Az igekötők gépi annotálásának problémái

Kalivoda Ágnes

PPKE BTK Nyelvtudományi Doktori Iskola
k.agnes92@gmail.com

Kivonat: Az automatikus szófaji egyértelműsítésben nehézséget okoznak a homográf szavak, vagyis amelyek írásképükben egyeznek, de a jelentésük és gyakran a szófajuk is eltérő. Kiugróan magas a tévesztések aránya az igekötők esetében: a *meg* például kötőszóként és igekötőként is gyakori, és gyakran félreelemzett. Nemcsak más magyar szófajokkal lehetnek átfedésben az igekötők, hanem idegen nyelvi szóalakokkal (pl. *be* mint angol létige és mint magyar igekötő), valamint rövidítésekkel, mozaikszókkal is (pl. *LE* mint lóerő). Gyakori és nehezen kezelhető probléma a félregépelés és az ékezetek hiánya (pl. *fél → fel*). Az annotáció javítása fontos feladat, mivel a jobb annotáció bármilyen magyar nyelvű, korpuszt használó kutatáshoz jobb alapanyagot biztosít. A kutatásom olyan megoldást ad a homográfia problémájára, amely – a munka jelen szakaszában – a rossz annotációk több mint felét képes kiszűrni. A javítási módszer szabályalapú (reguláris kifejezéseket használ), és az igekötőként elemzett szavak kontextusára támaszkodik.

1 Bevezetés

Az automatikus szófaji egyértelműsítésben a homográf szavak jelentik az egyik legnagyobb problémát. Különösen gyakran látni hibás annotációt az igekötők esetében: például a *ki* használata kérdő vagy vonatkozó névmásként is gyakori, az *át* névutó is lehet. Jelen tanulmány egy szabályalapú javítási módszert mutat be, melynek során automatikusan, az igekötőként annotált szó kontextusa alapján dől el, hogy helyes-e a megadott szófaji címke.

Jelen tanulmány célja az, hogy választ adjon a következő kérdésekre:

(1) Hogyan lehet automatikusan javítani az igekötők gépi annotációját?

(2) Mennyire lehet hatékony a többértelműségek feloldása?

A tanulmány felépítése a következő: A 2. fejezet ismerteti a felhasznált korpuszokat és az egyes munkafázisokat. A 3. fejezet a homográfia problémakörét járja körül. Ezt követi a szabályalapú megoldás bemutatása (4. fejezet), és a javasolt megoldás kiértékelése a tesztkorpuszon a pontosság, fedés és f-mérték megállapításával (5. fejezet). Végül rövid összegzés után a további kérdések és teendők ismertetése következik.

2 Módszertan

2.1 Korpuszok

A kutatásban két korpuszt használtam fel. A 1,2 milliárd token nagyságú Pázmány-korpusz (Endrédy 2016) alapján térképeztem fel az igekötőket érintő annotációs hibákat, és állítottam össze a kontextusra illeszkedő reguláris kifejezéseket, amelyek segítségével elvégezhető a javítás. Tesztkorpuszként a Magyar nemzeti szövegtár (MNSz.) 2.0.4 verziója (Oravecz–Váradi–Sass 2014) szolgált, amely írásjelekkel együtt 1,348 milliárd tokent tartalmaz.

A két korpuszban nagyon hasonló elven működik a szófaji egyértelműsítés (az MNSz. -hez l. Oravecz–Dienes 2002, a Pázmány-korpuszhoz l. Orosz–Novák 2013), így a javító script mindkettőn alkalmazható, és közel azonos teljesítményt nyújt. Két nagyobb különbségre hívnám fel a figyelmet a szófaji címkék kapcsán:

(1) Az MNSz.-ben 79 egytagú igekötő szerepel, a Pázmány-korpuszban 76. Az *alább, benn, közé* szavak ez utóbbiban sosem jelennek meg igekötőként, ha elváltak az igétől.

(2) Az ismeretlen szavak kezelése az MNSz.-ben hatékonyabb. Itt SKIP címke járul az idegen nyelvi szavak jelentős részéhez, míg a Pázmány-korpuszban minden szó kap valamilyen magyar szófaji címkét. Például az angol *must be* 'kell hogy legyen' mind a 195 előfordulásában tárgyesetes főnév + igekötő elemzést kap a Pázmány-korpuszban, az MNSz. 750 példájában 65 alkalommal alanyesetű főnév + igekötő, 695-ször SKIP + SKIP elemzés társul hozzá.

