

Angol és magyar nyelvű kérdések a számítógépes nyelvészetben

Vincze Veronika^{1,2}

¹Szegedi Tudományegyetem, Informatikai Tanszékcsoport

²MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat A cikkben korpuszalapú vizsgálatok segítségével bemutatjuk a magyar és angol nyelvű kérdések sajátosságait, különös figyelmet fordítva a közösségi médiában előforduló kérdésekre. Emellett a kérdések számítógépes nyelvészeti hasznosíthatóságára is rámutatunk, egyrészt többszavas kifejezések azonosításában, másrészt eldöntendő kérdésekre felajánlott automatikus válaszlehetőségek továbbfejlesztésében.

Kulcsszavak: kérdések, szemantika, pragmatika, többszavas kifejezések, alkalmazás

1. Bevezetés

A kérdések szerepe a természetes nyelvekben főként az, hogy információt kérjünk másoktól, így központi fontossággal bírnak az emberi kommunikációban. A kérdések a legtöbb nyelvben eltérően viselkednek az állításoktól, és sajátos nyelvi (logikai, interperszonális és szintaktikai-szemantikai) jellemzőkkel bírnak, amire több nyelvészeti tanulmány is rámutatott (pl. [1,2,3]), a számítógépes nyelvészetben viszont – a kérdések megválaszolása (question answering) témakört leszámítva – kevésbé vizsgált témának számít.

Jelen cikk célja, hogy korpuszalapú vizsgálatok segítségével megvizsgáljuk a magyar és angol nyelvű kérdések sajátosságait, különös figyelmet fordítva a közösségi médiában előforduló kérdésekre. Emellett a kérdések számítógépes nyelvészeti hasznosíthatóságára is rámutatunk, egyrészt angol nyelvű többszavas kifejezések azonosításában, másrészt magyar nyelvű eldöntendő kérdésekre felajánlott automatikus válaszlehetőségek továbbfejlesztésében.

2. Angol nyelvű adatok

Angol nyelvű adataink két forrásból származnak: az angol Univerzális Dependencia Treebank (UD) [4] és a QuestionBank (QB) [5]. Utóbbi 4000 kérdést tartalmaz, az UD treebankben pedig a mondatok 6,76%-a (1124 darab) kérdés. A kérdőszavak részletes elemzéséből (l. 1. táblázat) kiderül többek között, hogy például a *how* és *why* szavak használata sokkal gyakoribb az UD treebankben,

amelyben szerepelnek közösségi médiából származó szövegek is: a felhasználók gyakran kérnek egymástól segítséget a weben különféle ügyekkel kapcsolatban, ugyanakkor ritkábban tesznek fel személyekre vagy helyekre, időpontokra irányuló tényszerű kérdéseket, ellentétben a QB kérdéseivel.

1. táblázat. Az angol kérdőszavak eloszlása.

| | UD | % | QB | % | Összesen | % |
|------------|-----|--------|------|--------|----------|--------|
| what | 173 | 43,14 | 2333 | 60,63 | 2506 | 58,98 |
| who | 21 | 5,24 | 445 | 11,56 | 466 | 10,97 |
| how | 72 | 17,96 | 233 | 6,06 | 305 | 7,18 |
| where | 25 | 6,23 | 243 | 6,31 | 268 | 6,31 |
| when | 5 | 1,25 | 199 | 5,17 | 204 | 4,80 |
| how many | 6 | 1,50 | 173 | 4,50 | 179 | 4,21 |
| which | 15 | 3,74 | 87 | 2,26 | 102 | 2,40 |
| why | 48 | 11,97 | 49 | 1,27 | 97 | 2,28 |
| how much | 12 | 2,99 | 49 | 1,27 | 61 | 1,44 |
| how long | 5 | 1,25 | 26 | 0,68 | 31 | 0,73 |
| how about | 11 | 2,74 | 0 | 0,00 | 11 | 0,26 |
| what about | 8 | 2,00 | 0 | 0,00 | 8 | 0,19 |
| whose | 0 | 0,00 | 8 | 0,21 | 8 | 0,19 |
| whom | 0 | 0,00 | 3 | 0,08 | 3 | 0,07 |
| Összesen | 401 | 100,00 | 3848 | 100,00 | 4249 | 100,00 |

