28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat Két nagy méretű, magyar nyelvi erőforrást teszünk közzé. Az egyik a régi MNSZ [1] tagmondatainak sekély szintaktikai elemzéssel ellátott változata, mely a Mazsola [2] lekérdező adatbázisaként szolgál; a másik pedig az ebből az adatbázisból automatikusan származtatott igeiszerkezet-lista, melyből a Magyar Igei Szerkezetek című szótár [3] is született. Az erőforrások elérhetők a `http://corpus.nytud.hu/isz` címen.

Kulcsszavak: nyelvi erőforrás, szintaktikai elemzés, igei szerkezetek, Mazsola, méret

1. Az erőforrások létrehozása

A *Mazsola adatbázis* a Magyar Nemzeti Szövegtár 187 millió szavas régi változatának teljes anyagát tartalmazza, melyet a feldolgozás során tagmondatokra bontottunk és részleges szintaktikai elemzésnek vetettünk alá. Utóbbi során (1) megállapítottuk a tagmondat igéjét (főnévi igeneves szerkezet esetén a főnévi igenév a tagmondat igéje), és hozzákapcsoltuk az igéhez az esetleges odatartozó elváló igekötőt; (2) számba vettük az ige mellett felső szinten megjelenő névszói és névutói csoport bővítményeket (tehát a határozószói csoportokat például nem), ezeket a fej szótövével és esetével (értsd: esetragjával vagy névutójával) reprezentáltuk. A részleteket lásd [4] 2.2. fejezetében. A Mazsola [2] felületén keresztül lekérdezhető adatbázishoz képest a jelen adatbázis tartalmaz bizonyos javításokat, továbbfejlesztéseket: (1) a birtokos szerkezetek jobb kezelésének, valamint a főnévi igenév mellett *-nAk* raggal megjelenő alany funkciójú bővítmény alanyként való reprezentálásának köszönhetően csökkent a helytelen *-nAk* esetű bővítmények száma; (2) a *maga mögött* és a *mögöttem* típusú szerkezetek helyesen névutós névmásként elemződnek; valamint (3) szerepel egy további információ is az annotációban: hogy az adott bővítmény birtokos személyjeles-e.

Az *igeiszerkezet-lista* a fenti adatbázis alapján egy speciális igeiszerkezet-kinyerő algoritmussal automatikusan meghatározott igei szerkezeteket tartalmazza. Az algoritmus lényege, hogy a fent leírt reprezentáció szerinti mondatvázakat, igei kereteket azáltal összesíti, hogy a ritka (legfeljebb 5-ször előforduló) mondatvázakat egy rövidebb, illeszkedő mondatvázhoz rendeli hozzá; majd az eljárás végén lévő ellenőrző lépésben a túl általános mondatvázhoz került mondatokat a lehető legspecifikusabb meglévő mondatvázhoz helyezi át. A módszer

képes feltárni, hogy az adott esetragos bővítmény általában jellegzetes-e, illetve ezen túl azt is, hogy a bővítményi helyen megjelenő egyes konkrét tartalmas szavak tipikusak-e. Ennek megfelelően vonzatokat (*hisz vmiben*), kollokatív igei szerkezeteket (*süt (a) nap*, *döntés születik*), illetve a két eset kombinációjaként vonzatos komplex igéket (*szó van vmiről*, *igényt tart vmire*) egyaránt eredményez. Az igeiszerkezet-kinyerő módszer részletes bemutatása és kiértékelése [4] 3.3. fejezetében olvasható.

2. Az erőforrások formai leírása

A Mazsola adatbázis egy egyszerű szöveges fájl, sorainak formátumát az 1. ábra mutatja be.

```
engem meg sem hallgattak . stem@@meghallgat ACC@@én
A hasmenéstől szenvedő betegeknek sokat kell inniuk , stem@@iszik ACC@@sok NOM@@beteg
A Profi egyik támadójátékosa elhúzta mellettem a labdát , stem@@elhúz ACC@@labda mellett@@én ...
                                                    ... NOM@@támadójátékosPOSS
```

1. ábra. A Mazsola adatbázis sorainak felépítése

A tagmondat után következik a fentiek szerint elvégzett sekély szintaktikai elemzés eredményeként kapott reprezentáció: először – stem@@ után – az ige, majd eset@@fej_szótöve formában a névszói és névutói csoport bővítmények eset szerinti ábécésorrendben. Az igét nem tartalmazó tagmondatokban stem@@NULL, a határozott ragozású igét, de explicit tárgyat nem tartalmazó tagmondatokban pedig ACC@NULL szerepel. Látjuk az *engem*, *mellettem* elemzését, az igekötő igéhez kapcsolását, a főnévi igenév főigekénti kezelését, a főnévi igenév melletti -nAk ragos szó alanyként való értelmezését, az igeneves (*A hasmenéstől szenvedő betegeknek*) és a birtokos szerkezet (*A Profi egyik támadójátékosa*) egy egységként való kezelését (a POSS a birtokos személyjelet kódolja).