2.2 Munkafázisok

A javítás megkezdése előtt fel kellett mérni, mely igekötők milyen más szófaji címkét kaphatnak. A Pázmány-korpuszból lekértem minden olyan mondatot, amely igekötőként annotált szavakat tartalmaz, ezután olyanokat, amelyek tartalmazzák a vizsgált szóalakokat, viszont valamilyen más szófaji címkével. Szavanként átlagosan két különféle elemzéssel találkozunk, de néhány esetben ennél jóval hosszabb a lehetséges elemzések sora (példaként ld. 1. táblázat).

Annotáció	Annotáció jelentése	Gyakoriság
NU	névutó	89 212
IK	igekötő	13 014
HA	határozószó	128
FN+POS+NOM	alanyesetű birtokos főnév (mié?)	57
FN+FAC	factivus ragos főnév (mivé?)	1

1. táblázat. A *mellé* szó lehetséges elemzései és ezek gyakorisága a Pázmány-korpuszban

A kapott eredmények alapján megállapítottam a hibatípusokat és az ezek szűrésére alkalmas, kontextusra illeszkedő szabályokat, amelyeket reguláris kifejezésekkel adtam meg. Ezeknek a megalkotása kísérletező munka eredménye volt. A vizsgált mondatokat először csoportokba rendeztem aszerint, hogy mely igekötőket (pontosabban: igekötőnek annotált szavakat) tartalmazzák. Ezeken belül további

csoportokat hoztam létre az igekötőcímkés szó maximum 5-5 szavas környezetének a szófaji címkéiből. Így könnyebben láthatóvá váltak a hibás mintázatok.

Miután a szabályok elkészültek, az MNSz.-ből kigyűjtött 5000 mondatos korpuszon teszteltem a hatékonyságukat. A tesztkorpusz összeállításának szempontjairól bővebben lesz szó az 5.1. alfejezetben.

3 A homográfia problémája

Az igekötők homográfiáját vizsgálva négy hibatípus különíthető el, ezek közül a legnagyobb gondot az elírásokból, ékezet nélkül gépelt szavakból adódó homográfia okozza. Az alábbi példában nem világos, hogy elírásról van-e szó (tehát a *meg* + *ad* igéről felesleges szóközzel) vagy a *meg* 'pedig' jelentéséről:

[1] Lemondani nem engedi mert amit *meg ad* számot azon nem lehet!!

A legtöbb (összesen 66) igekötőt érintő probléma az, hogy más szófajú magyar szóval lehet azonos az írásképe. Az alábbi példák esetén van egy preferált értelmezésünk, de automatikus módszerrel eldönthetetlen, hogy a vizsgált szónak melyik jelentése szerepel a mondatban:

[2] Ha a karbantartó sokáig vacakol, akkor az a konklúzió, hogy nem *ért hozzá*. → *Ért valamihez* vagy *hozzá + ért (valamihez)*?
[3] 4 telót rendeltem online tudom mit ír ki. → *Ki + ír* vagy *(vala)ki ír*?

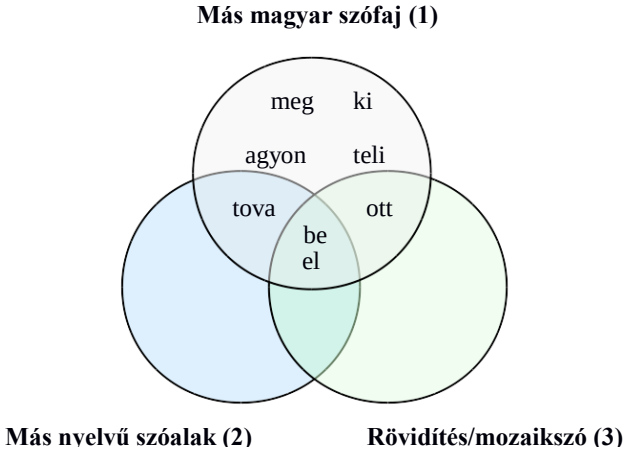
Bár magyar nyelvű korpuszokról van szó, viszonylag sok idegen nyelvű mondatot is találunk a szövegekben. Ekkor nem ritka eset, hogy például a francia *le*, a spanyol és arab *el* névelő vagy a bolgár *tova* mutató névmás igekötői címkét kap. Néhány példa, amelyben az idegen szó igekötőként elemzett:

[4] […] ugyanis a szám kerékpáros szakaszában összeütközött a kínai Ma Ming-hsziü, a dél-koreai Kim Hi Szun és a szingapúri Clara Vong *Van Ki*.
[5] […] helikopter *vitte* a Sharm *el* Sheiki kórházból a per helyszínére, ahol kerekes hordágyon tolták be a tárgyalóterembe.

Rövidítések, mozaikszók is kapnak tévesen igekötői címkét, például:

[6] *LE* (lóerő), *OTT* (oxytocin terheléses teszt)

A helyzetet tovább bonyolítja az, hogy egy igekötő akár több hibatípusban is érintett lehet, ilyenek láthatók az 1. ábra halmazainak metszetében:



1. ábra. Homográfiából adódó hibák az igekötők esetében

Homográfia szempontjából a *be* bizonyult a leginkább problémásnak, mivel megfelel neki egy rövidítés (*BE = Bowling Egyesület*), idegen nyelvű szóalak (az angol *be* 'lenni'), a magyar nyelven belül pedig kétféle határozószó (pl. *Se ki, se be.; Be rút teremtés!*) Bár elvileg több homográf párja lehet, az annotációja elég pontos (7% körül hibaaránnyal) az MNSz.-ben és a Pázmány-korpuszban is.

4 A javítás menete

A hibajavítás Python 3.4-ben írt scripttel történik. Ennek bemenete egy szövegfájl, amelyben minden sor egy mondat, a mondat szavainak a lemmatizált (szótövesített) alakjai és a morfológiai-szófaji elemzése is látható (az MNSz. esetében ez az `msd`, a Pázmány-korpuszban a `humor` mező), például:

```
tejet/tej/FN.ACC meg/meg/IK ilyeneket/ilyen/MN_NM.PL.ACC
```

A script megkeresi az `IK`, azaz igekötő címkével ellátott szót, megállapítja a lehetséges hibatípusokat (a *meg* esetében a hibatípus 1, tehát csak más magyar nyelvi szófajjal keveredik, a *be* igekötőnél 1, 2 és 3 is lehet, mert minden hibatípussal előfordulhat). Minden hibatípushoz tartozik egy függvény, a függvényen belül pedig számos további szabály (reguláris kifejezés), amelyeket egyesével ki kell próbálni az adott mondatra. Ha az egyik talált, a mondat feldolgozása véget ér.

A *meg*-hez Makrai (2007) már kidolgozott 7 szabályt, amelyeket kisebb módosításokkal felhasználtam. A többi igekötőhöz írt reguláris kifejezéssel együtt a script összesen 98 szabályt használ. A 2. táblázat a *meg*-re vonatkozó egyik szabályt mutatja be részletesen.

Pszeudokóddal	Reguláris kifejezéssel
névszó + esetrag$_i$	[^]+\/(FN|MN|SZN|MN_NM|NM_MN)[+.][^]*([^]{3})
0–3 szó, amely nem .?!	([^ .?!]+){0,3}
meg (mint igekötő)	meg\/meg\/IK
0–3 szó, amely nem .?!	([^ .?!]+){0,3}
névszó + esetrag$_i$	[^]+\/(FN|MN|SZN|MN_NM|NM_MN)[+.][^]*\2

2. táblázat. A *meg* egyik tipikus kötőszói használatára illeszkedő szabály: két azonos esetragot viselő névszó között a *meg* jellemzően kötőszó (a pszeudokódban használt indexelés jelzi az esetragok egyezését)

A script kimenete egy olyan szövegfájl, amely az egyes mondatok előtt további információkat is tartalmaz, ezek a következők:

1. hibás volt-e az eredeti annotáció (Y/N)
2. melyik címkéről mire kell váltani
   (ha nem volt hibás, akkor marad IK/IK, ha hibás, akkor IK/HA, IK/NU stb.)
3. melyik hibatípus áll fel
   (ha nincs hiba: 0, ha van, akkor az 1. ábra halmazainak megfelelő számok)
4. melyik szabály illeszkedett a mondatra (ez a megfelelő szabály számával jelölt)

