetőek.

A hétköznapi nyelvhasználat mint vizsgálati szempont kizárta a korpusz jogi, gazdasági és publicisztikai szövegeit (bonyolult mondatszerkezetek, speciális konstrukciók használata), megfeleltek viszont a kritériumnak a szépirodalmi és az iskolai fogalmazás részkorpuszok. A kézi annotálás miatt nem a teljes szépirodalmi és iskolai fogalmazás anyagot dolgoztuk fel, hanem csak a fogalmazások egy részét – az összefüggések pontosabb feltárásához a vizsgálatot a későbbiekben a maradék részkorpuszok adataira is ki kell terjeszteni.

A kívánt semleges szórendet a legjobban a semleges, fókuszt nem tartalmazó kijelentő mondatok kiválasztásával láttuk biztosítottnak. Mivel a mondatok írott formája miatt sokszor nem egyértelmű, hogy a mondat egy összetevője

fókuszpozícióban van-e, csak az egyértelműen eldönthető eseteket vettük figyelembe. Ezt úgy értük el, hogy a teljes korpuszból csak azokat a mondatokat gyűjtöttük ki, amelyekben ragozott igekötős ige szerepel, mégpedig úgy, hogy az igekötő az igével egybe van írva. A vizsgálat során nem vettük figyelembe a más konstrukciókban megjelenő topikalizációt, például a főnévi igeneves kifejezéseket (É. Kiss 1989; Szécsényi 2009). További feltételnek tekintettük, hogy a mondatokban legyen legalább két topikalizálható kifejezés, egy topikalizált és egy nem topikalizált is. Ennek érdekében a korpuszadatok kigyűjtése során azt a feltételt szabtuk, hogy a megvizsgálandó mondatokban az ige előtt és az ige mögött is legyen egy összetevő. A kigyűjtött korpusz így 515 mondatot tartalmazott.

Az így kiválogatott semleges mondatok természetesen nem fedik le a semleges mondatok teljes spektrumát, így nem is lehet teljesen reprezentatív mintának tekinteni: hiányoznak azok a mondatok, amelyekben az ige igekötő nélkül szerepel, vagy nem igekötői igemódosítóval, valamint azok a semleges mondatok, amelyek nem befejezett aspektusúak, hanem például progresszívek (ahol is az igekötő az ige mögé kerül, pl.: *Péter (éppen) mászott fel a fára...*), mint ahogy azok a mondatok is, ahol nincs egyszerre jelen ige előtti és ige mögötti összetevő is a mondatban. Jelen kutatásban azonban azzal a hallgatólagos feltételezéssel élünk, hogy ezek a tényezők nem befolyásolják a topikalizálhatóságot, mint ahogy É. Kiss sem említi ezeket a tulajdonságokat. Ha azonban mégis befolyásoló tényezők lennének, akkor a pontos leíráshoz minden esetben meg kellene határozni a mondatok aspektusát is, és még megannyi, most nem is említett tényezőt, mint például hogy a kérdéses mondat főmondat-e vagy alárendelt mondat, milyen modalitással rendelkezik stb. A kutatás jelenlegi formájában azonban nem tűzte ki célul topikalizálhatóságot befolyásoló összes tényező teljes leírását, csupán azt vizsgáltuk, hogy az 1. táblázatban felsorolt tulajdonságok hatással vannak-e a topikalizálhatóságra.

A vizsgálati anyagban a kigyűjtött mondatok szövegkörnyezet nélkül szerepeltek, így óhatatlanul bekerülhettek olyan adatok is, ahol az ige előtti összetevő nem topik pozícióban volt, hanem kontrasztív topik pozícióban – az ilyen mondatok azonban nem tekinthetőek semleges intonációjú mondatoknak. Azonban ez a kapott adatokat vélhetően nem befolyásolta nagy mértékben, mert (1) a vizsgált iskolai fogalmazásokban eleve is aránylag kevés a kontrasztív topikos mondat; (2) a kézi annotálás során az egyértelműen kontrasztív topikos mondatokat (mint például a *Péterrel, vele összefutottam*) kizártuk a vizsgálatból; (3) a maradék, nem egyértelmű mondatok esetében bár lehet, hogy kontrasztív

topik pozícióban kellett volna értelmezni a kontextus alapján, de a kontextus hiányában lehetett egyszerű topiknak is tekinteni a kifejezéseket.

