Argumentumszerkezet-variánsok korpusz alapú meghatározása

Szécsényi Tibor

Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék,
6722, Szeged, Egyetem u. 2.
szecsenyi@hung.u-szeged.hu

Kivonat: A tanulmány a lexikai egységek, tipikusan igék argumentumszerkezetének a leírására javasol egy új reprezentációs formát, ami nem a klasszikus kötelező vonzat – szabad bővítmény bináris oppozíciós lehetőségeket ragadja meg. Ehelyett az egyes bővítménytípusoknak a korpuszban való megjelenési gyakoriságai alapján a típusokhoz egy-egy valószínűségi értéket rendel, így az argumentumszerkezeti variánsok egy argumentumszerkezeti valószínűségi vektorral jellemezhetőek. A javasolt reprezentáció kizárólag a korpuszbeli adatok morfológiai és szintaktikai tulajdonságaira támaszkodik. Az argumentumszerkezeti variánsok argumentumszerkezeti vektorként való értelmezése új elméleti modellként a grammatikaelméletekben hozhat új eredményeket, másrészt a természetesnyelv-feldolgozásban is használható.

1 Bevezetés, célok

A teljes szintaktikai elemzés elengedhetetlen feltétele a szövegben található igék és más régensek argumentumszerkezetének valamilyen szintű ismerete, ez teszi lehetővé, hogy a mondatban a régens mellett kötelezően megjelenítendő kifejezések számát és azok tulajdonságait leírhassuk. Az argumentumszerkezet az igék/régensek egyedi, lexikai tulajdonsága, ami tulajdonságokat a nyelv nyelvelméleti igényű grammatikái explicit módon használnak fel a mondatszerkezet kialakítása során: a transzformációs nyelvtanokban például a projekciós elv [1], a HPSG-ben az alkategorizációs elv [2, 3] biztosítja, hogy az egyes igék/régensek mellett csak a megfelelő összetevők jelenhessenek meg vonzatpozícióban. Az ezen elméleteket alkalmazó számítógépes szintaktikai elemzők is hatékonyan használják a lexikai elemek argumentumszerkezeti információit, például a [4] tanulmányban bemutatott szabály alapú elemző az igék argumentumszerkezetének lexikai leírására támaszkodva csupán négy újraíró szabállyal képes elemezni a magyar mondatokat.

A leíró nyelvészeti munkák az argumentumszerkezetet az argumentumok, argumentumtípusok felsorolásával adja meg. Jelen dolgozatban egy olyan argumentumszerkezet-reprezentációra teszek javaslatot, amely nem ilyen bináris, argumentum – nem argumentum oppozícióként kezeli az argumentumszerkezetet, hanem a régensek lehetséges bővítményeihez egy-egy [0–1] intervallumon található értéket rendel, jelezve ezzel, hogy az adott bővítmény mekkora valószínűséggel jelenik meg a régens mellett egy mondatban. Ekkor a régens argumentumszerkezete *n* lehetséges bővítménytípus esetén

egy *n*-dimenziós egységkocka belsejébe mutató vektorral jellemezhető. Ez az argumentumszerkezeti vektor korpusz alapján automatikusan is meghatározható, illetve a bővítmények megjelenését befolyásoló tényezők feltérképezése után azok figyelembevételével a klasszikus argumentumszerkezeti lista is visszanyerhető.

A javasolt reprezentáció nem a szóbeágyazási modellek [5] egy változata, hanem a többdimenziós értelmezés miatt inkább Sass Bálint duplakocka-modelljével [6] rokonítható. Korábban Kálmán László is javasolta a vonzatság bináris felfogásának az elvetését [7, 8], de nála ez egyrészt vagy csak az ige és bővítmény közötti többféle lehetséges viszonyt jelentette, vagy ha a kapcsolatuk erősségének a gradualitását is megemlítette, ennek a gradualitásnak az értékét nem a bővítmények megjelenési valószínűségéhez kötötte. További különbség, hogy Kálmán az egyes ige-bővítmény kapcsolatokat egyenként elemezte, nem az ige teljes argumentumszerkezetét próbálta meg így leírni.

A dolgozat fő újdonságaként az első részben először bevezetem az argumentumszerkezeti vektorok fogalmát (2. szakasz), majd bemutatom, hogyan lehet egy ige argumentumszerkezeti variánsait korpusz alapján meghatározni (3. szakasz). Ezután az argumentumszerkezeti vektorokat befolyásoló néhány tényezőt mutatok be (4. szakasz). A dolgozat végén néhány lehetséges felhasználási területet is ismertetek.

2 Az argumentumszerkezetek megfigyelése természetes környezetükben

Az igék és más régensek argumentumszerkezetét hagyományosan egy táblázatban adhatjuk meg, ahol minden egyes ige minden vonzatszerkezeti variánsához egy sor tartozik, ahol jelöljük, hogy milyen tulajdonságú vonzatokkal kell együtt szerepelnie egy teljes mondatban. A Szeged Korpuszban [9, 10] a *bíz/bízik* ige 6 variánsával található meg, ezek az 1. táblázatban láthatóak (a főnévi vonzatokat az esetükkel jellemzem). A hat argumentumszerkezeti variánsra példák: *Péter bízik*$_1$ *Mariban*; *Péter megbízik*$_2$ *Mariban*; *Péter Marira bízta*$_3$ *a könyvet*; *Péter rábízta*$_4$ *a könyvet Marira*; *Péter megbízta*$_5$ *Marit a feladattal*; *Péter elbízta*$_6$ *magát*, bár itt nagyon speciális tárgyról beszélünk, csakis visszaható névmási tárgy lehet.

	igekötő	NOM	ACC	BAN	RA	VAL
bíz/bízik$_1$	-	+	-	+	-	-
bíz/bízik$_2$	meg	+	-	+	-	-
bíz/bízik$_3$	-	+	+	-	+	-
bíz/bízik$_4$	rá	+	+	-	+	-
bíz/bízik$_5$	meg	+	+	-	-	+
bíz/bízik$_6$	el	+	+	-	-	-

1. táblázat. a *bízik* ige argumentumszerkezeti variánsai

A Szeged Korpuszban ez a hat argumentumszerkezeti variáns nem mindig a vonzataival együtt jelenik meg, továbbá nem csak a vonzatai találhatóak mellette, hanem más szabad bővítmények is. A dependenciakorpuszból [10] saját korpusztranszformációval és kézi annotálással (MMAX2 [11]) a 2. táblázatban látható vonzat-előfordulási adatokat kapjuk.