Az igeiszerkezet-lista szintén egy egyszerű szöveges fájl, soronként egy szerkezetet tartalmaz a 2. ábrán látható, szemléletesebb, ember számára jobban olvasható formában: a Mazsola adatbázisban szereplő szokásos hárombetűs esetrövidítések helyett itt az esetragok szerepelnek (önállóan vagy a tartalmas szavak végéhez kapcsolva); a névutókat egyenlőségjel jelzi; a birtokos személyjelet pedig -A. A két formátum szükség esetén egyszerűen átalakítható egymásba. A 2. ábrán a *csap* karaktersorozatot tartalmazó néhány példát látunk.

Minden sor egy igei szerkezetet és egy gyakorisági mérőszámot tartalmaz. Az első elem mindig az igető, utána következnek a névszói és névutói csoport bővítmények. A fent leírt kinyerési módszernek köszönhetően a bővítmények között egyaránt megjelennek a szabad esetrag/névutó által képviseltek és a konkrét szóval kitöltöttek is. A fenti mér csapás-t -rA szerkezet mindkét esetet példázza: a mér komplex igét alkot a konkrét szóval kitöltött tárggyal, és ehhez a kéttagú szerkezethez járul még egy -rA ragos vonzat. A kitöltött alanyi bővítménnyel nem bíró szerkezetekhez az alanyi bővítményt implicite mindig odaértjük. A kinyerő algoritmus által szolgáltatott gyakorisági mérőszám

becsap -t 1248
lecsap -rA 620
mér csapás-t -rA 360
átcsap -bA 345
megcsappan 217
lesz csapadék 205
csap -t hón-A=alá 80
becsap ajtó-t maga=mögött 28
átcsap =fölött 20

2. ábra. Az igeiszerkezet-lista sorainak felépítése

jelentése: ennyi olyan mondat volt a korpuszban, ami megfelel az adott szerkezetnek, és nincs olyan specifikusabb szerkezet a listán, aminek megfelelne. Következésképpen ha azon mondatok számára vagyunk kíváncsiak, amikben például a *becsap* ige mellett van tárgy, akkor össze kell számolni a lista összes olyan bejegyzését, amiben ez a két elem (*becsap* + tárgy) szerepel.

3. Mennyiség és minőség

Ahogy a cím is kiemeli, igen jelentős méretű erőforrásokról van szó: ez pontosan 27970403 sekély elemzéssel ellátott tagmondatot és 535609 igei szerkezetet jelent. Mindkét mennyiség egyedülállónak mondható a magyar nyelv tekintetében. A szótárral [3] összevetve azt látjuk, hogy az igeiszerkezet-lista két nagyságrenddel bővebb anyag (a szótár csak a 250-nél nagyobb gyakorisági mérőszámmal bíró 6266 szerkezetet tartalmazza), ugyanakkor tisztítatlan, nyers adat, érvényesek rá a szótár bevezetőjében említett korlátok [3, 9-17. oldal] és természetesen nélkülözi a szótári példamondatokat, illetve mutatókat. Összevetettük az igeiszerkezet-listát egy kézzel annotált, gold sztenderd korpuszból származó félig kompozicionális szerkezeteket tartalmazó listával[1] [5] is. Azt látjuk, az igeiszerkezet-lista (a más típusú, illetve kompozicionális szerkezetek mellett) nagy mennyiségű félig kompozicionális szerkezetet tartalmaz. A nagyobb korpuszméret a gyakoriságok jobb becslésére ad lehetőséget. Kiemelendő, hogy az igeiszerkezet-listán a teljes szerkezetek (is) szerepelnek, azaz nemcsak a komplex igék, hanem a hozzájuk tartozó vonzatok is megjelennek: a *zsebre vág* szerkezetet vág -t zseb-rA formában, azaz a tárggyal együtt találjuk meg.