Ezekre azért van szükség, hogy fejleszteni lehessen a szabályrendszert, könnyen átlátható, melyik szabály milyen gyakran érvényesül és mennyire hatékony. Három főbb szabálytípust használok:

Leggyakrabban **általános, mintaalapú szabályokra** van szükség, ilyen a 2. táblázatban bemutatott reguláris kifejezés, vagy például az *át* javítására használt szabály, miszerint: ha superessivus ragos névszót követ az *át* (pl. *ezt hallgattuk órákon át*), akkor nem igekötő. Az általános szabályok hátránya az, hogy ritkán ugyan, de jól annotált szavakat is kiszűrnek. Például az előbbi *át*-tal kapcsolatos szabályra ez a mondatrész is illeszkedik: *esett a műtéten át*, ahol valóban arról van szó, hogy *át* + *esett a műtéten*, így nem kellene javítani.

Az idegen nyelvi szóalakkal való keveredés esetén a **speciális, lexikonalapú szabályok** bizonyultak a leghatékonyabbnak. Ez lényegében az idegen nyelvi kontextus, általában tulajdonnevek listázását jelenti, amelyet egy külön fájlból beolvas a script, és kivételként kezeli őket (pl. ***El*** *Greco, Sharm* ***el*** *Sheiki, Entre* ***el*** *amor y* ***el*** *odio*). Ennek hátránya az, hogy folyamatosan kerülhetnek be új, idegen nyelvi elemek a korpuszba, így a kivételek listája nem fedheti le az összes esetet.

A harmadik szabálytípus a **lehetséges ige-igekötő kombinációkat** veszi figyelembe (27 091 kombinációt ismer, a listát l. Kalivoda 2016). Ez olyan igekötők esetén hatékony, amelyek nem produktívak, és csak néhány igéhez járulhatnak, pl. *észre, észbe, egyet.* A szabály lényege az, hogy ha az igekötőként annotált szó a tagmondat valamelyik igéjével létező kombinációt alkot, akkor marad az igekötői címkéje, ellenkező esetben új szófaji címkét kap. Például:

[7] szépség mellett *észre* is szükség *van* → *észre* + *van* nem létező kombináció

[8] ezt a hibát korábban *észre* sem *vettem* → *észre* + *vesz* létező kombináció

Ennek hátránya az, hogy ha félreelemzett az ige, akkor a script tévesen javítja a mondatot, például:

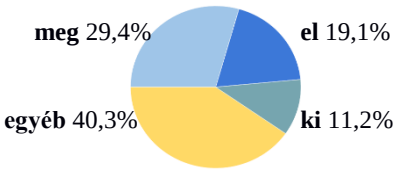
[9] *észre* is *vétette* → *vétet* helyett *vét* a lemma, *észre* + *vét* nem létezik

Mint látjuk, mindegyik szabálytípusnak megvannak a korlátai, és maga az igekötő, valamint annak hibatípusa határozza meg, mikor melyiket célszerű alkalmazni. A további technikai részletek és a forráskód elérhetők az alábbi linken: http://github.com/kagnes/prt_rules (letöltve: 2017. október 16.).

5 Kiértékelés

5.1 A tesztkorpusz

A legfontosabb szempont a tesztkorpusz összeállításánál az volt, hogy az igekötők mennyiségét tekintve az MNSz. kicsinyített mása legyen. Ehhez először megmértem, hogy a teljes MNSz. igekötő-állományának hány százalékát alkotják az egyes igekötők. Az eredményt a 2. ábra szemlélteti.



2. ábra. Az MNSz. igekötő-állománya: leggyakoribb igekötők a *meg, ki, el,* a további 76 igekötő összesen 40,3%-ot tesz ki

Az MNSz.-ből ezután lekértem véletlenszerűen 5000 mondatot, amely igekötőnek annotált szót tartalmaz – nem az ige vagy igenév közvetlen környezetében, ezzel növelve a hibás annotáció esélyét. Ügyeltem arra, hogy a teljes MNSz.-ben megfigyelt arányokat megőrizzem (így pl. a mondatok 29,4%-a, azaz 1470 mondat tartalmazta a *meg* igekötőt). A kapott adatokat átnéztem, és az 5000-ből 1033, tehát a tesztkorpusz egyötöde bizonyult hibásnak. Mivel a hibás mondatok esetén azt is meg kellett határozni, hogy mi lenne a helyes annotáció, a feladat 8-10 órát vett igénybe. A hibák arányát a 3. táblázat mutatja be részletesen.