			Argumentumtípus (X)				
			NOM	ACC	BAN	RA	VAL
bíz/bízik	n_V=157	n_X	64	66	91	44	23
		f_X	0,41	0,42	0,58	0,28	0,15
bíz/bízik$_1$	n_{V1}=66	n_{1X}	29	0	65	2	0
	f_{V1}=0,42	f_{1X}	0,44	0	0,98	0,03	0
bíz/bízik$_2$	n_{V2}=17	n_{2X}	9	0	17	0	0
	f_{V2}=0,11	f_{2X}	0,53	0	1,00	0	0
bíz/bízik$_3$	n_{V3}=37	n_{3X}	9	33	4	37	0
	f_{V3}=0,24	f_{3X}	0,24	0,89	0,11	1,00	0
bíz/bízik$_4$	n_{V4}=6	n_{4X}	2	4	1	3	0
	f_{V4}=0,04	f_{4X}	0,33	0,67	0,17	0,5	0
bíz/bízik$_5$	n_{V5}=28	n_{5X}	14	26	4	2	23
	f_{V5}=0,18	f_{5X}	0,5	0,93	0,14	0,07	0,82
bíz/bízik$_6$	n_{V6}=3	n_{6X}	1	3	0	0	0
	f_{V6}=0,02	f_{6X}	0,33	1,00	0	0	0

2. táblázat. a *bízik* ige bővítményeinek előfordulási száma és megjelenési gyakorisága a Szeged Korpuszban

A korpuszban összesen 157-szer szerepel az ige (n_V), ebből 64-szer szerepel vele egy tagmondatban alanyesetű maximális főnévi csoport (n_{NOM}), ami 0,41 relatív gyakoriságot jelent ($f_{NOM} = \frac{n_{NOM}}{n_V}$) stb. Ezek a korpuszból automatikusan, kézi annotálás nélkül kigyűjthető adatok.

Kézi annotálással meghatározható, hogy a hat argumentumszerkezet-variáns egyenként 66-szor, 17-szer stb. fordul elő (n_{Vi}), ami 0,42, 0,11 stb. relatív gyakoriságot jelent ($f_{Vi} = \frac{n_{Vi}}{n_V}$). A táblázat többi részében az egyes argumentumszerkezet-variánsok mellett megjelenő egyes bővítmények megjelenési száma (n_{iX}) és megjelenési gyakorisága ($f_{iX} = \frac{n_{iX}}{n_{Vi}}$) található. Láthatjuk, hogy a kötelező vonzatok nem jelennek meg minden esetben az ige mellett, az alany például, amely mindegyik variánsnak vonzata, csak 0,33–0,53 gyakorisággal. Ennek egyrészt az az oka, az alanyi és tárgyi vonzat sokszor elhagyható (pro-drop), máskor egyenesen tilos kitenni (főnévi igenév alanya), az ellipszis (pl. *Péter keringőzött ~~Marival~~, Lajos pedig foxtrottozott Marival*) is látszólagos vonzathiányt okoz, illetve vannak egyszerűen hiányos mondatok (rövid válasz, pl. – *Találkoztál Marival? – Találkoztam ~~Marival~~*). Az alanyon kívüli vonzatok azonban igen nagy gyakorisággal megjelennek (>0,6). Azonban az is megfigyelhető, hogy a vizsgált bővítmények akkor is megjelenhetnek az igék mellett, ha annak nem vonzatai. A BAN esetű bővítmény például helyhatározóként a 3., 4. és 5. variánst is módosíthatják, ezekben az esetekben azonban aránylag kicsi a megjelenési gyakoriságuk (<0,2).

Hogy az argumentumszerkezetet közvetlenül a korpuszban ténylegesen megfigyelhető adatok alapján értelmezhessük, ezáltal számot tudjunk adni az esetleges vonzatelmaradásokról is, továbbá hogy egységes keretben tudjuk kezelni a kötelező vonzatokat és a szabad bővítményeket úgy, hogy közben a két csoport tagjainak a megkülönböztethetősége megmaradjon, a továbbiakban

- az igék vonzatszerkezetét illetve a vonzatszerkezeti variánsait nem a vonzatok felsorolásával, bináris listaként jellemezzük (1. táblázat), hanem a skaláris argumentumgyakorisági értékek listájával (2. táblázat), vagyis egy-egy argumentumgyakorisági vektorral.

Tegyük fel, hogy a magyar nyelvben a lehetséges bővítménytípusok a *bíz/bízik* ige kapcsán tárgyalt [NOM; ACC; BAN; RA; VAL] listával adhatók meg. Ekkor az ige első argumentumszerkezeti variánsát a [0,44; 0; 0,98; 0,03; 0] ötdimenziós vektorral jellemezzük, a második variánsát a [0,53; 0; 1; 0; 0] vektorral stb. Jelöljük ezeket a vektorokat $\bar{v}_1$-gyel, $\bar{v}_2$-vel, … $\bar{v}_6$-tal, a 2. táblázatban látható összesített [0,41; 0,42; 0,58; 0,28; 0,15] előfordulási gyakorisági vektort pedig $\bar{v}$-vel. $\bar{v}$ a *bíz/bízik* ige Szeged Korpuszban való előfordulásaiból közvetlenül meghatározható, $\bar{v}_i$ pedig kézi annotáció utáni számlálással. Ekkor a következő összefüggés áll fenn:

$$\bar{v} = \sum_{i=1}^{6} f_{Vi} \cdot \bar{v}_i \qquad (1)$$

vagyis az ige korpuszban megfigyelhető bővítménygyakorisági vektora az ige argumentumszerkezet-variánsainak a variáns előfordulási gyakoriságának a súlyozásával vett vektori összegével egyenlő.

Az igei argumentumszerkezetek-variánsok vektorainak birtokában és a variánsok korpuszbeli előfordulási gyakoriságának ismeretében tehát megkaphatjuk az ige bővítményeinek a korpuszbeli előfordulási gyakoriságvektorát.

3 Argumentumszerkezeti vektor meghatározása korpuszból

Egy V ige bővítményeinek egy adott korpuszban megfigyelhető előfordulási gyakoriságát tehát a $\bar{v}$ vektorral jellemeztük. Ez a vektor a korpuszból közvetlenül kinyerhető, amennyiben a korpusz szavai megfelelő morfoszintaktikai annotálással vannak ellátva, illetve automatikusan meghatározhatók a korpuszban a mondat és tagmondathatárok és a maximális főnévi kifejezések. Ha az ige V_i argumentumszerkezeti variánsokkal rendelkezik, ezeket az alternánsokat a $\bar{v}_i$ vektorokkal kívánjuk jellemezni, illetve az egyes variánsok korpuszbeli előfordulási gyakoriságát f_{Vi}-vel. Ezek, vagyis a variánsok argumentumszerkezet-vektorai és a variánsok gyakorisági együtthatói a korpuszból közvetlenül nem meghatározhatóak, viszont tudjuk, hogy teljesül rájuk az (1) egyenlőség.

Automatikusan meghatározható viszont az az információ, hogy az ige a korpuszban ugyanabban a tagmondatban ténylegesen milyen bővítményekkel fordul elő. Ezeket a megfigyelhető ige–bővítmény előfordulásokat kombinációstípusonként összegezhetjük is, illetve az ige összes előfordulásához viszonyítva a gyakoriságukat is megadhatjuk. Jelöljük a megkülönböztetett bővítménytípusok halmázát *ArgType*-pal (vagy AT-val). (Az előbbi *bíz/bízik* példában *ArgType* = {NOM, ACC, BAN, RA, VAL} volt.)

