Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű szövegelemzőben

Miháltz Márton[1], Indig Balázs[2], Prószéky Gábor[1,2,3]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport, 1083 Budapest, Práter utca 50/a
[2] PPKE Információs Technológiai és Bionikai Kar, 1083 Budapest, Práter utca 50/a
[3] MorphoLogic, 1122 Budapest, Ráth György u. 36.
mmihaltz@gmail.com, indig.balazs@itk.ppke.hu,
proszeky@morphologic.hu

1 Bevezetés

Az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport célkitűzései közé tartozik egy olyan, pszicholingvisztikailag motivált nyelvfeldolgozó rendszer kifejlesztése, amely nyers magyar nyelvű szövegből képes szintaktikai és szemantikai reprezentációt felépíteni [6]. Ennek egyik kulcsfontosságú lépése a természetes nyelvi mondatokban található ige-vonzat viszonyok felismerése és osztályozása (nyelvtani szerepek, tematikus szerepek). Ebben a cikkben bemutatjuk jelenleg zajló munkánkat, melynek célja az MTA-PPKE szövegelemző konstrukciós tudástárának bővítése. A munkához igyekszünk felhasználni és újrahasznosítani több, korábban kifejlesztett nyelvi erőforrás anyagát is.

Az elemző egyik alapvető felépítési elve a párhuzamosság: több különböző „erőforrásszál" egymást felülbírálva-javítva dolgozik. Ezek az erőforrásszálak a kategoriális nyelvtanokéra [3] emlékeztető mechanizmussal működnek: a nyelvi egységek közötti viszonyokat „ajánlások" és „elvárások" párhuzamos, ún. „strukturális szálai" közötti megfeleltetések teremtik meg. Ebben a paradigmában a mondatbeli potenciális vonzatok (névszói csoportok) „felajánlásokat" fogalmaznak meg (lexikai, morfológiai és szemantikai tulajdonságokat), melyek a mondatbeli igék által előhívott vonzatkeretek vonzatpozíciónak „elvárásaival" (megkötések az előbbi típusú tulajdonságokra) tudnak összekapcsolódni [8].

A továbbiakban először röviden bemutatjuk nyelvi elemzőnk igei vonzatkereteket kezelő mechanizmusát, majd felvázoljuk annak lehetőségét, hogyan lehet az elemzőben használt vonzatkeret-adatbázist tematikusszerep-leírásokkal kibővíteni az angol VerbNet erőforrás felhasználása segítségével.

2 Igei szerkezetek elemzése

Az igei vonzatszerkezetek elemzésének támogatására felhasználjuk a *MetaMorpho* magyar–angol (és angol–magyar) fordítóprogram [7] névszói és igei adatbázisának környezetfüggetlen, jegystruktúrás újraíró szabályait. A szabályok egy részére jel-

lemző, hogy bennük *minden* jobboldali szimbólum tartalmaz lexikális megkötést. Ezek a szabályok írják le a lexikont, vagyis a főnévi, melléknévi és határozószói egy- vagy többszavas kifejezéseket és ezek szemantikai és egyéb tulajdonságait [5,10], pl. „házőrző kutya”: <főnév, megszámlálható, élőlény>, „tegnapelőtt”: <időhatározó>, „hindi”: <főnév/melléknév, nyelv>.

A számunkra érdekes MetaMorpho szabályok másik csoportja nem minden (de legalább egy) jobboldali szimbólumra tartalmaz lexikális megkötést, míg a többi összetevőkre csak szófaji, morfológiai, szemantikai stb. feltételeket. Ezek közé tartoznak az igei vonzatkereteket leíró szabályok, melyeknél legalább az igei pozíció kötött lexikálisan, a vonzatpozíciók közül legfeljebb csak az idiomatikus igemódosítók (pl. „<valaki> jövendöl <valamit> <valakinek>” vs. „*szó* esik <valamiről>”).