Igekötő	Darab	Helyes	Hibás	Hibásak %-a
meg	1470	960	510	34,7
el	956	901	55	5,8
ki	563	463	100	17,8
be	368	341	27	7,3
fel	355	332	23	6,5
vissza	123	96	27	21,9
át	118	21	97	82,2
hozzá	77	37	40	51,9
rá	70	22	48	68,6

3. táblázat. Néhány gyakori igekötő eloszlása a tesztkorpuszban (az igekötő mellett látható a tartalmazó mondatok mennyisége, a helyes és hibás annotáció mennyisége, végül a hibás annotáció aránya a helyeshez képest)

Három olyan igekötő van, amelyik viszonylag gyakori, és az esetek több mint felében hibás szófaji címkét kap, ezek az *át, rá* és *hozzá.* Az *át* 82,2%-ban igekötőként annotált akkor is, ha valójában névutó, így kimagasló hibaaránnyal bír. A *vissza* esete azért problémás, mert egyik korpuszban sincs soha határozószóként annotálva, pedig 21,9%-ban ez volna a megfelelő címke.

5.2 Eredmények

A szabályok hatékonyságát a számítógépes nyelvészetben hagyományosnak számító módszerrel, a pontosság, fedés és f-mérték segítségével értékeltem ki. A pontosság (1) azt fejezi ki, hogy a script által hibásnak jelölt annotációk közül mennyi valóban hibás. A fedésre (2) kapott százalékos érték azt jelzi, hogy az összes hibás annotáció közül hányat talált meg a script. Az f-mérték (3) a pontosság és fedés harmonikus közepe.

(1) pontosság = (valóban hibásak × 100) / hibásnak jelöltek

(2) fedés = (megtalált hibásak × 100) / összes hibás

(3) f-mérték = 2 × ((pontosság × fedés) / (pontosság + fedés))

A 4. táblázat tartalmazza a script eredményeit az 5000 mondatos tesztkorpuszon.

Pontosság	Fedés	F-mérték
88,2 %	57,5 %	69,6 %

4. táblázat. A javító script teljesítménye a tesztkorpuszon

A pontossága viszonylag jó, 88,2%-ban igaz, hogy hibás annotációt ismert fel, nem egy eredetileg jót rontott el. A fedése viszont gyenge, az összes hibás szófaji címkének valamivel több mint a felét sikerült felismernie. Ugyanakkor fontos, hogy a felismerésen túl 98%-ban helyesen is javította az annotációt (a tökéletlenség néhány olyan esetnél mutatkozik meg, ahol rosszul választ a névutó és a határozószó címke között).

A fedés két okból alacsony. Egyrészt, ha maga a kontextus is hibásan elemzett (pl. az *ír* ige melléknévi annotációt kap), akkor a megfelelő szabály nem illeszkedik rá. Másrészt a fel nem ismert mondatok átnézése során kiderült, hogy rengeteg köztük az elírás (pl. *el* helyett *le*), hibás szóközt tartalmazó, vagy egyszerűen ékezetek nélküli mondat (pl. *fél → fel, még → meg*). A script az elírásokat jelenleg semmilyen formában nem kezeli.

6 Összegzés, további kérdések

A kutatás eredménye egy olyan script, amellyel az igekötők annotációját utólag lehet javítani. A kutatási kérdésekre adott válaszaim a következők: (1) A kontextus mintázatai alapján hatékonyan javítható az igekötők annotációja, de a hibásan elemzett kontextus és elírások felismerése, javítása – egyelőre – nem megoldott. Tökéletes eredményt nem is várhatunk. (2) A hibás annotációt az esetek több mint

felében sikerült egyértelműen kiszűrni. Legnehezebbnek azok a szerkezetek bizonyultak – az elírásokat leszámítva –, ahol a kérdéses szó (főként a *meg* és *ki*) olyan pozícióban áll, ahol igekötő is várható, és a tagmondat igéjével létező kombinációt alkothat.