Az *ArgType* = {A, B, C stb.} *k* elemű bővítménytípus-halmaz esetén azoknak a mondatoknak a számát, ahol a *V* ige bővítmény nélkül jelenik meg, jelöljük n_{V+0}-val, relatív gyakoriságát f_{V+0}-val. Annak a számát, amikor csak A bővítménnyel fordul elő, jelöljük

n_{V+A}-val, relatív gyakoriságát f_{V+A}-val, amikor A-val és B-vel fordul elő, n_{V+A+B}-vel és f_{V+A+B}-vel és így tovább: n_{V+B} és f_{V+B}, n_{V+A+C} és f_{V+A+C} stb. Ha *k* darab különböző bővítménytípus veszünk figyelembe, akkor 2^k különböző kombinációban jelenhetnek meg ezek a bővítmények az ige mellett, vagyis ennyi előfordulási adatot és gyakorisági adatot kaphatunk a korpuszból, bár ezek nagy része valószínűleg egyszer sem fordul elő: például szinte nulla a valószínűsége annak, hogy egy ige az összes lehetséges bővítménnyel együtt jelenjen meg egy mondatban.

Azon mondatok számát, ahol az ige az A bővítménnyel jelenik meg, függetlenül más bővítménytípusok jelenlététől, jelöljük n_{V+A+*}-gal, relatív gyakoriságát f_{V+A+*}-gal, azon mondatok számát, ahol az ige A-val és B-vel jelenik meg, függetlenül más bővítménytípusok megjelenésétől, jelöljük $n_{V+A+B+*}$-gal stb. Azt várnánk, hogy $n_{V+A+*} = n_A$, vagyis az igét és az A bővítményt egyaránt tartalmazó mondatok száma megegyezik az A bővítmények számával, de ez nem szükségszerűen igaz: vannak eseteket, amikor a kettő eltérhet, pl. *A múlt évben még bíztam Mariban* esetében az ige egyszer fordul elő BAN bővítmény mellett, de a BAN bővítmény kétszer fordul elő az ige környezetében.

A *bíz/bízik* ige esetében nem csak a korábban bemutatott 5 lehetséges vonzattípussal kel számolni, hanem több szabad bővítménnyel is. Ezeket a további bővítményeket főnévi kifejezés esetében szintén az esetükkel lehet jellemezni, más esetekben pedig a megjelenő névutóval, az igenévi típussal (pl. főnévi igenév – NI) vagy a mondattípussal (pl. HKM). A Szeged Korpuszban az ige az említett bővítményeken kívül szerepelt még *hogy* kötőszavas mellékmondattal (HKM), szuperesszívusz esetű bővítménnyel (ON), terminatívuszi bővítménnyel (IG), ablatívuszi bővítménnyel (TÓL), különböző határozószókkal (ADV) és néhány névutós kifejezéssel (PP). Ez utóbbi öt típus csak néhányszor fordult elő, ezért most a határozószókat, illetve a névutós kifejezéseket összevontan kezelem. Az így kapott 11 bővítménytípussal összesen $2^{11} = 2048$ különféle bővítménykombinációt lehetne létrehozni, de a korpuszban – már csak azért is, mert összesen csak 157-szer szerepel a kérdéses ige – nem található meg mindegyik, hanem csak 40. Ebből a 40-ből is csak 10 olyan van, ami kettőnél többször fordul elő, ezekben pedig a HKM-en kívül csak a korábban ismertetett 5 bővítménytípus van jelen, ezek lefedik a *bíz/bízik* ige előfordulásának több mint a kétharmadát, összesen 116-ot a 157-ből.

Típus	előfordulási szám (n_{V+X})	gyakorisági szám (f_{V+X})
V+BAN	26	0,17
V+ACC+RA	21	0,13
V+NOM+BAN	21	0,13
V+BAN+HKM	12	0,08
V+NOM+BAN+HKM	11	0,07
V+NOM+ACC+VAL	7	0,04
V+ACC+VAL	6	0,04
V+NOM+ACC+RA	5	0,03
V+RA	4	0,03
V+ACC	3	0,02

3. táblázat. a *bízik* ige 10 leggyakoribb megjelenő bővítménykombinációja

A táblázatban szereplő adatok esetében nem tettem különbséget az igekötős és az igekötő nélküli igék között, csak az igével előforduló bővítménykombináció alapján összegeztem az adatokat. Az igekötők szerepére a 4.2.7 szakaszban térek vissza.

A 2. táblázat adatai az új bővítménytípusokkal kiegészítve a következő (a táblázat alsó sora $\vec{v}$):

	biz/bízik	NOM	ACC	BAN	RA	VAL	HKM	ON	IG	TÓL	ADV	PP
n	157	64	66	91	44	23	31	2	2	1	11	6
f	1,00	0,41	0,42	0,58	0,28	0,15	0,20	0,01	0,01	0,005	0,07	0,04

4. táblázat. a *bízik* ige összes bővítményének az előfordulási adatai

Korábban kézi annotációval, azaz a mondat értelmezésével és a mondatban szereplő bővítmények tulajdonságainak figyelembevételével határoztuk meg, hogy a korpusz egyes mondataiban melyik argumentumszerkezeti variáns található, ami alapján a 2. táblázatot összeállítottuk, vagyis az argumentumszerkezeti vektorokat és a variánsok előfordulási gyakoriságát meghatároztuk. A kérdés az, hogy meghatározhatjuk-e ezeket a vektorokat és gyakoriságokat automatikusan a korpuszból, kizárólag a hozzáférhető morfológiai és szintaktikai információkra hagyatkozva, a mondatok értelmezése nélkül, vagyis meghatározhatóak-e az 2. táblázatban látható adatok kizárólag a 3. és 4. táblázatban található információk alapján? Mivel itt már nem a megfigyelhető szerkezetek előfordulásait számoljuk, vagyis a relatív gyakoriságukat (f), hanem becsüljük azokat, ezért ezeket a meghatározandó értékeket előfordulási valószínűségnek (p) tekintjük.

3.1 Két triviális argumentumszerkezet-variáns

Az első probléma az argumentumszerkezeti vektorok automatikus meghatározásánál az, hogy nem tudjuk, hogy hány vektort keresünk, azaz hány variánsa van az igének. Erre a kérdésre két triviális válasz is lehetséges. A két triviális megoldás legtöbbször nem megfelelő leírása az adatoknak, de két fontos általánosítás megfogalmazására teremtenek lehetőséget.

3.1.1 Maximális variánsszámú ige

Tekinthetjük a 3. táblázatban felsorolt és a felsorolásból kihagyott, összesen 40 megfigyelhető bővítménykombinációt mind különálló argumentumszerkezet-variánsnak, ahol a megadott (megjelent) bővítmények előfordulási valószínűsége mind 1,00, a meg nem jelent bővítményeké pedig egyre 0,00, a variánsok előfordulási gyakorisága pedig megegyezik a megfigyelhető kombinációk előfordulási gyakoriságával, vagyis minden megjelenő bővítmény kötelező vonzat is egyben. Az argumentumszerkezeti variánsoknak ez a triviális listája így megfelel a (1) azonosságnak is. Azonban ekkor nem tudunk számot adni arról a jelenségről, hogy a természetes nyelvekre úgy tekintünk, hogy azokban vonzatok sem jelennek meg mindig, bizonyos esetekben a vonzatot is elhagyhatjuk. Továbbá szeretnénk olyan nyelvi leírást adni, ami a lehető leggazdaságosabb reprezentációt igényli, azaz

- Az ige argumentumszerkezeti variánsainak a számának minimalizálására törekszünk.