A Szeged Dependency Treebank kézzel annotált morfológiai és szintaktikai információkat is tartalmaz, ezen információk felhasználásával nemcsak a semleges szórendű mondatokat tudtuk automatikusan kigyűjteni és bennük a megfelelő igéket beazonosítani, hanem az igekötős igéhez tartozó bővítményeket is, előannotálva azok esetét, bizonyos szintaktikai típusát *(name, pron, dem)*, illetve a bővítmények szóhosszát is, valamint azt, hogy az adott bővítmény a mondatban az ige előtt (topik) vagy után található-e (nem topik). A korpusz mondatainak és előelemzésének a menete a Gyulai (2019)-ben ismertetett módon történt.

A kézi annotálás MMAX2 szoftver segítségével történt (Müller–Strube 2006), egyetlen annotátor közreműködésével. Az annotálás során egyrészt ellenőriztük, hogy a Szeged Dependency Treebankból automatikusan kigyűjtött információk helytállóak-e (semleges mondatról van-e szó, a megtalált bővítmények tényleg topikalizálhatóak-e, mindegyik ilyen elem meg lett-e találva, megfelelő *case, np_type, length* osztályozást kaptak-e), másrészt a hiányzó információkat is rögzítettük: *sem_type, theta.*

A kézi annotálás után végül összesen 424 semleges szórendű mondat maradt, mivel az eredetileg kigyűjtött 515 mondatból kizártuk azokat, amelyekben az ige előtti vagy az igét követő összetevő nem topikalizálható kifejezés volt, hanem például univerzális kvantor (pl. *Minden fiú meglátogatta Marit*). A 424 mondat összesen 931 topikalizálható összetevőt tartalmazott. Ebből 457 volt ténylegesen topik pozícióban, 474 pedig ige utáni pozícióban maradt.

Az 515 mondatból álló korpusz kézi annotált változata és az abból kinyert táblázatok a https://github.com/szecsenyi/topicalization repozitóriumban hozzáférhetőek.

4. A topikalizálhatóságot befolyásoló tényezők statisztikai elemzése

A korpuszból kinyert adatok tanulmányozása során kiderült, hogy a *place* és *time* thematikus címkéjű kifejezések kiugróan nagy számban kerültek topik pozícióba: a helyhatározók 78%-a, az időhatározóknak pedig 98%-a topikalizálódott. Az ilyen kifejezések más szempontokból is különböztek a többi vizsgált bővítménytől: (1) a hely- és időhatározók szabad bővítmények, míg a többi általában vonzata a mondat igéjének, (2) szinte kizárólag inherens esetben álltak

(other), (3) *def* és *indef* szintaktikai osztályúak voltak, (4) szemantikai típusuk pedig *thing* és *other* volt. Hogy a hely- és időhatározók kiértékelést torzító hatását kiküszöböljük, végül a statisztikai elemzésnél nem vettük őket figyelembe. Így összesen 723 topikalizálható kifejezés maradt, amelyből 262 volt topik pozícióban, 461 pedig ige utáni pozícióban.

A topikalizálható kifejezések osztályeloszlását a 2. táblázat mutatja be, külön feltüntetve a hely- és időhatározókkal együtt kapott adatokat és a nélkülük összeszámoltakat is.

Ahhoz, hogy megvizsgáljuk, van-e összefüggés a fent említett csoportok között, χ^2 statisztikai próbát végeztünk rajtuk az SPSS segítségével. Egyesével vizsgáltuk meg az egyes csoportok hatását a topikalizálásra, tehát az volt a függő változó, hogy az adott kifejezés topik-e vagy sem, az összes többi a független.

- A *case* és a topik összehasonlítása során a szabadságfok 5 volt, $\chi^2 = 363{,}466$, $p < 0{,}001$.
- Az *NP_type* és a topik összehasonlítása során a szabadságfok 4 volt, $\chi^2 = 54{,}224$, $p < 0{,}001$.
- A *sem_type* és a topik összehasonlítása során a szabadságfok 2 volt, $\chi^2 = 216{,}924$, $p < 0{,}001$.
- A *theta* és a topik összehasonlítása során a szabadságfok 3 volt, $\chi^2 = 294{,}961$, $p < 0{,}001$.
- A *length* és a topik összehasonlítása során a szabadságfok 4 volt, $\chi^2 = 44{,}964$, $p < 0{,}001$.