3. Kérdések a nem sztenderd szövegekben

A webes nyelvhasználat, különös tekintettel a felhasználók által írt szövegekre, számos olyan sajátossággal rendelkezik, melyek eltérnek a sztenderd nyelvtől [6,7,8]. Ide sorolhatjuk többek között a központozás hiányát vagy következetlen használatát, a rövidítések gyakori használatát, az újonnan alkotott szavakat, a gyakori elírásokat és emotikonokat. Bizonyos esetekben pedig a mondatok szintaktikai szerkezete is eltérhet a sztenderdtől. Például az angolban gyakran elmaradnak az alanyi funkciót betöltő első személyű névmások:

Can't believe you left last night.

A mondatok sokszor ellipszist tartalmaznak, és pusztán egy frázisból állnak:

Very professional.

Reasonable rate.

Az elliptikus kérdések számos társalgási funkciót tölthetnek be, céljuk lehet a visszakérdezés vagy a tényleges kérdésfeltevés. Utóbbi esetben egyes, a sztenderd nyelvben kötelező mondatrészek hiányozhatnak¹.

Az alábbiakban közlünk néhány példát a korpuszból, kiegészítve a teljes, sztenderd nyelvnek megfelelő kérdésváltozattal:

Any feedback from Rick Buy? vs. **Is there** any feedback from Rick Buy?
 Sushi tonight? vs. **How about eating** sushi tonight?
 Any help? vs. **Could you please give me** any help?
 Weather in december in Tremblant? vs. **What is the weather like** in December in Tremblant?
 Paris or England while studying aboard? vs. **Should I choose** Paris or England while studying abroad?
 Dwarf Hamster Making Too Much Noise On Wheel at Night? vs. **What should I do if my dwarf hamster is** making too much noise on **its** wheel at night?

A fentihez hasonló kérdések automatikus feldolgozását több tényező is nehezíti. Egyrészt a kontextus ismerete nélkül nem könnyű azonosítani a kérdések szándékolt jelentését, másrészt ehhez gyakran világtudás is szükséges. Gyakran kérésként értelmezendők (*Any help?*), máskor ajánlatként vagy javaslatként (*Sushi tonight?*), érdeklődésként (*Any feedback from Rick Buy?* és *Weather in december in Tremblant?*) vagy pusztán problémaleírásként és közvetett segítségkérésként (*Dwarf Hamster Making Too Much Noise On Wheel at Night?*). A hasonló példák részletes elemzése igen kívánatos lenne a közösségi média szövegeinek automatikus feldolgozása céljából, jelen cikkben részletesebben azonban a nem elliptikus kérdésekkel foglalkozunk.

4. Magyar nyelvű adatok

A magyar adatok a Szeged Dependencia Treebankból [10] származnak, kiegészítve 2000 mondatnyi, webről származó szöveggel [11]. A szövegek összesen 5668 kérdést tartalmaznak. Arányaiban a legtöbb kérdés a webes szövegekben szerepel, hiszen a szövegek egy része a gyakorikerdesek.hu weboldalról származik, ahol felhasználók válaszolnak egymás kérdéseire. Az adatokból az is kiderül, hogy a legtöbb eldöntendő kérdés a webes és az irodalmi szövegekben fordul elő, ahol főleg interperszonális funkcióval bírnak: a hallgató egyetértésének vagy beleegyezésének kiváltása. Ezzel ellentétben a jogi szövegeket, illetve gazdasági híreket tartalmazó alkorpuszokban alig-alig találhatunk kérdéseket, természetesen ez az adott szövegtípusok leíró jellegéből fakad.

A kérdőszavak eloszlása doménenként változó, azonban leginkább a *mi*, *milyen* és *miért* kérdőszavak fordulnak elő az adatbázisban. Számítástechnikai szövegekben gyakori még a *hogyan* használata, ami azzal magyarázható, hogy a

¹ Hasonló jelenséget figyelhetünk meg az angol médiában használt címekben, melyek szintén sajátos szintaktikai szerkezeteket tartalmaznak [9].