Elemzőnk működése során szigorúan balról jobbra halad. Minden újabb mondatbeli token megjelenése a már ismert tokenek kapcsolati rendszerének (az elemzést reprezentáló gráf) időszerűsítéséhez (kiegészítéshez, és amennyiben szükséges, módosításához) vezet. A HuMor elemző felhasználásával minden tokent ellátunk morfológiai annotációval, a lehetséges elemzési szekvenciák közül egy, a PurePoS tagger [4] elvén működő, csak baloldali kontextust figyelő komponensünk választ, amely csak lokális, pontozott elemzési lehetőségeket produkál, Viterbi beam search nélkül. Az N legjobb elemzésen ezután párhuzamosan futtatjuk nyelvi elemzőnk központi elemző komponensét [8].

Az ige–vonzat viszonyok azonosítása a mondat feldolgozása során a keresletkínálat elv szerint történik. Névszói tokenek megjelenésekor ezekre szófaji és morfológiai tulajdonságaik segítségével igyekszünk új lexikális szabályokat illeszteni, vagy ezekkel korábban megkezdett többszavas mintákat folytatni. Ha egy egy- vagy többszavas névszói kifejezést teljesen felismertünk, megkíséreljük vele a mondatban korábban már szerepelt igék betöltetlen vonzatpozícióit „kielégíteni”. Igei tokenek megjelenése az igetőhöz tartozó lehetséges vonzatkeretek betöltésével jár, melyek vonzatpozícióit az elemző ekkor megkísérli „kielégíteni” az addig a pozícióig a mondatban már szerepelt (és teljesen befejezett) névszói elemek „felajánlásaival”.

3 Tematikus szerepek azonosítása

Az ige-vonzat viszonyok azonosításán túl a vonzatkeretek alkalmasak azok jellemzésére is. A szemantikai reprezentáció kialakítása szempontjából hasznos **tematikusszerep-leírások** azonban csupán a MetaMorpho magyar vonzatkeret-leírások mintegy 10%-ához állnak rendelkezésre (ezeket egy, a MetaMorpho fejlesztéséhez kapcsolódó független projekt, történelmi szövegek narratív pszichológiai elemzésének támogatásához készítettek [9]. Ugyanakkor a vonzatkeret-leírások fontos tulajdonsága, hogy angolul és magyarul is tartalmazzák e szabályok megfelelőit: minden forrásnyelvi (magyar) elemző szabályhoz rendelkezésre áll egy célnyelvi (angol) generáló szabály is, amely az adott nyelvi konstrukció fordítása. Ezért lehetőség van arra, hogy szabadon hozzáférhető, angol nyelvű erőforrásokra támaszkodva először a szabályok angol, majd a későbbiekben a szabályok magyar oldalát is kibővítsük az angol erőforrás MetaMorpho szabálypárokhoz kapcsolásával és az angol–magyar megfeleltetések helyes kezelésével (linked resource).

Ilyen erőforrás például a SemLink projekt [2] termékeként előállt egységes angol lexikai adatbázis, ami a *VerbNet* igei szótár , a PropBank szintaktikai és szemantikai jegyekkel annotált korpusz és a FrameNet szemantikaikeret-adatbázis összekapcsolása, amelyben az angol igék vonzatkeretei, szintaktikai és szemantikai információi is elérhetők megfelelő minőségben. Célunk a MetaMorpho szabályminták ezen egységes erőforráshoz kapcsolása és a MetaMorpho angol nyelvű, igei vonzatkereteket leíró szabályainak tematikus szerepekkel való minél teljesebb automatikus annotálása. Ezt követően megvalósíthatjuk az így nyert információk hatékony átvitelét a magyar nyelvű vonzatkeret-elemekre, kibővítve a magyar nyelvi elemzéshez rendelkezésre álló erőforrást. Fontos, hogy ez olyan módon történjen, hogy minél inkább megkönnyítse a magyar nyelvű szabályok emberi javítását, hogy a tisztán emberi módszerekkel létrehozott erőforrás minősége ne romoljon.

A következő részben bemutatjuk a MetaMorpho és a VerbNet összekapcsolásának és a tematikus szerepek gépi úton történő átvitelének kezdeti gyakorlati problémáit.