3.1.2 Egyvariánsos ige

Feltételezhetjük, hogy az igének csak egyetlen variánsa van. Ekkor mondhatjuk azt, hogy az ige egyetlen argumentumszerkezet-variánsa a 4. táblázatban látható argumentumszerkezeti vektorral jellemezhető, és a variáns gyakorisági együtthatója 1,00.

Ebben az esetben nem tudjuk megmagyarázni azt a tényt, hogy bár a korpuszban a vizsgált ige környezetében a BAN bővítmény (f_{BAN} = 0,58) és a tárgyi bővítmény (f_{ACC} = 0,42) a két leggyakrabban előforduló, együtt mégis csak nyolc mondatban találjuk meg mindkettőt (kb. 5%). A kisebb előfordulási gyakoriságú RA bővítmény (f_{RA} = 0,28) tárggyal együtt viszont sokkal többször, 37-szer szerepel (kb. 24%).

Feltételezzük ugyanis, hogy egy egyvariánsos ige különböző bővítményeinek a megjelenési valószínűsége független egymástól, az egyik megjelenése nem befolyásolja a másik megjelenési valószínűségét. Ez igaz a többvariánsos igék egyes variánsa esetében is:

- Egy ige egy argumentumszerkezeti variánsa esetében a variáns különböző bővítményeinek a megjelenési valószínűségei függetlenek egymástól.

Vegyük a V igének egy V_i variánsát (vagy egy egyvariánsos igét), ami mellett az A, B, C és D bővítmények jelenhetnek meg. Annak a valószínűsége, hogy a variáns mellett megjelenik az A bővítmény, p_{iA} (illetve p_{iB}, p_{iC}, p_{iD}), annak a valószínűsége pedig hogy az A bővítmény nem jelenik meg mellette, $1\text{-}p_{iA}$ (illetve $1\text{-}p_{iB}$, $1\text{-}p_{iC}$, $1\text{-}p_{iD}$). Ekkor a V+A+C bővítménykombináció előfordulási valószínűsége a V_i variáns mellett $p_{iV+A+C} = p_{iA}\cdot(1\text{-}p_{iB})\cdot p_{iC}\cdot(1\text{-}p_{iD})$, az A és a C bővítmény együttes előfordulásának a valószínűsége (függetlenül attól, hogy a B és a D megjelenik-e) $p_{iV+A+C+*} = p_{iA}\cdot p_{iC}$.

A *bíz/bízik* ige mellett a tárgy és a BAN bővítmény együttes előfordulásának a valószínűsége egyvariánsos igének feltételezve így $p_{V+ACC+BAN+*} = p_{ACC}\cdot p_{BAN} = 0{,}42\cdot 0{,}58 = 0{,}24$, a tárgy és a RA bővítményé pedig $p_{V+ACC+RA+*} = p_{ACC}\cdot p_{RA} = 0{,}42\cdot 0{,}28 = 0{,}12$ kellene hogy legyen, a megfigyelt 0,05 és 0,28 helyett.

3.2 Az argumentumszerkezeti vektor és a korpuszban megfigyelhető gyakoriságok közötti összefüggések

Az előző részben használt számolás mögötti összefüggések általánosítva a következők:

Legyen *ArgType* (vagy AT) a lehetséges bővítménytípusok halmaza, C pedig ennek egy részhalmaza. Jelöljük V+C-vel a azokat a bővítménykombinációkat, amikor az ige a C-ben levő bővítményekkel együtt jelenik meg (pl. ha $C = \{c_1, c_2, c_3\}$, akkor $V+C = V+c_1+c_2+c_3$). Ekkor

- a V ige V_i argumentumszerkezet-variánsa melletti V+C bővítménykombináció megjelenési valószínűsége

$$p_{iV+C} = \prod_{c \in C} p_{ic} \cdot \prod_{c \in AT \setminus C} (1 - p_{ic}) \qquad (2)$$

- ha az igének *k* különböző argumentumszerkezeti variánsa van, akkor az ige melletti V+C bővítménykombináció megjelenési valószínűsége

$$p_{V+C} = \sum_{i=1}^{k} \left(p_{Vi} \cdot \prod_{c \in C} p_{ic} \cdot \prod_{c \in AT \setminus C} (1 - p_{ic}) \right) \quad (3)$$

3.3 Argumentumszerkezeti vektor meghatározása – egyszerű példa

Vegyünk egy egyszerűsített példát, a *bíz/bízik* ige első (*vki bízik vmiben*) és harmadik (*vki bíz vmit vkire*) variánsát, és csak az ACC, BAN és RA bővítményeket vegyük figyelembe. A két variáns a korpuszban összesen 103-szor fordul elő, ebből 66 az első variáns, 37 a harmadik variáns előfordulási száma, vagyis $p_{V1} = 0{,}64$ és $p_{V3} = 0{,}36$. Tárgyi bővítmény 33-szor jelenik meg az ige mellett, mind a harmadik variánsnál, BAN bővítmény 69-szer, 4 kivételével az első variánsnál, RA bővítmény pedig 39-szer, kettő kivételével a harmadik variánsnál.

A korpuszból automatikusan kigyűjthető adatokat az 5. táblázat tartalmazza, kiemelve az adatok száma, illetve ezekből kiszámolhatóak a bővítménykombinációk gyakorisági értékei és az összesített $\bar{v}$ argumentumszerkezeti vektor. Megjegyzem, hogy ebben a példában az egyes igei bővítménykombinációk egyes korpuszbeli megjelenései minden esetben ugyanahhoz az argumentumszerkezeti variánshoz tartoztak, nevezetesen az első három sor a $bíz_1$, a második három pedig a $bíz_3$ variánshoz, de ez nem szükségszerű. Az ige csak ACC, vagy csak ACC és BAN bővítményekkel egyszer sem fordul elő.

kombinációk	ACC	BAN	RA	n	
V+BAN		+		63	$f_{V+BAN} = 0{,}611650$
V+BAN+RA		+	+	2	$f_{V+BAN+RA} = 0{,}019417$
V+0				1	$f_{V+0} = 0{,}009709$
V+ACC+RA	+		+	29	$f_{V+ACC+RA} = 0{,}281553$
V+ACC+BAN+RA	+	+	+	4	$f_{V+ACC+BAN+RA} = 0{,}038835$
V+RA				4	$f_{V+RA} = 0{,}038835$
V+ACC				0	$f_{V+ACC} = 0{,}0$
V+ACC+BAN			+	0	$f_{V+ACC+BAN} = 0{,}0$
össz.	33	69	39	103	
$\bar{v}$	0,320388	0,669903	0,378641		

5. táblázat. a *bízik* ige megfigyelhető előfordulási adatai három bővítménytípussal kombinálva

Ezen adatok ismeretében az a feladatunk, hogy meghatározzuk azokat a $\bar{v}_1 = [p_{1ACC}; p_{1BAN}; p_{1RA}]$ és $\bar{v}_3 = [p_{3ACC}; p_{3BAN}; p_{3RA}]$ vektorokat és a p_{V1} és p_{V3} valószínűségi együtthatókat ($p_{V1} + p_{V3} = 1$), amelyekkel a két argumentumvariánst jellemezhetjük. A kézi annotálás segítségével megszámolt értékek a 6. táblázatban találhatóak, nekünk ezt most azonban becsülnünk kell.

