Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból

Novák Attila, Laki László János, Novák Borbála

Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat Cikkünkben egy olyan algoritmust mutatunk be, amelynek segítségével elemzett angol-magyar párhuzamos korpuszból gyűjtöttünk idiomatikus és félig kompozicionális igei szerkezeteket a szómegfeleltetések felhasználásával. Mivel a kutatás kontextusát egy kérdések megfogalmazását lehetővé tevő elemző megalkotása jelentette, ezek a szerkezetek elsősorban abból a szempontból érdekeltek minket, amennyiben a szerkezet egyes elemeire a kompozicionalitás hiányából kifolyólag nem lehet kérdezni. Az eredményeket összevetettük egy ilyen szerkezeteket tartalmazó létező erőforrással, és azt találtuk, hogy algoritmusunk sok új számunkra érdekes nem vagy részben kompozicionális igei szerkezetet azonosított.

1. Bevezetés

A nem kompozicionális igei szerkezetek olyan igéből és főnévből álló kifejezések, ahol a kifejezés jelentése nem kiszámítható a szerkezet tagjainak jelentéséből. Ezek lehetnek teljesen idiomatikus kifejezések, de olyan szerkezetek is, ahol ugyan a kifejezés egyik vagy akár mindkét tagjának a jelentése fontos mozzanatot visz a komplex kifejezés jelentésébe, az utóbbi mégsem egészen a szokásos kompozicionális jelentéskombinációs műveletek eredményeként áll elő, hanem vagy tartalmaz valami pluszt, vagy valamelyik elem (rendszerint az ige) jelentéséből legfeljebb valamilyen mozzanat jelenik meg. A nem teljesen idiomatikus kifejezéseket szokás félig kompozicionális igei szerkezeteknek hívni. Bár a ma egyre inkább teret hódító neurális architektúrák kevésbé érzékenyek ezekre a szerkzetekre, kezelésük szinte minden nyelvtechnológiai feladatban külön figyelmet igényel.

Az ebben a cikkben bemutatott módszerrel angol-magyar párhuzamos korpuszból gyűjtöttünk ki félig vagy még kevésbé kompozicionális igei szerkezeteket. Ezekkel az volt a célunk, hogy egy fejlesztés alatt álló elemzőrendszerhez készített igei vonzatkeret-adatbázis építését és ezeknek a korpuszhoz való illesztését támogassuk [1].

2. A félig kompozicionális igei szerkezetek

Formailag a félig kompozicionális igei szerkezetekben egy ige és egy vagy több (ragozott, prepozíciós vagy névutós) névszói vonzat szerepel. A nyelvészetben számos csoportosítása ismert a félig kompozicionális, illetve idiomatikus szerkezeteknek olyan paraméterek mentén, hogy a kifejezés egyes tagjainak önálló szintaktikai és szemantikai szerepe mennyiben és hogyan járul hozzá a kifejezés egészének jelentéséhez.

Az egyik ilyen csoportosítás [2]-ben szerepel, ahol a szerző négy csoportot definiál:

- idiómák: ahol a kifejezés tagjainak jelentéséből egyáltalán nem számítható ki az egész jelentése, pl. *feldobja a talpát* 'meghal'
- mind az ige, mind a főnév eredeti jelentésükben járul hozzá a kifejezés jelentéséhez, de a kifejezés hordoz valami pluszt, pl. *iskolába jár* 'ott tanul'[1]
- idiómaszerű, de az idiómáknál szabadabb kifejezések, ahol az egyik tagra nem lexikális, hanem egy szemantikai kategóriát meghatározó megkötés van, pl. *főbe, hasba, lábon lő*
- azok az állandó fordulatok, ahol az ige jelentése nem jelenik meg a kifejezés jelentésében a maga teljességében, a vonzatszerű névszói elem apportálja a szemantikai tőke javát, az ige szinte pusztán a grammatikai kategóriáját viszi vásárra, jelentéstartalmából legfeljebb valamilyen mozzanat jelenik meg, pl. *lehetőség nyílik valamire*