4 MetaMorpho és VerbNet vonzatkeretek megfeleltetése

Az összekapcsolás megvalósítása során figyelembe kellett venni, hogy az erőforrások számtalan ponton különböznek és ezeket az eltéréseket egységesíteni kell. Mindkét erőforrás tartalmazhat hibákat, amik befolyásolhatják a párhuzamosságok megtalálásának sikerességét. A MetaMorpho rendszer szabályainak készítői csak lazán voltak szabványokhoz és konvenciókhoz kötve, részben saját elképzeléseik alapján is dolgoztak, írásos dokumentáció a fejlesztési elvekről nem maradt fenn. Vizsgálataink során kiderült, hogy jópár elütésből és emberi hibából származó elem is található a szabályok között. Az elírási hibák egy része természetesen helyesírás-ellenőrzővel javítható automatikusan, de az előzetes tesztek azt mutatták, hogy sokszor ritka szavakról van szó, amit a helyesírás ellenőrző sem ismer.

Egy másik lényeges probléma, ami megnehezíti a két erőforrás harmonizálását az amerikai és a brit nyelvváltozatok írásmódjának kérdése: míg a MetaMorpho-t az eredeti szándékok szerint a brit angol ortográfiájának megfelelően fejlesztették, ezzel szemben a VerbNet az amerikai angol írásmódját követi.

A VerbNet összesen 6343 igét tartalmaz, ebből 2057 ige csak felsorolásszinten van jelen, mivel a többi, VerbNettel összekapcsolt erőforrásban előfordul. Az ilyen igéknek nem volt vonzatkeret-információja, ezért nem tudtuk őket a kutatás jelen állapotában hasznosítani. A maradék 4286 ige közül, melyekhez van vonzatkeret-információ, 2957 ige csak egy vonzatosztályban szerepel.

További problémát okozott az összetett igék kezelése. Az angol WordNetben összesen 7440 igéből 1410 *phrasal verb*, ami 549 igetövet érint. A VerbNetben 404 darab többszavas kifejezés található, melyekben 223 ige szerepel. A *MetaMorpho* 30 292 igei vonzatkeretes szabályába 3505 egyedi angol igető tartozik, amiből 920 darab nincs benne a VerbNetben (ebből 143 a helyesírás-ellenőrző számára hibás, illetve feltehetőleg ismeretlen szó). A vonzatkeretek és az angol igetövek száma közötti közel 10-szeres MetaMorpho-beli különbség egyik oka az, hogy kicsit több mint a szabályok egyharmadában idiomatikus vagy más lexikális megszorítás található (10 694 angol, 8347 magyar vonzatkeretben). Másfelől a magyar–angol

fordítórendszer fejlesztésekor nem volt cél, hogy a célnyelvi (angol) igék fedése jó legyen, elég volt a célnyelvi nyelvi hűség, jó fedésre a forrásnyelven (a magyar oldalon) volt inkább szükség. Szem előtt kell tartanunk, hogy ez a tulajdonság később még okozhat problémákat.

Az eddigieket figyelembe véve 2600 olyan egyedi ige van, ami mindkét erőforrásban szerepel és a VerbNetben osztályozva van. Ezekből 1545 csak egy, 622 kettő és 246 három VerbNet osztályba is beletartozik, továbbá van 10 darab olyan ige, ami 7 osztályba is be van sorolva. Ebből látszik, hogy az igék mintegy 42%-a esetén még az osztályok egyértelműsítésére is szükség van az ige minden egyes MetaMorpho-beli előfordulásánál.

A VerbNetben minden igeosztály tartalmaz egy vagy több, a VerbNet formalizmusában megadott szintaktikai vonzatkeret-leírást. Fő célunk ezeknek a VerbNetes vonzatkereteknek az egyértelmű azonosítása a MetaMorpho szabályok szintjén és az argumentumok kölcsönös leképezése a két erőforrás között.

Az egyértelműsítés során segítségünkre voltak a két erőforrás által definiált jeghalmazok, amik leírják az egyes argumentumok megszorításait a szintaxis és a szemantika oldaláról egyaránt. Míg a VerbNet a COMLEX formalizmust [1] alkalmazza, addig a MetaMorpho egy saját szempontrendszert [5,7,10]. Ezeket a leíró rendszereket kellett feltérképeznünk, illetve egymásra leképeznünk a különféle előfordulásaikat.