Minden egyes teszt szignifikáns eredményt hozott, ami azt jelenti, hogy a vizsgált tényezők befolyásolják annak valószínűségét, hogy a kifejezés topik lesz-e vagy sem. Azért, hogy megvizsgáljuk, mennyire szoros az összefüggés az egyes csoportok és a topik között, Cramer's V tesztet végeztünk szintén az SPSS segítségével. A teszt eredménye egy 0 és 1 közötti mérőszám, ahol a 0 a leggyengébb, az 1 pedig a legerősebb kapcsolat a két változó között.

- A *case* és a topik viszonylatában Cramer's V = 0,709, tehát szoros az összefüggés.
- A *theta* és a topik esetében Cramer's V = 0,639, tehát szoros, de nem olyan szoros, mint az eset tekintetében.
- A *sem_type* és a topik összehasonlításában Cramer's V = 0,548, ami közepes összefüggést jelent.
- Az *NP_type* és a topik esetében a Cramer's V teszt eredménye Cramer's V = 0,274, tehát gyenge összefüggés mutatható ki.
- A *length* és a topik esetében Cramer's V = 0,249, tehát a vizsgált csoportok közül ez a leggyengébb összefüggést mutató két csoport.

A vizsgált 424 mondatban a 723 topikalizálható kifejezés közül 262 volt topik pozícióban, ez az összes topikalizálható kifejezés 36,24%-a.

2. táblázat: A topikalizálhatóságot befolyásoló osztályok eloszlása

Tényező	Osztály	Time és place osztályokkal			Time és place osztályok nélkül		
		topik	nem topik	összesen	topik	nem topik	összesen
case	nominative	186	46	232	186	46	232
	accusative	13	148	161	13	148	161
	pp	101	11	112	5	9	14
	instrumental	42	21	63	38	21	59
	dative	4	7	11	4	7	11
	other	111	241	352	16	230	246
np_type	def	299	348	647	164	337	501
	indef	67	51	118	18	50	68
	name	32	53	85	21	52	73
	pron	50	19	69	50	19	69
	dem	9	3	12	9	3	12
sem_type	live	192	85	277	192	85	277
	thing	69	291	360	42	285	327
	other	196	98	294	28	91	119
theta	theme	70	366	436	70	366	436
	agent	161	21	182	161	21	182
	time	162	4	166			
	patient	20	56	76	20	56	76
	place	33	9	42			
	experient	11	18	29	11	18	29
length	1	99	62	161	86	61	147
	2	181	280	461	107	271	378
	3	101	80	181	36	78	114
	4	54	23	77	19	22	41
	more	22	29	51	14	29	43

A topikalizálható kifejezések esetét figyelembe véve az alanyesetű kifejezések kimagasló arányban kerültek topik pozícióba (80,17%), amit az *instrumental* esetcímkéjű elemek követnek (64,41%), a datívuszi és a névutós kifejezések az

átlagnak megfelelő topikalizálhatóságot mutatnak (36,36% és 35,72%), a tárgyesetű és az egyéb esetű kifejezések nagyon kis topikalizálhatósági értéket mutatnak (8,08% és 6,05%). Az alanyeset nélküli topikalizálhatósági ráta 15,48%.

Thematikus szerepe alapján az *agent* topikalizálható legkönnyebben (88,46%), amit az *experient* (37,93%) és a *patient* (26,32%) követ, a legkevésbé a *theme*/egyéb thematikus szerepű kifejezések topikalizálhatóak (16,16%).

Szemantikai jegyeiket tekintve az élő jeggyel rendelkező kifejezések topikalizálhatóak leginkább (*live* 69,31%), majd az absztrakt jelentésű kifejezések (*other* 23,53%), végül pedig az élettelen dolgokra referáló kifejezések (*thing* 12,84%).

Szintaktikai felépítésük és kategóriájuk alapján a névmások (*pron* 72,46%) és a demonstratívumok (*dem* 75,00%) kerültek leginkább topik pozícióba, a határozott névelős kifejezések (*def* 32,73%), a nevek (*name* 28,77%) és a határozatlan kifejezések (*indef* 26,47%) mind átlag körüli, de azt el nem érő topikalizációs rátát mutatnak.

A topikalizálható kifejezések hossza alapján csak az egy- (58,50%) és a négyszavas (46,34%) kifejezések topikalizálódnak az átlagtól lényegesen eltérő mértékben.