A nemzetközi szakirodalomban is hasonló csoportosítások jelennek meg. Sag szerint [3] például a többszavas kifejezések két fő csoportot alkotnak: lexikalizálódott és intézményesült kifejezések. Lexikalizálódott kifejezéseknek azokat a kifejezéseket tekintik, amiknek a szintaktikai vagy szemantikai felépítése legalább részben idioszinkratikus, vagy olyan szavakat tartalmaznak, amik önmagukban nem fordulnak elő azzal a jelentéssel. Ezek a fajta kifejezések a lexikai kötöttségük szempontjából további alcsoportra bonthatók: teljesen kötött kifejezések, félig kötött kifejezések és szintaktikailag rugalmas kifejezések. A Sag rendszerezésében szereplő intézményesült kifejezések szintaktikailag és szemantikailag ugyan kompozicionálisak, de adott kontextusban az átlagosnál gyakrabban jelennek meg együtt.

Célunk nem az volt, hogy ezeknek az előre definiált csoportosításoknak megfelelő kifejezéseket gyűjtsünk, hanem egy saját kritériumrendszert állítottunk fel. Mivel a célunk egy olyan elemzőrendszer létrehozása, amely ténylegesen alkalmas arra, hogy releváns kérdéseket tegyen fel azzal a szöveggel kapcsolatban, amit feldolgoz [1], ezért a félig kompozicionális szerkezetek azonosítása során is ez volt a fő szempont.

Az idiomatikus és félig kompozicionális szerkezetek azonosításakor is azt a célt tartottuk tehát szem előtt, hogy egy kifejezés az arra vonatkozó releváns kérdés megfogalmazása szempontjából hogyan viselkedik. A *döntést hoz* kifejezés esetén nem jó kérdés a *Mit hoz?*, hacsak nem viccet szeretnénk csinálni belőle, pl.:

[1] A portás vagy a tanári kar hiába jár szintén oda, ők munkába járnak, nem iskolába.

```
- Mit hozott Édesapám?
- Döntést.
```

A kérdezés szempontjából ugyanakkor például az egyébként szintén nem kompozicionális *csinálja a fesztivált* kifejezés kevésbé tűnik érdekesnek, mert a minden ágenses igére használható *Mit csinál?* kérdést lehet ezzel kapcsolatban is feltenni. Az utóbbi esetben is megfigyelhető ugyanakkor, hogy némileg humoros hatást kelt a csak a tárgyat megnevező válasz: *Mit csinál? A fesztivált.*, ami egy kompozicionális ige-vonzat kapcsolat esetében teljesen normális lenne. Mindazonáltal a *csinál* ige a mi szempontunkból kevésbé tűnik „veszélyesnek", mint más idiomatikus vagy félig kompozicionális szerkezeteket alkotó igék.

Szintén nem érdekesek a mi szempontunkból az *iskolába jár, fát vág* típusú szerkezetek, ahol ugyan van valami jelentéstöbblet, mind az ige, mind a névszói vonzat jelentése viszonylag csorbítatlanul jelen van a kifejezés jelentésében, ezért nem ostobaság és nem vicc sem a vonzatra *(Mit vág?)* sem az igére *(Mit csinál a fával?)* kérdezni.

A szemantikailag kötött vonzatú igék esetében változatos a kép a kérdezés szempontjából. A *főbe/fejbe/hasba/ülepen/fenékbe/lábon lőtték* esetében nem lehet azt kérdezni, hogy *Mibe/min/hova lőtték?*, hanem csak a *Hol/melyik testrészén lőtték meg?* kérdés lehetséges. Tehát van két alternatív minta, mindkettőben testrészekre korlátozódik az egyik argumentumként megjelenő elemek köre, de az egyik minta (ahol egyébként a rag is változik az adott testrész függvényében) nem teszi lehetővé a testrészre kérdezést, a másik viszont (ahol a rag fixen a szuperesszívusz), lehetővé teszi azt. Látjuk tehát, hogy a szemantikai/lexikai kötöttség hol nyitva hagyja, hogy pedig nem teszi lehetővé az adott vonzatra kérdezést.