További teendők a fedés arányának javítása, az összetett igekötők (pl. *bele-bele*, *ki-be*) annotációjának ellenőrzése, valamint a javítási irány megfordítása: olyan szabályokat is létre kell hozni, amelyek nem igekötőként annotált szavakról döntik el, hogy igekötők-e.

Jelen kutatásban olyan mondatokkal foglalkoztam, ahol az igekötőnek elemzett szó közelében ige is szerepelt. Viszont az ige jelenléte nem kötelező egy elliptikus szerkezetben, az eldöntendő kérdésre adott rövid válaszban. A problémát az alábbi két példa szemlélteti, az elsőben a *meg* igekötő, a másodikban diskurzuspartikula (jobb híján KOT, azaz kötőszó címkét lehetne hozzárendelni a korpuszokban):

[10] Megírtad a cikket? Én *meg*.
[11] Te megint ülsz le játszani… Én *meg*?

Végül nehézséget jelent az is, hogy az igekötő-állomány az MNSz.-ben és a Pázmány-korpuszban sem konzisztens. Például igekötőként elemzett a *zokon*, de a *cserben* inessivus ragos főnévként szerepel, pedig a mondatbeli viselkedése és a produktivitás szempontjából ez a két szó nem különbözik egymástól. Sok esetben nem dönthető el, hogy az adott szó határozószó vagy igekötő-e, pl. *oda, haza*; illetve névmás vagy igekötő, pl. *rá*. Mivel ezeknek nyelvészeti szempontból sem egyértelmű a besorolása, érdemes lehet egy vagylagos szófaji címke használata, az *oda* esetén például [IK|HA]. Ez könnyen kivitelezhető módosítás – és az MNSz.-ben már most is találunk hasonlót.

Források

MNSz. = *Magyar Nemzeti Szövegtár* v2.0.4. Elérhető: http://corpus.nytud.hu/mnsz. Letöltve: 2017. szeptember 26.
Pázmány-korpusz. Nyilvánosan nem elérhető.

Irodalom

Endrédy, I. 2016. *Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz.* PhD-értekezés. Elérhető: https://itk.ppke.hu/uploads/articles/163/file/EI_PhD_Disszert%C3%A1ci%C3%B3.pdf. Letöltve: 2017. szeptember 26.

Kalivoda, Á. 2016. *A magyar igei komplexumok vizsgálata.* Mesterszakos szakdolgozat. Elérhető: https://github.com/kagnes/hungarian_verbal_complex. Letöltve: 2017. szeptember 26.

Makrai, M. 2007. *Többértelműségek magyar mondatok számítógépes elemzésében – a „meg" szó szófajának vizsgálata gyakoriságokkal.* Témalabor-dolgozat. Elérhető: http://hlt.bme.hu/media/pdf/makrai07_temalabor_meg.pdf. Letöltve: 2017. szeptember 26.

Oravecz, Cs., Dienes, P. 2002. *Efficient Stochastic Part-of-Speech Tagging for Hungarian.* In: Calzolari, N., Choukri K., Maegaard B., Mariani J., Municio A. M. Tapias, D. Zampolli, A. (szerk.) *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002).* Párizs: ELRA. 710–717. Elérhető: https://www.researchgate.net/profile/Csaba_Oravecz/publication/2481141_Efficient_Stocha

stic_Part-of-Speech_Tagging_for_Hungarian/links/0912f5139e302a44c0000000/Efficient-Stochastic-Part-of-Speech-Tagging-for-Hungarian.pdf. Letöltve: 2017. szeptember 26.

Oravecz, Cs., Váradi T., Sass B. 2014. The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. (szerk.) *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014).* Elérhető: http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf. Letöltve: 2017. szeptember 26.

Orosz, Gy., Novák, A. 2013. *PurePos 2.0: a hybrid tool for morphological disambiguation.* In: Angelova, G., Bontcheva K., Mitkov R. (szerk.) *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013).* Sumen: INCOMA Ltd. 539–545. Elérhető: http://aclweb.org/anthology/R13-1071. Letöltve: 2017. szeptember 26.