A MetaMorpho angol szabályok jobboldalának formája a legegyszerűbb esetben a SUBJ TV[1] [OBJ] mintát (egyszerű intranzitív/tranzitív szerkezet) követi. Az ilyen típusú szabályok az összes szabály kétharmadát teszik ki (kb. 20 000 szabály). Ezek az esetek képezték vizsgálódásunk első lépését. Az ilyen típusú szabályok esetén csak a sorrend és a típus figyelembevételével sikerült 1658 egyértelmű és 2908 többértelmű párosítást találni. A többértelmű párosítások feloldhatónak látszanak, amennyiben figyelembe vesszük az argumentumokra előírt megszorításokat mindkét oldalon. A vizsgált alakú szabályok és a közülük megtalált párosítások közötti több mint 4-szeres különbség a prepozíciók eltérő kezelésének tudható be, ugyanis a *MetaMorpho* a prepozíciókat egy egységként kezeli az őket követő szerkezetekkel, míg a VerbNetben a prepozíció önálló egységet alkot. Az ilyen különbségek helyes kezelése még folyamatban van.

Vannak olyan esetek is, amikor maga a vonzatkeret nem ad egyértelmű leképezést, annak ellenére, hogy a tematikus szerepek egyértelműek, mivel a VerbNet-ben a szemantikai információk is fel vannak tüntetve (és a szintaktikai szerkezettel együtt adják a rendezés kulcsát) és bár a szintaxis szintjén megegyeznek, a szemantika szintjén többértelműség keletkezik, amit fontosnak tartottak jelölni. Ilyen például a *meet* ige: a VerbNet különbséget tesz szemantikai szinten, amikor két ember találkozik, illetve amikor egy csoport tagjai találkoznak, míg szintaktikai szinten két azonos vonzat keret áll rendelkezésre. Ezen két eset között géppel a rendelkezésre álló információk segítségével nem lehet dönteni.

[1] TV (transitive verb): az igének megfelelő szimbólum a vonzatkeretben.

5 Összefoglalás

Jelen cikkben bemutattuk kereslet-kínálat elvű nyelvi elemzőnk igei vonzatkereteket azonosító működését, valamint megvizsgáltuk a *MetaMorpho* és a *VerbNet* összekapcsolásának és a tematikus szerepek átvitelének lehetőségét gépi úton. Fontos, hogy ez olyan módon történjen, hogy minél inkább megkönnyítse a magyar nyelvű szabályok emberi javítását, hogy a tisztán emberi módszerekkel létrehozott erőforrás minősége ne romoljon. A lexikai információk és nyelvi tulajdonságok további kétirányú megosztását tervezzük az összekapcsolt erőforrások között, amennyiben valamelyikben változás (verzióváltás, formátumváltozás stb.) állna be. Az ehhez szükséges lépéseket a tervezés során figyelembe vesszük. A jövőben megvizsgáljuk továbbá a lehetőségét annak, hogy a tematikus szerepekkel bővített igei konstrukciók segítségével hogyan lehet pszicholingvisztikailag reális módon szemantikai struktúrát kapcsolni a szintaktikai elemzéshez.

Hivatkozások

1. Grishman, R., Macleod, C., Meyers, A.: COMLEX Syntax: Building a Computational Lexicon. In: Proceedings of Coling, Kyoto (1994)
2. Loper, E., Yi, Sz.-t., Palmer, M.: Combining lexical resources: Mapping between PropBank and VerbNet. In: Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg (2007)
3. Morrill, G.V.: Categorial Grammar: Logical Syntax, Semantics, and Processing. Oxford University Press (2010)
4. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (2013)
5. Orosz, K.: Főnevek szemantikai jegyei és kódolásuk a MetaMorpho projektben. In: Alexin, Z., Csendes, D., eds.: IV. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, (2006) 157–166
6. Prószéky, G., Indig, B., Mihältz, M., Sass, B.: Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In: Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, (2014) 79–90
7. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: Foundation for International Studies (2004) 138–142
8. Sass, B.: Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (2015), ld. jelen kötetben.
9. Vincze, O., Gábor, K., Ehmann, B., László, J.: Technológiai fejlesztések a NooJ pszichológiai alkalmazásában. In: VI. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged (2009) 285–294
10. Vincze, V., Lucza, M., Csendes, D., Kiss, G.: Szótárazási dilemmák a MetaMorpho magyar–angol névszói adatbázisának építésében. In: Alexin, Z., Csendes, D., eds.: IV. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged (2006) 180–189