3. Módszer

A statisztikai gépi fordítás fénykorában számos, nem feltétlenül gépi fordítási feladatra, elkezdték alkalmazni a statisztikai gépi fordító rendszerek egyes alkotóelemeit. Az egyik ilyen alkotóelem a szómegfeleltetési modell (word alignment), ami a rendszer tanításához használt párhuzamos korpusz mondatain belüli szavakat megfelelteti egymással. Ez a megfeleltetés lehet $n : m$ vagy $m : n$, ahol $n \geq m$. A [4] cikkben a többszavas kifejezések szómegfeleltetési modell alapján történő azonosításának a következő definíciója szerepel:

Mivel a két nyelv közötti automatikus szómegfeleltetési modell a forrásnyelvi mondat szavainak ekvivalensét keresi a célnyelvi mondatban, ezért ha egy S forrásnyelvi szósorozat (ahol $S = s_1...s_n, n \geq 2$) megfeleltethető egy T célnyelvi szósorozatnak (ahol $T = t_1...t_m, m \geq 1$), azaz S és T egymás megfeleltetései, akkor feltételezhetjük, hogy S és T szemantikai tartalma legalább részben hasonló, és hogy S egy potenciális többszavas kifejezés. Más szóval ez azt jelenti, hogy az S szósorozat egy kifejezésjelölt, ha egy egy vagy több szóból álló T sorozatnak feleltethető meg a célnyelven (azaz egy $n : m$ típusú megfeleltetésről van szó, ahol $n \geq 2, m \geq 1$). Fontos tehát, hogy a forrásnyelvi kifejezés több szóból

áll, ami a célnyelven egy vagy több szónak felel meg. Az általunk megvalósított algoritmus ebből a definícióból indul ki, de az eredeti célunk érdekében további megszorításokat tettünk, amiket a továbbiakban részletezünk.

3.1. A korpusz előkészítése

A keresett félig kompozicionális illetve idiomatikus igei kifejezéseket tehát egy párhuzamos korpuszból szerettük volna összegyűjteni. Ehhez egy 644,5 millió token méretű angol-magyar párhuzamos korpuszt [5] használtunk. Mivel a célunk az volt, hogy nem vagy félig kompozicionális igei szerkezeteket gyűjtsünk egy igei vonzatkeret-adatbázis építéséhez, ezért csak azokra a kifejezésekre koncentráltunk, amikben igék szerepelnek. Ehhez először a párhuzamos korpusz mindkét oldalának elemzésére volt szükség. Az angol oldalt a Stanford taggerrel [6] szófaji egyértelműsítettük és a morpha [7] lemmatizálóval lemmatizáltuk. A magyar oldalon a PurePos [8] szófaji egyértelműsítőt és lemmatizálót alkalmaztuk, amely a Humor morfológiai elemző [9,10] elemzéseit használja. Ezek után mindkét oldalon úgy alakítottuk át az elemzett szöveget, hogy minden eredeti tokent két token reprezentál: (1) a lemma a fő szófajcímkével és (2) a szóhoz tartozó további morfoszintaktikai címkék.

Az alábbi példa a *Szeretlek, kedvesem. – I love you, dear.* mondatpár így előfeldolgozott változatát mutatja:

```
szeret[IGE]  [Ie1] ,[PUNCT] kedves[FN] [PSe1][NOM]
I#PRP love#VB [P] you#PRP ,#, dear#RB
```

Ez az átalakítás veszteségmentesen megőrzi az eredeti szóalakokra vonatkozó információkat, így a végeredményben visszaalakíthatóak a szóalakok. Ez azért is fontos, mert sok esetben a félig kompozicionális kifejezésekben egy-egy szónak csak bizonyos alakja(i) megengedett(ek), tehát nem lenne elegendő a lemma azonosítása. Ugyanakkor, a kifejezések ilyen formában nem kötött részeire robosztusabb statisztikát kapunk a lemmák egységes kezelése miatt. Fontos szempont volt továbbá, ahogy már említettük, az igék azonosítása, amit ez az átalakítás szintén lehetővé tett.