2. táblázat. Statisztikai adatok a magyar korpuszban található kérdésekről.

| Alkorpusz | mondat | kérdés | % | kérdőszó | k.szó/kérdés | eldönt. | eldönt./kérdés |
|------------------|--------|--------|-------|----------|--------------|---------|----------------|
| iskolás | 24720 | 1435 | 5,81 | 1352 | 94,22 | 83 | 5,78 |
| irodalom | 18558 | 2408 | 12,98 | 1660 | 68,94 | 748 | 31,06 |
| számítástechnika | 9627 | 532 | 5,53 | 498 | 93,61 | 34 | 6,39 |
| újtság | 10210 | 689 | 6,75 | 601 | 87,23 | 88 | 12,77 |
| rövidhír | 9574 | 71 | 0,74 | 70 | 98,59 | 1 | 1,41 |
| jog | 9278 | 243 | 2,62 | 243 | 100,00 | 0 | 0,00 |
| web | 1935 | 290 | 14,99 | 217 | 74,83 | 73 | 25,17 |
| Összesen | 83902 | 5668 | 6,76 | 4641 | 81,88 | 1027 | 18,12 |

3. táblázat. A leggyakoribb magyar kérdőszavak eloszlása.

| kérdőszó | iskolás | irodalom | sz.tech. | jog | újtság | hír | web | Összesen |
|----------|-----------|-----------|-----------|----------|-----------|----------|----------|------------|
| mi | 397 28,75 | 605 36,01 | 109 19,96 | 22 12,43 | 155 26,32 | 8 13,33 | 77 33,77 | 1373 29,46 |
| milyen | 263 19,04 | 141 8,39 | 109 19,96 | 65 36,72 | 70 11,88 | 13 21,67 | 12 5,26 | 673 14,44 |
| miért | 167 12,09 | 211 12,56 | 33 6,04 | 4 2,26 | 48 8,15 | 2 3,33 | 35 15,35 | 500 10,73 |
| ki | 105 7,60 | 105 6,25 | 32 5,86 | 19 10,73 | 60 10,19 | 6 10,00 | 25 10,96 | 352 7,55 |
| hogyan | 54 3,91 | 72 4,29 | 62 11,36 | 8 4,52 | 47 7,98 | 1 1,67 | 5 2,19 | 249 5,34 |
| mennyi | 60 4,34 | 67 3,99 | 39 7,14 | 15 8,47 | 38 6,45 | 5 8,33 | 9 3,95 | 233 5,00 |
| hogy | 72 5,21 | 114 6,79 | 2 0,37 | 0 0,00 | 6 1,02 | 0 0,00 | 13 5,70 | 207 4,44 |
| hol | 53 3,84 | 77 4,58 | 10 1,83 | 2 1,13 | 25 4,24 | 0 0,00 | 16 7,02 | 183 3,93 |
| mikor | 47 3,40 | 35 2,08 | 11 2,01 | 9 5,08 | 26 4,41 | 6 10,00 | 6 2,63 | 140 3,00 |
| melyik | 29 2,10 | 23 1,37 | 35 6,41 | 13 7,34 | 18 3,06 | 5 8,33 | 5 2,19 | 128 2,75 |

számítástechnikai kézikönyv gyakran tartalmaz technikai jellegű leírásokat arra nézve, hogy mit hogyan kell beállítani a számítógépen. Érdekesképpen megfigyelhetjük azt is, hogy informális nyelvhasználatban (webes, iskolás és irodalmi szövegek) jóval gyakrabban fordul elő a *hogy* kérdőszó, mint a hosszabb, szten-derd változata (*hogyan*).

A magyar webes szövegekben is találunk arra példákat, hogy a hallgató szá-ndéka nem információkérésre irányul, hanem például javasol vagy kér valamit (*Nem lehetne háromnegyed 9kor?*), esetleg az aktuális témával kapcsolatos ellen-szenvét fejezi ki (*Szalagavatón táncolni én?*). E típusú kérdések természetesen pragmatikai jelentéstartalommal bírnak, így az elvárt válasz sem mindig egysze-rűsíthető le az igen/nem válaszok egyikére. A későbbiekben szeretnénk e kér-déstípusokat részletes vizsgálatnak is alávetni mind elméleti, mind számítógépes nyelvészeti szempontból.