		ACC	BAN	RA
bíz$_1$ (n_1=66)	n_{1X}	0	65	2
p_{V1}=0,640777	$\bar{v}_1$	0	0,984848	0,030303
bíz$_3$ (n_3=37)	n_{3X}	33	4	37
p_{V3}=0,359223	$\bar{v}_3$	0,891892	0,108108	1,00

6. táblázat. a *bízik* ige két argumentumszerkezeti vektora kézi annotálással

Ha feltételezzük, hogy 2 argumentumszerkezeti variáns van, akkor a megbecsülendő adatokból az (1) és a (3) képletek szerint a következő számolt valószínűségi értékek határozhatóak meg:

$$p_{ACC} = p_{V1} \cdot p_{1ACC} + p_{V3} \cdot p_{3ACC} \quad (4)$$
$$p_{BAN} = p_{V1} \cdot p_{1BAN} + p_{V3} \cdot p_{3BAN}$$
$$p_{RA} = p_{V1} \cdot p_{1RA} + p_{V3} \cdot p_{3RA}$$
$$p_{V\text{-}0} = p_{V1} \cdot (1\text{-}p_{1ACC})\,(1\text{-}p_{1BAN})\,(1\text{-}p_{1RA}) + p_{V3} \cdot (1\text{-}p_{3ACC})\,(1\text{-}p_{3BAN})\,(1\text{-}p_{3RA})$$
$$p_{V+ACC} = p_{V1} \cdot p_{1ACC}\,(1\text{-}p_{1BAN})\,(1\text{-}p_{1RA}) + p_{V3} \cdot p_{3ACC}\,(1\text{-}p_{3BAN})\,(1\text{-}p_{3RA})$$
$$p_{V+BAN} = p_{V1} \cdot (1\text{-}p_{1ACC})\,p_{1BAN}\,(1\text{-}p_{1RA}) + p_{V3} \cdot (1\text{-}p_{3ACC})\,p_{3BAN}\,(1\text{-}p_{3RA})$$
$$p_{V+RA} = p_{V1} \cdot (1\text{-}p_{1ACC})\,(1\text{-}p_{1BAN})\,p_{1RA} + p_{V3} \cdot (1\text{-}p_{3ACC})\,(1\text{-}p_{3BAN})\,p_{3RA}$$
$$p_{V+ACC+BAN} = p_{V1} \cdot p_{1ACC}\,p_{1BAN}\,(1\text{-}p_{1RA}) + p_{V3} \cdot p_{3ACC}\,p_{3BAN}\,(1\text{-}p_{3RA})$$
$$p_{V+ACC+RA} = p_{V1} \cdot p_{1ACC}\,(1\text{-}p_{1BAN})\,p_{1RA} + p_{V3} \cdot p_{3ACC}\,(1\text{-}p_{3BAN})\,p_{3RA}$$
$$p_{V+BAN+RA} = p_{V1} \cdot (1\text{-}p_{1ACC})\,p_{1BAN}\,p_{1RA} + p_{V3} \cdot (1\text{-}p_{3ACC})\,p_{3BAN}\,p_{3RA}$$
$$p_{V+ACC+BAN+RA} = p_{V1} \cdot p_{1ACC}\,p_{1BAN}\,p_{1RA} + p_{V3} \cdot p_{3ACC}\,p_{3BAN}\,p_{3RA}$$

A célunk tehát az, hogy a $\bar{v}_1 = [p_{1ACC}; p_{1BAN}; p_{1RA}]$ és $\bar{v}_3 = [p_{3ACC}; p_{3BAN}; p_{3RA}]$ vektorokra és a p_{V1} és p_{V3} valószínűségi együtthatókra olyan becslést adjunk meg, amelyek alapján a (4)-ben számolt valószínűségi tényezők a ténylegesen megfigyelt f_{ACC}, f_{BAN}, f_{RA}, $f_{V\text{-}0}$, $f_{V\text{-}ACC}$, $f_{V\text{-}BAN}$, $f_{V\text{-}RA}$, $f_{V\text{-}ACC\text{-}BAN}$, $f_{V\text{-}ACC\text{-}RA}$, $f_{V\text{-}BAN\text{-}RA}$, $f_{V\text{-}ACC\text{-}BAN\text{-}RA}$ gyakorisági tényezőket legjobban megközelítik, vagyis az azokhoz viszonyított különbségeik négyzeteinek összege minimális:

$$(f_{ACC}\text{-}p_{ACC})^2 + (f_{BAN}\text{-}p_{BAN})^2 + (f_{RA}\text{-}p_{RA})^2 + (f_{V+0}\text{-}p_{V+0})^2 + (f_{V+ACC}\text{-}p_{V+ACC})^2 + (f_{V+BAN}\text{-}p_{V+BAN})^2 + (f_{V+RA}\text{-}p_{V+RA})^2 + (f_{V+ACC+BAN}\text{-}p_{V+ACC+BAN})^2 + (f_{V+ACC+RA}\text{-}p_{V+ACC+RA})^2 + (f_{V+BAN3RA}\text{-}p_{V+BAN+RA})^2 + (f_{V+ACC+BAN+RA}\text{-}p_{V+ACC+BAN+RA})^2 \quad (5)$$

Mivel most 3 bővítménytípus és 2 variáns van, ez egy $2\cdot(3+1)$ dimenziós térben való minimumkeresés. k bővítménytípus és n variáns esetében ez a keresés $n\cdot(k+1)$ dimenziós térben történik.

Természetesen elvégezhetjük a számítást több argumentumszerkezeti variánst feltételezve is. A helyes variánsszám meghatározásánál figyelembe kell venni azt, hogy egyrészt törekednünk kell a minél kisebb variánsszámra (3.1.1 szakasz), de azért az adatokat minél jobban megmagyarázni képes modellt szeretnénk kialakítani (3.1.2 szakasz).

4 Az argumentumszerkezeti vektort befolyásoló tényezők

Az argumentumszerkezeti vektor értékét több tényező is befolyásolja, például a korpusz egyedi tulajdonságai, ami alapján meghatározzuk a vektort, de vannak grammatikai befolyásoló tényezők is. Ezen tényezők számbavétele és a hatásuk leírása egyrészt a hatás kiküszöbölésével pontosíthatja az argumentumszerkezeti vektor meghatározását, másrészt feltárásukkal hasznos összefüggésekre lelhetünk a nyelv és a nyelvtan működését illetően.

4.1 Korpuszhatások

Ha az argumentumszerkezeti vektort korpusz alapján határozzuk meg, akkor a vektor a korpusz adatait fogja visszatükrözni, más korpuszt választva más értékeket kaphatnánk. A korpusz mérete is befolyásolja ezt a folyamatot, nagyobb korpusz esetén csökken az adatok esetlegességének a mértéke.