3.2. A szómegfeleltetési modell létrehozása

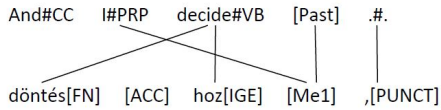
Az átalakított párhuzamos korpuszból a szómegfeleltetési modell létrehozásához a fast align programot [11] használtuk. Ennek kimenetéből a Moses SMT-készlet [12] frázistábla-építőjével készítettünk maximum 7-gram méretű frázispárokat. Bár a Moses különböző pontszámokat is rendel a megfeleltetett kifejezésekhez, amik többek között az adott fordítás feltételes valószínűségét jelzik, ezeket a pontszámokat végül nem használtuk fel, csupán magát a megfeleltetést, hogy melyik szavakat melyikhez rendelte hozzá.

Ezen kívül további szűréseket alkalmaztunk. A kapott listából kidobtuk azokat a frázispárokat, amikben nem szerepelt ige, vagy több ige szerepelt valamelyik oldalon, illetve azokat is, amikben mondat- vagy tagmondathatárra utaló jel (írásjelek) szerepeltek. Mivel a megfeleltetésben szereplő frázisok n-gram

alapúak, ezért ezek a szűrések nem jártak információveszteséggel, hiszen még a kidobott frázisok releváns részletei is megmaradtak másik frázis részeként. A szűrésnek köszönhetően viszont nem kerültek az eredménybe olyan hamis pozitív kifejezések, amik átlépnek például tagmondathatárokon. Hasonlóképpen, azoknak a frázisoknak az igéit tartalmazó részfrázisok, amikben több ige szerepelt, megjelentek másik frázisokban.

3.3. A félig kompozicionális szerkezetek kigyűjtése

A kifejezésjelölteknek a szűrés után megmaradt frázispárlistából való kiszűrése során azzal a feltételezéssel éltünk, hogy a magyar oldalon szereplő félig kompozicionális szerkezetek minden eleme az angol oldalon szereplő igéhez van kötve. Például ha az angol oldalon a *decide* ige szerepel, a magyar oldalon pedig a *döntést hoz* kifejezés, akkor ezek kötött áll fenn a megfeleltetés (1. ábra).



1. ábra. Példa egy szómegfeleltetésre az angol és magyar elemzett frázisok elemei között.

Az ilyen típusú megfeleltetéseket magyar igénként külön-külön kigyűjtöttük és az egyes gyűjteményekben szereplő angol igékhez összesítettük, hogy melyik angol igéhez melyik magyar ige milyen főnévvel szerepel. Ekkor már a magyar főneveket egyesítettük az utánuk külön tokenként szereplő morfoszintaktikai címkéjükkel és visszaalakítottuk az eredeti formájukra a Humor generátor funkciójával [9,10]. Ez a lista már önmagában érdekes abból a szempontból, hogy az angol igék a magyar felsorolásban szereplő kifejezéseknek tulajdonképpen mint egyszavas definíciói jelennek meg. Az 1. táblázatban a magyar *hoz* igéhez összegyűlt néhány angol ige és a szómegfeleltetés alapján a *hoz* mellett az adott igéhez kötött főnév visszaalakított alakjai láthatók. Így például a *fix* angol ige tulajdonképpen definiálja a *rendbe hoz, helyre hoz, működésbe hoz* magyar kifejezéseket. Ebbe a sorba bekerült a *dolgokat* vonzat is, ami a *rendbe hozza a dolgokat* kifejezésből ered, de mivel a szintaktikai viszonyokat nem vizsgáltuk, ezért nem tudunk különbséget tenni a különböző típusú vonzatok között. Az is látható a példából, hogy az algoritmusnak ezen a pontján olyan angol igékhez tartozó listák is létrejönnek, ahol az angol oldalon is félig kompozicionális igei szerkezet szerepel. Például a *make* ige esetén a *make a decision* a *döntést hoz* párja. Látható továbbá még az is, hogy a korpuszban megjelenő elírások, nem sztenderd szóalakok is megjelentek a listán.

Az algoritmus további lépéseiben az ilyen formán definícióként szereplő angol igékkel nem foglalkozunk, de a megfelelő tisztítás után jó alapanyaga lehet egy olyan szótárnak, amiben az összegyűlt többszavas kifejezések szerepelnek.

angol ige	magyar főnevek
fix	rendbe, helyre, dolgokat, működésbe, stb.
scare	frászt, szívbajt, szívinfarktust, fraszt, szivbajt, stb.
make	döntést, változást, nyilvánosságra, hasznot, áttörést, világra, stb.
embarrass	zavarba, helyzetbe, szégyent, szégyenbe, stb.
freak	frászt, szívbajt, szivbajt, fraszt, stb.
connect	kapcsolatba, összefüggésbe, összeköttetésbe, stb.