Tudni kell, hogy a Mazsola adatbázis bemutatott sekély szintaktikai elemzése részletesség és hibamentesség tekintetében nem közelíti meg a kézzel készített elemzések minőségét [6], ugyanakkor az erőforrás a nagy méret miatt fontos előnyös tulajdonsággal bír: a nagy korpusz lehetőséget ad a ritka jelenségek, szerkezetek jellemzésére [7, 323. oldal]. Emiatt és a kinyerő módszernek köszönhetően, a nem hibátlan elemzés ellenére van lehetőség olyan ritkább szerkezetek felfedezésére, azonosítására és gyakoriságának becslésére, mint a visz prím-t -bAn, ter-

[1] http://rgai.inf.u-szeged.hu/project/nlp/research/mwe/fx_list_hu.txt

jeszt rémhír-t, telik erő-A-bÓl -rA vagy tapos -t sár-bA. Az igeiszerkezet-lista a Mazsola adatbázis elemzési hibái ellenére képes megbízható adatokat szolgáltatni az igei szerkezetekről. A Mazsola adatbázis alapvetően az igeiszerkezet-lista elkészítése érdekében jött létre, ugyanakkor hasznosnak gondoljuk erőforrásként önmagában is közzétenni a további felhasználás érdekében. A fentiek is mutatják a kis plusz hozzáadott információt tartalmazó (például a fenti sekély elemzéssel ellátott), de nagy méretű korpuszok hasznosságát, összevetve akár a még sokkal nagyobb POS-taggelt, akár a kisebb méretű gazdag annotációval bíró korpuszokkal.

4. Példák

Alább néhány példával világítjuk meg, hogy mi mindent tartalmaz az igeiszerkezet-lista, és mire lehet alkalmas. Mint említettük, az igei szerkezetek kinyerése gyakorisági alapon történik. Ennek következtében az idiomatikus kollokációk (komplex igék) mellett megjelennek a listán az igével kompozicionális szerkezetet alkotó gyakori szavak is, a vonzatok mellett pedig az egyéb bővítmények is (eset/névutó által képviselve). Jól látszik ez, ha egy gazdag vonzatszerkezettel bíró igét vizsgálunk meg. Nézzük a *száll* legjellegzetesebb szerkezeteit a 3. ábrán.

1. száll -rA 610	11. száll =mellett sík-rA 94
2. száll 463	12. száll vonat-rA 80
3. száll vita-bA -vAl 359	13. száll maga-A-bA 72
4. száll -bA 292	14. száll -n 71
5. száll -ért sík-rA 150	15. száll sík-rA 69
6. száll -ért harc-bA 142	16. száll -bA -vAl 67
7. száll -bAn 141	17. száll -ért ring-bA 65
8. száll -vAl 134	18. száll part-rA 64
9. száll ring-bA 103	19. száll harc-bA 63
10. száll fej-A-bA 101	20. száll -rÓl -rA 61

3. ábra. A *száll* első húsz szerkezete

A 18. szerkezet (száll part-rA) tipikus komplex ige, a 12. (száll vonat-rA) talán kevésbé idiomatikus, mindenesetre itt a bővítményi helyen egyéb szavak is megjelenhetnek (villamos, busz, hajó), ahogy ez a lista további részéből kiderül. Látjuk, hogy ezek a szavak egy szemantikailag koherens osztályt alkotnak, jelen esetben a (tömeg)közlekedési eszközökét. Ilyen szóosztályokkal általában akkor találkozunk, ha egy igének egy vonzati helyén jelennek meg az odaillő, literális jelentésű szavak (vö: az eszik tárgyaként megjelenő különféle ételek). Az is gyakori megfigyelés, hogy az ilyen szemantikailag koherens osztályokból kakukktojásként ugranak ki a komplex igék, idiómák, szólások, mint például az eltörik alanyaiként szereplő testrészek közül a mécses. Vonzatra példát itt a komplex igék mellett látunk: száll vita-bA -vAl, száll sík-rA/harc-bA/ring-bA -ért, illetve

sík-rA =mellett. Az ige mellett megjelenő -bAn, -n stb. esetek különféle szabad határozók jelenlétére utalnak. Az effajta gyakori esetragok a szabad határozók miatt lényegében minden ige mellett megjelennek, vonzati funkciójukra a sokkal prominensebb megjelenés utal, például a szerepel esetében a kiemelkedően magas gyakorisági mérőszámmal bíró szerepel -bAn. A száll fej-A-bA alanyaként a teljes listában a dicsőség, vér és ital szavakat találjuk. E három szó nagyjából meg is adja azt a három fogalmi kört, ami itt előfordulhat, ez a Mazsola adatbázison ellenőrizhető. A legelöl álló száll -rA szerkezet nagyon heterogén, több különböző jelentésű szerkezetet foglal magába. A lejjebb lévő specifikusabb szerkezetek utalnak rá, hogy miféléket, de ahogy a gyakorisági mérőszám meghatározásánál erről volt szó, az itt lévő 610-es érték csakis olyan tagmondatokból állt elő, melyek mondatváza a listán szereplő egyéb szerkezetekre nem illeszkedik.