Az argumentumszerkezeti variánsok egymáshoz viszonyított előfordulási valószínűsége például erősen korpuszfüggő. A különböző variánsok ugyanis különböző jelentést hordozhatnak, ezért a korpuszban szereplő szövegek típusa, témája meghatározza, hogy mely variánsok lesznek a gyakoribbak benne. A hivatalos, jogi vagy gazdasági szövegekben várhatóan kevesebbszer fordul elő a *bíz/bízik* ige 6. variánsa: *vki elbízza magát*, az iskolások fogalmazásaiban vagy a szépirodalmi szövegekben, a *vki megbíz vkit vmivel* viszont gyakoribb lesz a gazdasági hírekben.

A korpuszban szereplő szövegek típusa az egyes vektorokban megjelenő argumentumok előfordulási valószínűségét is befolyásolja. Az iskolai fogalmazásokban sokkal többször jelenik meg az első és második személyű névmás, ugyanígy a szépirodalmi művekben is, mint a formálisabb szövegekben, a névmások viszont hajlamosabbak a meg nem jelenésre, mint a kifejtett főnévi kifejezések. Ezért ezekben fogalmazásokban várhatóan kisebb lesz az alanyi és tárgyi bővítmények megjelenési valószínűsége, mint a jogi szövegekben (ha csak ezt a különbséget vesszük figyelembe). De a fogalmazások és az irodalmi művek között is találhatunk különbséget, például a nem kötelező bővítmények megjelenési valószínűségét illetően.

Az előző bekezdésben ismertetett hatások azonban nem közvetlenül szövegtípusok és az argumentumszerkezeti vektorok között érvényesülnek, hanem a következő szakaszban ismertetett grammatikai hatásokon keresztül. Az egyes szövegtípusokra jellemző ugyanis azok névszó- és bővítményhasználata, és ha ezeknek tényezőknek az argumentumszerkezeti vektorokra való befolyását elkülönítve tudjuk jellemezni, akkor már csak azt kell megállapítani, hogy ezek a tényezők mennyire jellemzők a korpuszokra.

4.2 Grammatikai hatások

4.2.1 Pro drop

A magyarban a hangsúlytalan alany és tárgy esetű névmások elhagyhatóak. Ha korpuszvizsgálattal meghatározzuk, hogy az alanyi, illetve a tárgyi vonzattal rendelkező igék alanya, illetve tárgya mekkora valószínűséggel lesz (megjelenő vagy elhagyott) személyes névmás ($p_{pron\text{-}NOM}$, $p_{pron\text{-}ACC}$), továbbá meghatározzuk, hogy névmási alany és tárgy mekkora valószínűséggel kerül elhagyásra ($p_{prodrop\text{-}NOM}$, $p_{prodrop\text{-}ACC}$), akkor a névmáselhagyás hatása kiküszöbölhető. Ha ugyanis az ilyen alanyok és tárgyak nem lennének elhagyva, akkor a ténylegesen megfigyelhető adatokból számolt p_{iNOM}, illetve p_{iACC} alanyi és tárgyi valószínűség helyett a $p'_{iNOM} = p_{iNOM} + p_{pron\text{-}NOM} \cdot p_{prodrop\text{-}NOM}$ stb. korrigált alanyi előfordulási valószínűséggel dolgozhatunk.

A $p_{pron\text{-}NOM}$ és $p_{prodrop\text{-}NOM}$ valószínűségek nem csak egy igére vagy igevariánsra jellemző értékek, hanem az összes igére és variánsra: $p_{pron\text{-}NOM}$ korpuszfüggő valószínűség, $p_{prodrop\text{-}NOM}$ viszont korpuszfüggetlen.

4.2.2 Ellipszis

Nem csak az alanyi és a tárgyi névmás hagyható el a magyarban, hanem más vonzatelhagyási jelenségek is megfigyelhetőek. Az összetett mondatokra, különösen a mellérendelésekre jellemző, hogy ha ugyanaz a kifejezés több tagmondatban is jelen van, akkor csak az egyik tagmondatban jelenik meg: *Péter csak találkozott ~~Marival~~, de Pál beszélgetett is Marival.* A különböző típusú vonzatok elliptálhatósága a névmáselhagyási jelenséghez hasonlóan egy valószínűségi értékkel jellemezhetők, bár ebben az esetben a korpuszhatás nehezebben elhatárolható a teljes valószínűségi értéktől, és a különböző igék is különböző mértékben hajlamosak az ellipszisben való részvételre.

4.2.3 Szabad bővítmények igefüggetlen megjelenése

A 3. szakaszban bevezetett argumentumszerkezeti vektor nem tesz különbséget vonzat és szabad bővítmény között, azonban a vonzat és a szabad bővítmény ebben az értelmezésben is jól elkülöníthető egymástól: a hagyományosan vonzatnak tekintett bővítmények nagy valószínűséggel megtalálhatóak az ige mellet ($p>0,6$), míg a szabad bővítmények előfordulási gyakorisága kicsi ($p<0,4$). Ez alól csak az alanyi és tárgyi bővítmények jelenthetnek kivételt, de azok meg mindig vonzatok.

Míg a vonzatok esetében a vektor megfelelő értékének értelmezésekor azt kell megindokolni, hogy mikor, mekkora valószínűséggel nem jelenik meg mégsem az ige mellett, a szabad bővítményeknél a megjelenést kell alátámasztani: mivel a szabad bővítmény nem kötelező, mikor jelenik meg mégis, mekkora ennek a valószínűsége. A szabad bővítményeket nem az igék szelektálják, ezért egy adott szabad bővítménytípus megjelenés valószínűsége csak kis mértékben igefüggő, a különböző igék és argumentumszerkezeti variánsok melletti megjelenési gyakorisága állandónak tekinthető. Az igék különböző szabadbővítmény-felvevő hajlandósága csak közvetetten köthető az igéhez: a szabad bővítmények jellemzője az, hogy milyen típusú, milyen jelentéskategóriájú igéhez tudnak kapcsolódni, ezáltal az igék osztályozása áttételesen ad magyarázatot a varianciára. Mindazonáltal egy szabad bővítménytípus megjelenési valószínűségét több ige vizsgálatával korpusz alapján egységesen lehet megállapítani, az egyes alternánsok esetében pedig ezt lehet irányadónak venni.

4.2.4 Szabadbővítmény-csoportok

A korábbiakban az egyes bővítménytípusokat a bővítmény esetével vagy névutójával jellemeztük. Azonban vannak olyan esetcsoportok, amelyeket érdemesebb együtt kezelni, ugyanannak a bővítménytípusnak a különböző megnyilvánulásainak tekinteni őket. Például a helyhatározói funkciójú bővítmények hasonlóan működnek, ugyanolyan predikátumtípusokhoz illeszthetőek, egymással helyettesíthetőek, bár a morfológiai esetük többféle is lehet: BAN, ON, NÁL vagy MELLETT stb. Az ugyanolyan funkciójú, de különböző morfológiai esetű szabad bővítményeket ezért kívánatos egy bővítménytípusnak tekinteni és egységesen meghatározni a megjelenési valószínűségét, gyakoriságát: f_{HELY}. Ugyanakkor az ugyanolyan funkciójú, de különböző esetű bővítmények egyenként is jellemezhetőek aszerint, hogy az adott funkciójú megjelenő szabad bővítmény mekkora valószínűséggel realizálódik egy bizonyos esetű kifejezésként. Ez esetenként változó nagyságú lehet, a realizálódási értékek nagysága független az igétől, ami mellett megjelennek. Ha egy argumentumvariáns mellett a kérdéses bővítménytípusok (esetek) a funkcióra jellemző valószínűségekkel jelennek meg egymáshoz képest, akkor az adott funkciót betöltő szabadbővítmény-csoport tagjainak tekinthetőek.