1. táblázat. A *hoz* igéhez tartozó kifejezésjelöltek listájára néhány példa az angol igék szerint csoportosítva

3.4. A kifejezések rangsorolása

Ahogy az előző fejezetben látható volt, sok olyan kifejezés is megjelent a listán, amik nem feltétlenül alkotnak félig kompozicionális kifejezést az éppen vizsgált magyar igével (pl. *dolgokat hoz*). Ezért különböző statisztikai mérőszámok lineáris kombinációjával meghatároztunk minden frázisjelölthöz egy pontszámot. A pontszámításhoz használt mérőszámok a következők voltak:

- az angol és magyar igepár közös előfordulásainak a száma, azaz hogy hányszor voltak összekötve egymással a szómegfeleltetési modellben
- hányszor volt a vizsgált igepár úgy összekötve egymással, hogy a magyar oldalon egy főnév is az angol igéhez volt kötve
- a különböző magyar főnevek száma, amivel az angol ige megfeleltetésben állt egy adott magyar ige esetén
- az adott magyar-angol ige megfeleltetés esetén, az angol igéhez kötött magyar főnevek mindegyikénél meghatároztuk azt, hogy hányszor fordult elő az adott igével, majd ezt elosztottuk az összes főnevek számával, ami az adott igével megfeleltetésben állt (normalizált gyakoriságérték)

A vágási küszöbértéket két szempont mentén határoztuk meg a fenti értékek alapján. Először az igepárokra vonatkozó szűrést végeztünk. Ekkor minden egyes angol-magyar igepárhoz a megfeleltetett magyar főnevek közül a legnagyobb normalizált gyakoriságértékkel rendelkező főnévhez tartozó értéket megszoroztuk a második paraméterrel, azaz azzal az értékkel, hogy hányszor volt a vizsgált igepár úgy összekötve egymással, hogy a magyar oldalon egy főnév is az angol igéhez volt kötve. Ez biztosította azt, hogy a szómegfeleltetési modellben csak nagyon ritkán egymáshoz rendelt igék, illetve főnevek ne kerüljenek a listába. Ez helyettesítette a frázistáblában eredetileg szereplő, a megfeleltetés valószínűségét tükröző pontszámokat is.

Az egyes igékhez tartozó főnevek listájában is meghatároztunk egy küszöbértéket az alapján, hogy a normalizált gyakoriságértékek szerint rendezett listában hol van hirtelen nagy esés. Erre azért volt szükség, mert ezekben a listákba sokszor bekerültek olyan szavak, amik ugyan tényleg szinte mindig az adott igével

voltak összekötve, de ez nem azért volt, mert annak magyar megfelelőjével valamilyen kifejezést alkotnának, hanem csupán azért, mert a korpuszban szereplő néhány előfordulásuk mindig azzal az igével szerepelt.

A két vágás segítségével tehát eliminálni tudtuk a nem megfelelő ige-ige és a nem megfelelő főnév-ige párosításokat.

4. Eredmények

Az algoritmus kiértékeléséhez a Szeged Korpuszból és a SzegedParalell korpuszból készült félig kompozicionális igei szerkezeteket tartalmazó listát használtuk [13]. Az ebben a listában félig kompozicionális kifejezések részeként szereplő igékre futtattuk le a fenti algoritmust.

A számunkra érdekes igei kifejezések (illetve a megfelelő ige-vonzat párok) azonosításához, és egyben az algoritmus kiértékeléséhez a korábban ismertetett kérdeztesztet alkalmaztuk. Azaz azokat a kifejezéseket tekintettük helyes találatnak, ahol az adott igének az adott névszó valóban vonzata, és a névszóra vonatkozó *kit/mit/hol/hova* stb. típusú kérdés az adott igével nem lehetséges, vagy vicces hatást kelt.