5. Összefoglalás

Tanulmányunkban egy korpuszadatokon nyugvó statisztikai vizsgálatot mutattunk be, amely É. Kiss Katalinnak azt a megállapítását igyekezett ellenőrizni, hogy bár a magyar mondatokban az ige mellett megjelenő bővítmények bármelyike topikalizálható bizonyos feltételek teljesülése mellett, a bővítmények bizonyos szintaktikai, szemantikai tulajdonságai összefüggésbe hozhatók-e azzal, hogy mennyire lesz semleges az adott bővítmény topkalizálásával kapott mondat. Vizsgálatunk megállapította, hogy a topikalizálhatóságra legnagyobb hatással a kifejezés esete van, ezt követi a bővítmény thematikus szerepe, szemantikai jegyei és a szintaktikai felépítése és osztálya. Ezeket a topikalizálhatóságra ható tényezőket egymástól függetlenül vizsgáltuk. Az, hogy ezek a tényezők milyen együttes hatást fejtenek ki, további vizsgálatokat igényel.

A statisztikai vizsgálat tehát alátámasztotta É. Kiss Katalinnak azt az állítását, hogy az ige bővítményeinek a topikalizálhatóságát, a topikalizálással kapott mondat semlegességét befolyásolják a bővítmény egyéb tulajdonságai is.

Irodalom

Ariel, Mira 1990. Accessing noun-phrase antecedents. London & New York: Routledge.

É. Kiss, Katalin 1987. Configurationality in Hungarian. Budapest: Akadémiai Kiadó.

É. Kiss Katalin 1989. Egy főnévi igeneves szerkezetről. In: Telegdi Zsigmond – Kiefer Ferenc (szerk.): Általános Nyelvészeti Tanulmányok XVII. Tanulmányok a magyar mondattan köréből. Budapest: Akadémiai Kiadó. 153–169.

É. Kiss Katalin 1992. Az egyszerű mondat szerkezete. In: Kiefer Ferenc (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Budapest: Akadémiai Kiadó. 79–177.

É. Kiss, Katalin 2002. The syntax of Hungarian. Cambridge: Cambridge University Press.

Gyulai Lívia 2019. Nem kompozicionális igekötős igék argumentumszerkezetének korpuszalapú vizsgálata. In: Ludányi–Gráczi (2019, 45–60).

Kenesei István 2000. Van-e segédige a magyarban? Esettanulmány a grammatikai kategória és vonzat fogalmáról. In: Kenesei István (szerk.): Igei vonzatszerkezetek a magyarban. Budapest: Osiris Kiadó. 157–196.

Kenesei, István 2001. Criteria for auxiliaries in Hungarian. In: István Kenesei (szerk.): Argument structure in Hungarian. Budapest: Akadémiai Kiadó. 73–106.

Kovács Viktória 2019. Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In: Ludányi–Gráczi (2019, 113–121).

Ludányi Zsófia – Gráczi Tekla Etelka (szerk.) 2019. Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből. Budapest: MTA Nyelvtudományi Intézet.

Müller, Christoph – Michael Strube 2006. Multi-level annotation of linguistic data with MMAX2. In: Sabine Braun – Kurt Kohn – Joybrato Mukherjee (szerk.): Corpus technology and language pedagogy: New resources, new tools, new methods. Frankfurt am Main: Peter Lang. 197–214.

Szécsényi, Krisztina 2009. On the double nature of Hungarian infinitival constructions. Lingua 119: 592–624.

Vincze, Veronika – Dóra Szauter – Attila Almási – György Móra – Zoltán Alexin – János Csirik 2010. Hungarian Dependency Treebank. Proceedings of the Seventh Conference on International Language Resources and Evaluation. Valletta, Malta: European Language Resources Association. 1855–1862.

A statistical analysis of topicalization factors

Abstract: Katalin É. Kiss mentions in several places that the ability of a constituent to be topicalized and the neutrality of the sentence obtalned by topicalization are influenced by other features of the constituent: case, definiteness, semantic features and thematic role (É. Kiss 1987; 1992; 2002). In this paper, we examine whether this intuitive statement can be supported by actual language use. In simple, neutral sentences obtained from corpus we annotated these properties of topicalized and potentially topicalizable constituents, and we show by statistical analysis that these factors really correlate with whether a component is placed in the topic position.

Keywords: topic; theta roles; definiteness; corpora