5. Angol prepozíciós igék azonosítása

A többszavas kifejezések több tokenből állnak össze, melyekre jellemző, hogy a teljes egység jelentése (részben) különbözik az egységek saját jelentésétől [12]. A többszavas kifejezések közé tartoznak az úgynevezett angol prepozíciós igék (VPC-k), melyek egy ige és egy (vagy több) prepozíció kombinációjából állnak (*set up* vagy *come in*). A VPC-k felszíni szintaktikai szerkezete gyakran hasonlít más kompozicionális szintaktikai frázisokra: például a *to set up the rules* és a *to run up the road* kifejezések első látásra hasonló felépítésűnek tűnnek, azonban míg az elsőben VPC-t találunk, addig a második példa pusztán egy ige + prepozíciós frázis kombinációja.

A fentiek alapján tehát a felszíni szintaktikai szerkezet nem feltétlenül bír megkülönböztető erővel a VPC-k automatikus azonosításában. Azonban még létezik néhány olyan szintaktikai és szemantikai teszt, melyek segíthetnek abban, hogy a VPC-eket elkülönítsük más hasonló felépítésű, ám kompozicionális egységektől, melyek egyike az aktuális mondat kérdéssé alakítása. A VPC-t tartalmazó mondatokból képzett kérdések gyakran tartalmazzák a *who* és *what* kérdőszavakat, míg a prepozíció a mondat végén helyezkedik el (*What did you set up?*). Ezzel szemben a kompozicionális egységekből létrehozott kérdésekben többnyire a *where* és *when* kérdőszavak fordulnak elő (*Where did you run?*). Mindemellett fontos hangsúlyozni, hogy a **Where did you set?* és a **What did you run up?* kérdések nyelvtanilag nem elfogadhatók.

A következőkben bemutatjuk, hogy az angol prepozíciós igék (VPC-k) automatikus azonosítására pozitív hatással bírnak a kérdésekre irányuló nyelvi jellemzők.

5.1. Gépi tanulási módszerek

Gépi tanulási kísérleteinkben nagyrészt [13] eredményeire támaszkodtunk. Rendszerünk kiértékeléséhez a Tu & Roth korpuszt [14] választottuk. Hogy eredményeink teljes egészében összevethetők legyenek a korpuszon elért korábbi eredményekkel [14,13], egy SVM modellt tanítottunk [15], ötszörös keresztvalidációval, a Weka csomag [16] alapbeállításait használva.

A kiértékelésben a pontosság (accuracy) metrikát használtuk.

5.2. Felhasznált jellemzők

Méréseink során a következő egyszerű jellemzőket használtuk fel:

1. Kérdésekre vonatkozó jellemzők:

- (a) kérdőszó;
- (b) a kérdőszó szófaji elemzése;
- (c) a kérdőszó mondatbeli helye (a mondat élén áll-e vagy sem);
- (d) a kérdőszó távolsága a megelőző igétől;
- (e) a kérdőszó távolsága a megelőző főnévtől;
- (f) a kérdőszó szintaktikai szerepe.

2. Igei jellemzők:

- (a) megvizsgáltuk, hogy az ige lemmája megegyezik-e a leggyakoribb angol igék egyikével, mivel általában a leggyakoribb igék szerepelnek VPC-kben;
- (b) megvizsgáltuk, hogy az ige mozgást fejez-e ki, mivel a VPC-kben sokszor mozgást jelentő igét találhatunk (pl. *come, go*).

3. A prepozícióra vonatkozó jellemzők:

- (a) megvizsgáltuk, hogy a prepozíció egyike-e az angol nyelv leggyakoribb prepozícióinak;
- (b) megvizsgáltuk, hogy a prepozíció irányt jelöl-e;
- (c) megvizsgáltuk, hogy a prepozíció a-val kezdődik-e, mivel etimológiailag az *a* prefixum mozgást jelöl (pl. *across*);
- (d) a prepozíció mondatbeli helye;
- (e) a prepozíció nyelvtani szerepe;
- (f) megvizsgáltuk, hogy a prepozíciónak van-e gyermek csomópontja a függőségi fában, és amennyiben volt, felvettük annak a nyelvtani szerepét is.

