Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban

Vincze Veronika

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
vinczev@inf.u-szeged.hu

Kivonat: A félig kompozicionális főnév + ige szerkezetek számítógépes nyelvészeti kezelésének megkönnyítésére hozzuk létre a Szeged Korpusz egy olyan változatát, amelyben e kifejezések és altípusaik annotálva vannak. Az elkészült korpusz tanító adatbázisként szolgálhat a szerkezetek automatikus azonosításához, így hozzájárulhat többek között a gépi fordítás és az információkinyerés eredményességéhez.

1 Bevezetés

A számítógépes nyelvészeti alkalmazások számára az egyik legnagyobb kihívást a kollokációk megfelelő kezelése jelenti. Kollokációk gyakran előfordulnak a nyelvhasználatban, és viselkedésük sokszor eltér a kompozicionális kifejezésekétől, ezért különleges bánásmódot igényelnek.

2 Félig kompozicionális szerkezetek

A kollokációk egyik altípusának tekinthetők a félig kompozicionális főnév + ige szerkezetek (*tanácsot ad, döntést hoz, virágba borul…*) [1], melyekben a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában. E szerkezetek számítógépes nyelvészeti kezelése nem problémamentes. Mivel jelentésük nem teljesen kompozicionális, a szerkezet részeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Továbbá, a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*), így azonosításuk nem valósulhat meg pusztán szintaktikai mintákat figyelembe véve. Végül, mivel a szerkezet szintaktikai és szemantikai feje nem azonos, a szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni – az angol vonzatos igékhez (phrasal verbs) hasonlóan. Mindezen jellemzők miatt a félig kompozicionális főnév + ige szerkezetek felismerése és megfelelő kezelése kulcsfontosságú a számítógépes nyelvészeti alkalmazásokban, például a gépi fordításban és az információkinyerésben.

Egy félig kompozicionális szerkezeteket tartalmazó adatbázis létezése igencsak megkönnyítené az ilyen szerkezetek automatikus felismerését (így azok megfelelő kezelését is). Más nyelvekre léteznek már ilyen korpuszok: például hozzáférhető egy többszavas igéket tartalmazó adatbázis az észtre [2, 3] és a prepozíciós vonzattal rendelkező igék adatbázisa a németre [4]. Ezek nyomán hozzuk létre az első olyan magyar nyelvű korpuszt, melyben a félig kompozicionális főnév + ige szerkezetek be vannak jelölve. Az annotáció alapját a Szeged Treebank 2.0 képezi [5], mivel ez az adatbázis már tartalmaz morfoszintaktikai annotációt és szintaktikai elemzést is. Az annotáció során a szerkezet <FX></FX> tagek közé kerül, és jelölni lehet a szerkezet altípusát is. Jelenleg az üzleti hírek és az újsághírek annotációja készült el teljesen, a jogi szövegeke annotációja folyamatban van, azonban terveink szerint a teljes korpusz anyagára kiterjesztjük az annotációt.

A félig kompozicionális szerkezetek a prototipikus főnév + ige mintán kívül előfordulhatnak más szintaktikai mintázatban is, például igenévi alakban vagy főnévi (képzett) változatban. A korpuszban az alábbiak szerint vannak megjelölve a különféle altípusok (példákkal illusztrálva):

Főnév + ige kombinációja <verb>: *bejelentést tesz*

Igenevek <part>
Folyamatos melléknévi igenév: *életbe lépő (intézkedés)*
Befejezett melléknévi igenév: *csődbe ment (cég)*
Beálló melléknévi igenév: *fontolóra veendő (ajánlat)*
Főnévi igenév: *forgalomba hozni*
Határozói igenév: *ajánlatot téve*
Igei igenév: *(jogszabály) adta lehetőség*

Főnévi változat <nom>: *bérbe vétel*

Előfordulhat, hogy a főnévi és az igei komponens nem egymás mellett fordul elő a mondatban. Ezeket az eseteket is jelöljük, és a <split> altípusba soroljuk őket:

Különálló szerkezet <split>: *előadást fog tartani*

Mivel a Szeged Treebank már eleve tartalmaz szintaktikai annotációt, a félig kompozicionális szerkezetek jelölése során figyelembe vesszük a frázishatárokat is: a szerkezet főnévi komponensének legkülső határát jelöljük meg mint a szerkezet részét, nem csak pusztán a főnévi fejet. Ennélfogva a főnévi komponens esetleges jelzői is bekerülnek a szerkezetbe:

<FX>**nyilvános** ajánlatot tesz</FX>

A melléknévi igeneves alakban előforduló szerkezetek esetében pedig könnyen előfordulhat, hogy a szerkezetben más NP is szerepel:

<FX>**Nyíregyházán** tartott ülésén</FX>

A tárgyesetű főnévi komponenst tartalmazó szerkezetek nominalizációja kétféleképpen is történhet: összetett szóval, illetve birtokos szerkezettel:

<FX>szerződéskötés</FX>
<FX>adásvételi szerződések megkötése</FX>

A korpuszban mindkét típust jelöljük.

3 Statisztika

Az adatbázis jelenlegi formájában 407 félig kompozicionális szerkezetet tartalmaz 1745 előfordulásban az alábbi eloszlásban:

1. táblázat: A félig kompozicionális szerkezetek száma típus szerint.

	verb	**part**	**nom**	**split**	**összesen**
üzleti hírek	565	270	90	40	965
újsághírek	205	92	31	24	352
összesen	770 58.5%	362 27.5%	121 9.2%	64 4.8%	1317 100%

4 A korpusz hasznosíthatósága

A korpusz eredményesen használható mint tanító adatbázis a szerkezetek gépi úton történő azonosításához, melynek nyomán a különféle számítógépes nyelvészeti alkalmazások – például gépi fordítás és információkinyerés – pontossága is javulhat.

A gépi fordítás során a programnak először is fel kell ismernie, hogy az adott főnév és ige összetartozik (egy kollokáció két részét alkotják), továbbá – mivel egy adott szerkezet és idegen nyelvű megfelelője esetében a főnévi komponens megegyezik (azaz általában szó szerint fordítható), míg az ige eltérő [6] – a fordítóprogram az adott főnévhez társított megfelelő igét egy célnyelvi tanulókorpusz alapján készített gyakorisági mutató segítségével tudja kiválasztani.

Információkinyerésnél, különösen relációk kinyerésekor rendkívül fontos a mondatok megfelelő szintaktikai elemzése. A félig kompozicionális szerkezetek főnévi komponensének és a szerkezet egyéb vonzatainak szintaktikai státusa azonban vitatott [7]. Információkinyerés szempontjából a komplex predikátum feltételezése a legígéretesebb, azaz a szerkezetet egy egységként kezeljük, és ennek vannak vonzatai. Így például *A cég bérbe vette a raktárt* mondatból kinyerhető viszonyok a következők: **bérbe vétel** esemény, szereplői: **a cég**, **a raktár**. Ezzel szemben, ha az elemző nem ismeri fel a félig kompozicionális szerkezetet, így a főnévi komponens különleges szintaktikai státusát sem, a következő (helytelen) eredményt adja: **vétel** esemény, szereplői: **a cég**, **bér**, **a raktár**. Az elemző program betanítására szintén jól használható a létrehozott korpusz.

Köszönetnyilvánítás

A szerző köszönetet mond Szarvas Györgynek az annotációs eszköz kifejlesztésében nyújtott önzetlen segítségéért.

A kutatást – részben – a TUDORKA és MASZEKER programok keretében az NKTH támogatta.

Hivatkozások

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) Proceedings of Conference on Intelligent Text Processing and Computational Linguistics 2002. Mexico City (2002)
2. Kaalep, H.-J., Muischnek, K.: Multi-Word Verbs in a Flective Language: The Case of Estonian. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Context. Trento, Italy (2006) 57-64
3. Kaalep, H.-J., Muischnek, K.: Multi-Word Verbs of Estonian: a Database and a Corpus. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco (2008) 23-26
4. Krenn, B.: Description of Evaluation Resource – German PP-verb data. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco (2008) 7-10
5. Csendes D., Csirik J., Gyimóthy T., Kocsor A.: The Szeged Treebank, in Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005), Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123-131
6. Vincze V.: Angol–magyar főnév + ige szerkezetek és igei párjaik. In: Váradi T. (szerk.): II. Alkalmazott Nyelvészeti Doktorandusz Konferencia. Budapest: MTA Nyelvtudományi Intézet (2009) 113-123
7. Alonso Ramos, M.: Towards the Synthesis of Support Verb Constructions. In: Wanner, L. (ed.): Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk. Benjamins, Amsterdam / Philadelphia (2007) 97-138