Az algoritmus 309 igére adott eredményt, ezekhez összesen 6531 névszójelöltet generált. Meglepően sok új a kérdezés szempontjából számunkra érdekes idiomatikus illetve félig kompozicionális kifejezést hozott felszínre, amelyek a Szeged Korpuszból és a SzegedParalell korpuszból készült listán nem szerepeltek. Ugyanakkor az utóbbi listán szereplő kifejezések egy része a mi tesztünk szerint nem volt problematikus. A cikk beadási határidejéig a lista 1/4-ét sikerült feldolgozni. Ezen az anyagon a Szeged Korpuszból és a SzegedParalell korpuszból készült, illetve a saját algoritmusunk által generált listán szereplő összes számunkra érdekes igei kifejezést alapul véve (ebből számoltunk fedést) a szegedi lista pontossága 83,6%-osra, fedése 32,2%-osra, a sajátunk pontossága 28,6%-osra, fedése 84,2%-osra adódott. A végeredményként előállt lista elemeinek 2/3-a tehát az itt leírt eljárás eredményeként került horogra, ami nagyon jó eredmény. A viszonylag alacsony pontosság miatt ugyanakkor mindenképpen az eredmények alapos kézi átvizsgálására van szükség. A pontosságot rontja többek között az is, hogy az igéhez tartozó igekötő nem minden esetben jelenik meg a frázistábla építésekor meghatározott 7 tokenes ablakban. Ez a hiba azonban kézi javítás során általában viszonylag könnyen orvosolható. Mivel eleve csak a kérdezés szempontjából problematikus idiomatikus és félig kompozicionális ige–névszó szerkezetek azonosítását tűztük ki célul, az algoritmus nem azonosítja ezeknek a szerkezeteknek a teljes vonzatkeretét sem (a lexikálisan kötött névszói elem melletti egyéb vonzatokat). Ezeket a tematikus szerepükkel együtt kézzel adjuk hozzá az algoritmus által azonosított szerkezetek leírásához.

5. Konklúzió

Cikkünkben egy elemzett angol-magyar párhuzamos korpuszból idiomatikus és félig kompozicionális igei szerkezeteket azonosító algoritmust mutattunk be. Az

itt bemutatott kutatás része egy kérdések megfogalmazását lehetővé tevő elemző megalkotására irányuló folyamatban levő projektnek, melynek részletesebb leírását l. a szintén jelen kötetben megjelent cikkünkben [1]. Az eredményeket egy ilyen szerkezeteket tartalmazó létező erőforrással összevetve azt találtuk, hogy algoritmusunk sok új számunkra érdekes nem vagy részben kompozicionális igei szerkezetet azonosított.

Köszönetnyilvánítás

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program finanszírozásában a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással valósult meg.

Hivatkozások

1. Novák, A., Laki, L.J., Novák, B., Dömötör, A., Ligeti-Nagy, N., Kalivoda, A.: Egy magyar nyelvű kérdezőrendszer. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019), Szeged, SZTE (2019)
2. Komlósy, A.: Régensek és vonzatok. In Kiefer, F., ed.: Strukturális magyar nyelvtan 1. Akadémiai Kiadó (1992) 299–527
3. Sag, I.A., Baldwin, T., Bond, F., Copestake, A.A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '02, Berlin, Heidelberg, Springer-Verlag (2002) 1–15
4. de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., Villavicencio, A.: Alignment-based extraction of multiword expressions. Language Resources and Evaluation **44**(1-2) (2010) 59–77
5. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
6. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 173–180
7. Minnen, G., Carroll, J.A., Pearce, D.: Applied morphological processing of english. Natural Language Engineering **7**(3) (2001) 207–223
8. 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 (RANLP 2013), Hissar, Bulgaria, Incoma Ltd. Shoumen, Bulgaria (2013) 539–545
9. Novák, A.: Milyen a jó Humor? [What is good Humor like?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia [First Hungarian conference on computational linguistics], Szeged, SZTE (2003) 138–144

10. Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 1068–1073 ACL Anthology Identifier: L14-1207.
11. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of ibm model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (2013) 644–648
12. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: ACL, The Association for Computer Linguistics (2007)
13. Vincze, V.: Semi-Compositional Noun + Verb Constructions : Theoretical Questions and Computational Linguistic Analyses. PhD thesis (2012)