4. Mondatszintű jellemzők:

- (a) a mondat hossza;
- (b) külön jellemzőként jelöltük, ha az ige és a prepozíció egyaránt mozgást és irányt jelölt, mivel ezen kombinációk gyakran kompozicionális jelentésűek (pl. *go out*);
- (c) van-e tárgya az igének;
- (d) van-e névmási tárgya az igének;
- (e) van-e névmási alanya az igének.

Megjegyezzük, hogy a kérdésekre vonatkozó jellemzők és a prepozícióra vonatkozó jellemzők közül az utolsó három új, azaz tudomásunk szerint a VPC-k azonosítására vonatkozó hatásukat ezidáig még nem vizsgálták.

5.3. Eredmények

Az eredmények a 4. táblázatban láthatók.

4. táblázat. Gépi tanulási eredmények a Tu& Roth korpuszon.

| SVM | SVM kérdések nélkül | Tu & Roth | Nagy T. & Vincze |
|-------|---------------------|-----------|------------------|
| 80,05 | 77,46% | 78,6 | 81,92 |

Tu és Roth eredeti cikkükben [14] 78,6%-os pontosságot értek el, Nagy T. és Vincze [13] pedig 81,92%-ot ugyanezen a korpuszon mérve. A jelen cikkben közölt eredmények meghaladják [14] eredményeit, azonban [13] eredményeinél valamivel

alacsonyabbak, itt elsődlegesen azonban a kérdésekre épülő jellemzők hozzáadott értékére voltunk kíváncsiak.

Az új jellemzők bevezetése hozzájárult a rendszer jó teljesítményéhez. Modellünket újratanítottuk pusztán a szakirodalomban már korábban is használt jellemzőket hasznosítva, azaz az általunk bevezetett új jellemzőket mellőztük. Így 77,46%-os pontosságot, azaz 3,81 százalékponttal alacsonyabb teljesítményt értünk el. E kísérletünk is igazolja a kérdésekre épülő jellemzők hozzáadott értékét egy számítógépes nyelvészeti feladatban.

6. Automatikus válaszadás kérdésekre

A Yako alkalmazás fő célja, hogy egységes felületet biztosítson a telefonra beérkező üzeneteknek, legyen azok formája SMS, e-mail vagy pedig Messenger-üzenet [17]. Az érkező üzenetekre a felhasználó természetesen választ is írhat, szintén egy egységes felület segítségével. A felhasználó kényelmét biztosítandó az alkalmazás az üzenetben érkezett kérdésekre automatikus válaszlehetőségeket ajánl fel, melyek közül a felhasználó egy mozdulattal kiválaszthatja a szándékolt választ. E funkció eredetileg egyszerű eldöntendő kérdésekre működött, melyekre igennel vagy nemmel lehet válaszolni, továbbá olyan kérdésekre, ahol két lehetőség közül lehet választani.

E cikkben célunk, hogy a Yako által kezelhető kérdések körét bővítsük, azokhoz megfelelő nyelvi reprezentációt nyújtva. Ehhez megvizsgáltuk a magyar korpuszban előforduló eldöntendő kérdéseket, és a leggyakoribb szintaktikai és morfológiai mintázatokra építve felállítottunk néhány újabb lehetséges sémát az eldöntendő kérdésekre. E sémákat az 5. táblázatban foglaltuk össze, ahol a morfológiai információkat MSD-kódok formájában jelenítjük meg.