4.2.5 Argumentumszerkezet-típusok, argumentumszerkezet-változtató műveletek

Az argumentumszerkezeti vektorok segítségével az egyes igei lexikai egységek is összevethetőek: megvizsgálhatjuk, hogy melyek azok a lexikai egységek, variánsok, amelyek azonos vagy nagyon hasonló argumentumszerkezeti vektorral jellemezhetőek. Ezek – az igék jelentésének az előzetes vizsgálata és ismerete nélkül – utalhatnak arra, hogy a talált hasonló lexikai egységek valamilyen szintaktikai vagy szemantikai tulajdonságukban megegyeznek, ugyanabba a szintaktikai vagy szemantikai csoportba tartoznak.

Továbbá megvizsgálható, hogy vannak-e olyan igealakok, amelyek hasonló argumentumszerkezeti variánsokkal rendelkeznek, van-e értelmezhető grammatikai kapcsolat a több argumentumszerkezeti variánssal rendelkező kifejezések variánsai között.

Érdekes grammatikai általánosítások megfogalmazásához vezethet annak vizsgálata, hogy a nyilvánvaló morfológiai kapcsolatot mutató tövek különböző argumentumszerkezeti variánsai között van-e valamilyen kapcsolat. A *készül-készít, hárul-hárít, gurul-gurít* unakkuzatívuszi-akkuzatívuszi párok argumentumszerkezet-variánsai például egyértelműen párba állíthatóak, de a párhuzamokon túl érdekesek az egyes argumentummegjelenési valószínűségek változásai is, illetve az egyediségek is: melyek azok a variánsok, amik csak az egyik párnál jelennek meg, a többinél nem. Ezek az egyedi variánsok idiomatikus variánsai.

Például a *készül-készít* (és a *gurul-gurít* stb.) párok esetében megfigyelhető, hogy a *készül* ige alanya a *készít* ige tárgyának feleltethető meg (pl. *elkészült a leves – Péter elkészítette a levest* vagy *a cipő bőrből készült – a cipész bőrből készítette a cipőt*), vagyis az igék alanyi és tárgyi bővítményeinek a megjelenése korrelál. Vannak azonban olyan bővítményi környezetek, ahol ez a korreláció nem figyelhető meg (pl. *Péter Debrecenbe készül* – ?*Mari Debrecenbe készíti Pétert*), így ezek a variánsok nem célpontjai részt az argumentumszerkezet-változtató műveletnek: idiomatikusabbak.

Hasonlóan lehet jellemezni az egyes igeképzők argumentumszerkezet-változtató képességét is.

4.2.6 Örökölt vonzatok

Nem csak az ige argumentumai, vagyis a kötelező és szabad bővítményei jelenhetnek meg az ige mellett ugyanabban a tagmondatban, hanem más szintaktikailag önálló összetevők is. Ilyenek például az alany jelenlétében, de nem vele egy összetevőt alkotó VAL típusú bővítmények (pl. *Péter elment Marival a moziba*), vagy a szétváló birtokos kifejezések esetében a DAT birtokos (pl. *Péternek elment a barátja a moziba*). Ezek a bővítmények nem argumentumai az igének, nincsenek azzal szemantikai kapcsolatban, de az olyan argumentumszerkezeti modellekben, ahol csak a morfológiai és szintaktikai tényezőket vesszük figyelembe a bővítménység megállapításánál, ezek nem különböztethetőek meg egyszerűen a szabad bővítményektől. (Az ilyen jellegű problémák egyedi kezelésére lásd pl. Sass Bálint disszertációjának 2.2. szakaszát: [12].)

Azonban arról, hogy ezek milyen argumentumok mellett jelenhetnek meg (annak típusával vagy szótövével azonosítva), lehet feltételeket meghatározni, mint ahogy ahhoz is lehet valószínűségi értéket rendelni, hogy a megadott feltételek teljesülése esetén mekkora valószínűséggel jelenik meg az ilyen örökölt bővítmény.

Külön említést érdemelnek azok az igék, amelyeknek főnévi igeneves vonzatuk is van: ezeknek a főnévi igeneveknek a vonzatai, argumentumai megszorítás nélkül kerülhetnek a mátrix igével azonos tagmondatba is a mátrix ige bővítményeként (pl. *A házi feladatot tegnap elfelejtettem megcsinálni*, ahol a tárgy nem az *elfelejt* ige saját tárgya).

4.2.7 Igekötők

A 2. szakaszban az igekötős igéket és az igekötő nélkülieket megkülönböztettük, külön argumentumszerkezeti variánsnak tekintettük őket, ezáltal az igekötőket az igekötős ige részeként elemeztük, nem az ige önálló argumentumaként. Az igekötők azonban időnként átveszik valamely kötelező argumentum szerepét, a *rá* igekötő jelenlétében például nem jelenhet meg a *néz* ige mellett az egyébként kötelező névmási *rá* vonzat: *Péter ránézett *rá*. Az első személyű *rám* névmási vonzat esetében viszont az igekötő az, amit nem tehetjük ki: *Péter* (**rá*)*nézett rám*. Egyébként pedig, főnévi fejű RA bővítmény mellett, az igekötőt szabadon megjelenhet vagy elhagyható: *Péter Marira nézett/Péter ránézett Marira*. Ezekben az esetekben nem egyértelmű, hogy két argumentumszerkezeti variánst látunk-e, egy igekötőset és egy igekötő nélkülit, vagy pedig hármat, ahol a harmadik egy olyan igekötős *néz* variáns van, aminek nincs RA vonzata. És bármelyik megoldást is választjuk, a *Péter rám nézett* mondat igéjének besorolása elméleti szempontból is kérdéses.

Vagy választhatjuk azt a leírási módot is, hogy az igekötőket mint önálló mondatbeli összetevőket bővítménynek tekintjük, és megjegyezzük, hogy a *rá* igekötői bővítmény és a RA esetű bővítmény hajlamosak együtt megjelenni, mintegy szétváló, de szemantikailag összetartozó bővítményt alkotva. Ekkor hasonló leírást kívánnak meg, mint az elváló birtokos és a birtok: az egyik a bővítmény, a másik pedig annak a vonzata, ami esetlegesen az ige bővítményeként jelenik meg, példánkban a RA esetű bővítmény vezeti be a klitikumszerű *rá* igekötői bővítményt.

Máskor viszont az igekötős ige argumentumszerkezetében olyan vonzat jelenik meg, ami az igekötő nélküli esetben nem engedélyezett: **Péter megy az ajtón* de *Péter átmegy az ajtón*. Ebben az esetben az igekötő megjelenése az, ami engedélyezi az ON vonzat megjelenését.

