Morfológiai idioszinkrázia többszavas kifejezésekben

Oravecz Csaba, Varasdi Károly, Nagy Viktor[1]

[1] MTA Nyelvtudományi Intézet,
Budapest 1068, Benczúr u. 33.
{oravecz,varasdi,nagy}@nytud.hu

Kivonat: A dolgozatban megvizsgáljuk, hogy magyar nyelven egyes szókapcsolatok morfológiailag idioszinkratikus viselkedése, mint lehetséges információforrás, használható-e többszavas kifejezések korpuszból történő kinyerésére. Megmutatjuk, hogy legalábbis egyes TSZK típusok esetén, a toldalékolás idioszinkráziája jól jelzi a szókapcsolat TSZK státuszát illetve idiomatikusságát.

1. Bevezető

A számítógépes nyelvfeldolgozásban az utóbbi időkben számos módszert fejlesztettek ki többszavas kifejezések (TSZK) korpuszból történő kinyerésére illetve azonosítására [3]. Többségük a korpuszból kinyert pozíciós illetve relációs jelöltlisták [2] tagjait rangsorolja valamilyen asszociációs mérték segítségével. Olyan nyelvek esetében azonban, melyek morfológiája pl. az angolnál sokkal gazdagabb információforrást jelent, a kutatás éppen csak elkezdődött az „együtt előfordulás" mellett egyéb információ felhasználására [1]. Magyar nyelvre egyes szókapcsolatok morfológiailag idioszinkratikus viselkedése természetesen adódik, mint lehetséges további információforrás, melyet TSZK-k bizonyos csoportjainak azonosításában fel lehet használni.

A dolgozatban megvizsgáljuk a szókapcsolat jelöltek tagjainak toldalékeloszlásából kinyerhető információ felhasználhatóságát, és esettanulmányokon keresztül megmutatjuk, hogy legalábbis egyes TSZK típusok esetén, a toldalékolás idioszinkráziája jól jelezheti a szókapcsolat TSZK státuszát illetve idiomatikusságát.

2. A kivonatoló módszer

Egy szósorozatot akkor tekintünk morfológiailag illetve morfoszintaktikailag idioszinkratikusnak, ha egyes tagjainak toldalékeloszlása az adott szókapcsolatban jelentősen eltér a tagok összes előfordulásra vetített toldalékeloszlásától. Ez a megközelítés bizonyos mértékben eltér [1] módszerétől, ahol adott inflexiós jegyek csupán a már azonosított TSZK-n belül kerülnek összehasonlításra, és a jegyek egyes értékeinek (pl. egyes vagy többes szám) aránya a TSZK morfoszintaktikai preferenciájának jelzésére szolgál. Az általunk alkalmazott eljárás más megközelítésben, általános módszerként kívánja felhasználni a szókapcsolaton kívüli illetve belüli toldalékelosz-

lást, és az ebből kinyert információ segítségével próbálja azonosítani a TSZK-t. Ezáltal független osztályozóként az együtt előforduláson alapuló mértékek helyett, és nem utánuk, mint további feldolgozó lépés kíván szerepelni.

A munkahipotézis a következő. A jelöltlista valamilyen szintaktikai viszonyban álló 2 szavas kombinációkat tartalmaz, ahol szabad morfoszintaktikai jegyeknek nevezzük azokat a jegyeket, amelyeket nem ez a viszony kényszerít ki (egyeztetéssel vagy kormányzással). Ezek akkor vagy a tagok inherens jegyei, vagy a mondatszerkezet másik frázisa írja elő meglétüket. Pl. a „bedobja ... törölközőt" TSZK-ban a tárgyrag nem szabad, mert az állítmány-tárgy viszony írja elő, viszont a törölköző szám- stb. jegye szabad. Az ige minden lehetséges jegye szabad. A hipotézis az, hogy egy TSZK tag szabad jegyekre vett statisztikai eloszlása eltér az ő összesített (itt a szótő vagy a nem szabad jeggyel ellátott szótő összes előfordulását tekintjük) eloszlásától, ha a TSZK tagjaként fordul elő, és ez az eltérés jól jelzi a szemantikai átlátszatlanságot. Lehetséges viszont, hogy pusztán az is megváltoztatja a jegyeloszlást, hogy valamilyen szintaktikai viszonyban áll a tő. Ezért szűkebb környezetre kell az eloszlást vizsgálni, és a csupán az ugyanazon szintaktikai viszonyban álló alakok eloszlásának különbözőségét figyelembe venni. A tesztet a TSZK mindkét tagjára külön végre lehet hajtani, és így azt is megkaphatjuk, melyik tag jelentése változott meg a TSZK-ba kerüléskor.

3. Statisztikai vizsgálat

Az inflexiós elemzést az 1. táblázat szerint osztjuk fel dimenziókra.

1. táblázat. A különböző szófajoknál figyelembe vett inflexiós jegyek

Szófaj	**Dimenziók**				
Névszók	szám	birtokos szám/személy	anafonikus possessivus	eset	fok (melléknévnél)
Igék	mód/idő	határozottság	szám/személy	–	–

Minden potenciális többszavas kifejezésben (C) a tagok inflexiós eloszlását ezen jegyek mentén parametrizáljuk. Egy paraméter egy jegy (F) – érték (v) párt képvisel. Minden paraméterhez hozzárendeljük a jegy-érték pár relatív gyakoriságát:

$$P(F_i = v_j \mid w_k, C) = \frac{c(F_i(w_k) = v_j \ ha \ w_k C \ tagja)}{c(C)} \quad (1)$$

Nyilvánvalóan fennáll a következő összefüggés: $\sum_j \frac{c(F_i = v_j)}{c(C)} = 1$. Ezt az eloszlást kell összehasonlítani az összesített $P(F_i = v_j \mid w_k)$ eloszlással, vagyis amikor a tagszó előfordulásait nem korlátozzuk arra, hogy tagja legyen a többszavas kifeje-

zésnek, viszont feltételül kell szabni, hogy ugyanolyan szintaktikai szerkezeti pozícióban legyen, mint C-ben (pl. ha a TSZK-ban a tagszó főnevet módosító melléknév, akkor az összesített eloszlásban nem vesszük figyelembe azokat az előfordulásokat, amikor állítmányi szerepű).

$$P\ (F_i = v_j \mid w_k) = \frac{c\ (F_i(w_k) = v_j)}{c\ (w_k)} \tag{2}$$

A vizsgálatokban több szókapcsolatjelölt toldalékolási mintáját elemeztük. Számos esetben volt felfedezhető összefüggés az eloszlás egyenetlensége és a szósorozat idiomatikussága között, mely mutatja, hogy a tárgyalt megközelítés mindenképpen biztató eredményeket ad. További kutatást igényel viszont az eloszlások összehasonlítását végző legjobb mérték kiválasztása, illetve a poliszémiából származó torz adatok ki-küszöbölésének módja is.

Bibliográfia

1. Evert, S., Heid, U., Spranger, K.: Identifying morphosyntactic preferences in collocations. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 907–910
2. Evert, S., Krenn, B.: Computational approaches to collocations. Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003) (2003) Vienna.
3. Krenn, B.: The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 7. PhD thesis, Universität des Saarlandes, Department of Computational Linguistics (2000)