Az erőforrás hasznos lehet a vonzatok kötelezőségével foglalkozó vizsgálatokban. A listán sok olyan szerkezetpárral találkozunk, hogy az egyiket a másikból egy bővítmény/vonzat elhagyásával kaphatjuk meg. Ez arra utalhat, hogy az adott vonzat nem kötelező, elhagyható, vagy – és ez a két eset pusztán a lista alapján nem különíthető el – hogy a szerkezet sok esetben elliptikusan manifesztálódik. A felszólít, felkér és tanít esetében a sima tárgyas keret gyakoribb, mint a -t -rA keret, ez a nem kötelező -rA ragos vonzat vagy bővítmény gyanúját veti fel; a bíz, kényszerít és alapoz esetében fordított a helyzet, ekkor kötelező -rA ragos vonzatot sejthetünk.

Adott bővítményi szavakat vizsgálva megkapjuk a szót tartalmazó jellegzetes igei szerkezeteket. A *vagyon* esetében például a *rendelkezik*, *szert tesz*, *felél*, *megfoszt*, *gyarapít*, *elkoboz*, *kiforgat*, *felhalmoz* igékkel együttállva; a *tej* esetében többek között a *kifut (a) tej* vagy az *aprít (a) tejbe vmit*; a *kenyér* esetében pedig *eszik/süt/szel kenyeret*-től a *vmivel keresi (a) kenyerét*-en át a *visszadob kenyérrel*-ig.

5. A közzététel módja

A bemutatott két erőforrást oktatási, kutatási és magáncélra – az üzleti felhasználás külön megállapodás tárgyát képezheti – szabadon letölthetővé tesszük a `http://corpus.nytud.hu/isz` címen. A pontos felhasználási feltételek a honlapon olvashatók. A Mazsola adatbázist alkotó tagmondatokat ábécérend szerint, az igei szerkezeteket pedig gyakoriság szerint rendezve közöljük. Terveink szerint az erőforrások később a META-SHARE repozitóriumba is be fognak kerülni.

Néhány megjegyzés a közzététel és a szabad hozzáférés kapcsán. Van olyan álláspont [8, 4. rész], miszerint a weben szabadon elérhető anyagok korpuszépítési célú felhasználása lényegében korlátozás nélkül megengedett, főleg, ha feldolgozott, származtatott erőforrásról van szó. Ennél óvatosabb az a megközelítés, mikor az eredeti szöveg visszaállítását lényegében lehetetlenné téve ábécérendbe teszik a korpusz mondatait [9, 1. rész, „Literary texts”]. Az által, hogy esetünkben az alapegység a tagmondat, még egy lépéssel továbbmegyünk a visszaállíthatóság csökkentésében, így eljárásunk semmilyen értelemben nem tekinthető az MNSZ-ben lévő művek újraközlésének.

Azon túl, hogy a Mazsola korpuszlekérdező, illetve a Magyar igei szerkezetek szótár létrehozása során közvetlenül a bemutatott erőforrásokra építettünk, más kutatások is használták már azokat [10,11] most pedig megnyílik a lehetőség a széleskörű felhasználás előtt.

Hivatkozások

1. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
2. Sass, B.: ,,Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest (2009) 117–129
3. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
4. Sass, B.: Igei szerkezetek gyakorisági szótára - egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD thesis, PPKE ITK (2011)
5. Vincze, V., Csirik, J.: Hungarian corpus of light verb constructions. In: Proceedings of COLING 2010, Beijing, China (2010) 1110–1118
6. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In Matoušek, V., ed.: Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005). Springer (2005) 123–131 Springer LNAI 3658.
7. Kornai, A.: Probabilistic grammars and languages. Journal of Logic, Language, and Information (20) (2011) 317–328
8. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3) (2009) 209–226
9. Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., eds.: Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05. John Benjamins (2007) 247–258
10. Miháltz, M., Sass, B., Indig, B.: What do we drink? Automatically extending Hungarian WordNet with selectional preference relations. In: Proceedings of Joint Symposium on Semantic Processing, Trento (2013) 105–109
11. Pléh, Cs., Németh, K., Varga, D.: The possible role of entropy in processing argument dependencies in Hungarian. In: 16th International Morphology Meeting, Information Theory in Morphology workshop. (2014)