Az igekötős igék argumentumszerkezeti vektorainak a vizsgálatával megállapíthatjuk, hogy egy adott igekötő milyen bővítménytípusokkal szokott együtt megjelenni: ezeket az igekötő-bővítménytípus párokat így összetartozókként kezelhetjük. Ugyanezen igék igekötő nélküli változatainak a vizsgálatával leírhatjuk, hogy az igekötő megjelenése milyen argumentumszerkezeti változást okoz, mint ahogyan a képzők argumentumszerkezet-változtató képességét is leírjuk. Megállapíthatjuk, hogy milyen feltételek mellett, vagyis milyen argumentumszerkezeti variáns esetében lehet egy bizonyos igekötővel ellátni egy igét, és hogy az igekötő megjelenése hogyan változtatja meg az argumentumszerkezeti variáns argumentumvektorát. Ha a feltárt feltételeknek megfelelő variáns hiányában is megjelenhet egy igekötő, vagy nem az elvárt módon változtatja meg az ige argumentumszerkezetét, akkor idiomatikus igekötő-ige párt találtunk. Az *el* igekötőtől például azt várnánk, hogy ha van valamilyen argumentumszerkezet-változtató képessége, akkor valamilyen BÓL/BA jellegű bővítmény megjelenését erősíti, nem pedig mondjuk BAN bővítményt (*Péter elindult az iskoláBÓL*/*iskoláBA*/?*iskolában*). A korábban vizsgált *bízik* ige esetében azonban ACC bővítmény megjelenését tapasztalhatjuk: *vki elbízza magát*. Ez nem magyarázható az igekötő szokásos viselkedésével, vagyis az *elbízik* igekötős ige idiomatikus szerkezetű.

5 Összefoglalás, alkalmazási lehetőségek

A tanulmány a lexikai egységek, tipikusan igék argumentumszerkezetének a leírására javasol egy új reprezentációs formát, ami nem a klasszikus kötelező vonzat – szabad bővítmény bináris oppozíciós lehetőségeket ragadja meg. Ehelyett az egyes bővítménytípusoknak a korpuszban való megjelenési gyakoriságai alapján a típusokhoz egy-egy valószínűségi értéket rendel, így az argumentumszerkezeti variánsok egy argumentumszerkezeti valószínűségi vektorral jellemezhetőek. A javasolt módszer kizárólag a korpuszbeli adatok morfológiai és szintaktikai tulajdonságaira támaszkodik, nem is célja a lexikai elemek szemantikai jellemzése, továbbá nem a vizsgált lexikai elemek környezetében levő kifejezések alakját vagy szótövét veszi figyelembe, hanem csak néhány absztraktabb, általánosabb tulajdonságát, ezért nem tekinthető a szóbeágyazási modellek egy változatának [5]. Az argumentumszerkezet többdimenziós értelmezése miatt inkább Sass Bálint duplakocka-modelljével [6] rokonítható.

Az argumentumszerkezeti variánsok argumentumszerkezeti vektorként való értelmezése új elméleti modellként a grammatikaelméletekben hozhat új eredményeket: a 4.2. szakaszban bemutatott, az argumentumszerkezeti vektorokat befolyásoló grammatikai tényezők feltárásával korpuszra, vagyis valós nyelvi adatokra támaszkodó grammatikai összefüggéseket lehet megfogalmazni. Az elméleti eredményeken túl azonban az argumentumszerkezeti vektorok a nyelvfeldolgozás során is több helyen alkalmazhatóak:

- Az argumentumszerkezeti vektorok a bővítmények valószínűségi értékeinek felhasználásával közvetlenül átalakíthatóak valószínűségi frázisstruktúra nyelvtanná ([13] 494.o.).
- A régensek környezetét vizsgálva valószínűsíthetjük, hogy az adott mondatban melyik argumentumszerkezeti variánsát találjuk. Abban az esetben, amikor a különböző argumentumszerkezeti variánsok más jelentést hordoznak, ez a jelentésegyértelműsítést is magával hozza.
- Az alanyi és tárgyi névmások elhagyásának a valószínűségét ismerve egy vizsgált szövegben az is megállapítható lehetne a legvalószínűbb argumentumszerkezeti variáns megtalálásával, hogy mellette szerepel-e zéró névmás, ami az anaforafeloldás során fontos információ.
- Elég nagy korpusz segítségével a szövegtípusok argumentumszerkezetivektor-változtató képességét is megadhatjuk, aminek a segítségével egy ismeretlen szöveg típusára adhatunk becsléseket.
- A lexikai elemek szokásos argumentumvektorainak ismeretében egy nyelvhasználónál az azoktól eltérő vektorok meglétéből következtethetünk a beszélő nyelvhasználati tulajdonságaira, így például a beszélő korára, társadalmi helyzetére vagy a mentális képességeire, nyelvi zavaraira is.

Mindezek fényében a lexikai elemek argumentumszerkezeti variánsainak vektoros reprezentációja mind elméleti, mind gyakorlati szempontból átgondolni érdemesnek látszik.

Bibliográfia

1. Carnie, A.: Syntax: a generative introduction. Wiley-Blackwell, Hoboken, New Jersey (2013).
2. Pollard, C., Sag, I.A.: Head-driven phrase structure grammar. CSLI, University of Chicago Press, Stanford, Chicago (1994).
3. Szécsényi T.: Magyar mondatszerkezeti jelenségek elemzése HPSG-ben. In: Bartos H. (ed.) Új irányok és eredmények a mondattani kutatásban. pp. 99–138. Akadémiai Kiadó, Budapest (2011).
4. Kovács V., Simkó K., Szécsényi T.: Szabályalapú szintaktikai elemző szintaktikai szabályok nélkül. In: Tanács A., Varga V., and Vincze V. (eds.) XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). pp. 251–259. Szegedi Tudományegyetem, Szeged (2016).
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. ArXiv13104546 Cs Stat. (2013).
6. Sass B.: Az igei szerkezetek algebrai struktúrája, avagy a duplakocka modell. Argumentum. 14, 12–44 (2018).
7. Kálmán L.: Miért nem vonzanak a régensek? In: Kálmán L. (ed.) KB 120. A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósi András tiszteletére. pp. 229–246. MTA Nyelvtudományi Intézet, Tinta Könyvkiadó, Budapest (2006).
8. Kálmán L.: Bővítménykeretek mint konstrukciók. In: Kas B. (ed.) "Szavad ne feledd" Tanulmányok Bánréti Zoltán tiszteletére. pp. 61–72. MTA Nyelvtudományi Intézet, Budapest (2016).
9. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Matoušek, V., Mautner, P., and Pavelka, T. (eds.) Text, Speech and Dialogue. pp. 123–131. Springer Berlin Heidelberg, Berlin, Heidelberg (2005).
10. Vincze, V., Szauter, D., Almási, A., Móra, G., 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).
11. Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., and Mukherjee, J. (eds.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006).
12. Sass B.: Igei szerkezetek gyakorisági szótára - Egy automatikus lexikai kinyerő eljárás és alkalmazása, http://real-phd.mtak.hu/342/, (2011).
13. Jurafsky, D., Martin, J.H.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Pearson Education Internat, Upper Saddle River (2009).