5. táblázat. Eldöntendő kérdések sémái és lehetséges válaszok.

| Séma | Válaszjavaslat | Példa |
|--|---|---|
| Rp + Va* + Vmn | igen/Rp + nem | El akarsz jönni? Igen/El. Nem. |
| Rm + Vm* | de + nem | Nem jössz velünk? De. Nem. |
| Rm + Va* + Vmn | de + nem | Nem akarsz eljönni? De. Nem. |
| ugye | igen + nem | Ugye eljössz az MSZNY-re? Igen. Nem. |
| vajon | igen + nem | Vajon eljön a karácsony? Igen. Nem. |
| -e | igen + nem | Eljön-e az MSZNY-re? Igen. Nem. |
| , nem? (mondat végén) | de + nem | Eljössz az MSZNY-re, nem? De. Nem. |
| Vm* (mondat elején) | igen + nem | Eljössz az MSZNY-re? Igen. Nem. |
| N* ₁ vagy N* ₂ | N* ₁ + N* ₂ | Sört vagy bort? Sört. Bort. |
| N* ₁ , N* ₂ vagy N* ₃ | N* ₁ + N* ₂ + N* ₃ | Sört, bort vagy kólát? Sört. Bort. Kólát. |

Reményeink szerint ezek az újonnan felállított sémák hozzájárulnak a Yako alkalmazás további tökéletesítéséhez.

7. Összegzés

Ebben a cikkben angol és magyar nyelvű kérdések számítógépes nyelvészeti elemzése és felhasználása felé tettük meg az első lépéseket. Először korpuszalapú vizsgálatok segítségével feltérképeztük az egyes kérdéstípusok és kérdőszavak gyakoriságát, majd a gyakorlati hasznosítás felől közelítettük meg a kérdéskört. Gépi tanulási kísérletekkel kimutattuk, hogy az angol prepozíciós igék automatikus azonosítására pozitív hatással bírnak a kérdésekre irányuló nyelvi jellemzők: az alaprendszerhez képest 3,8 százalékpontos javulást értünk el pontosság terén. Emellett a Yako alkalmazás egyik funkciójának – az automatikus válaszadási lehetőségeknek – továbbfejlesztésére tettünk javaslatot, a korpuszban található eldöntendő kérdések nyelvi jellemzőinek feltérképezése segítségével, ezzel is igazolva a kérdések szerepének fontosságát számítógépes nyelvészeti alkalmazásokban.

A jövőben célunk, hogy a közösségi médiában előforduló kérdések interperszonális és pragmatikai funkcióit részletesebben is feltérképezzük, illetve azok számítógépes nyelvészeti hasznosíthatóságát is megvizsgáljuk. Mindemellett szeretnénk a kérdések nyelvészeti jellemzőit más számítógépes nyelvészeti alkalmazásokba is beépíteni és azok hasznosságát megvizsgálni.

Köszönetnyilvánítás

A jelen kutatás a PARSEME COST Action (IC1207) projekt keretében az Európai Unió támogatásával valósult meg.

Hivatkozások

1. Groenendijk, J., Stokhof, M.: Questions. In van Benthem, J., ter Meulen, A., eds.: *Handbook of Logic and Language*, Amsterdam/Cambridge, MA, Elsevier/MIT Press (1997) 1055–124
2. Graesser, A.C., Person, N.K., Huber, J.D.: Mechanisms that generate questions. In Lauer, T.E., Peacock, E., Graesser, A.C., eds.: *Questions and information systems*, Hillsdale, NJ, Lawrence Erlbaum Associates (1992) 167–187
3. Dikken, M.d.: On the morphosyntax of wh-movement. In Boeckx, C., Grohmann, K., eds.: *Multiple wh-fronting*, Amsterdam, John Benjamins (2003) 77–98
4. Nivre, J., Bosco, C., Choi, J., de Marneffe, M.C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., Zeman, D.: *Universal dependencies 1.0* (2015)
5. Judge, J., Cahill, A., Genabith, J.V.: Questionbank: Creating a corpus of parse-annotated questions. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*. (2006) 497–504

6. Seddah, D., Sagot, B., Candito, M., Moulleron, V., Combet, V.: The French Social Media Bank: a treebank of noisy user generated content. In: Proceedings of COLING 2012, Mumbai, India, The COLING 2012 Organizing Committee (2012) 2441–2458
7. Mott, J., Bies, A., Laury, J., Warner, C.: Bracketing Webtext: An Addendum to Penn Treebank II Guidelines. Linguistic Data Consortium (2012)
8. Bies, A., Mott, J., Warner, C., Kulick, S.: English Web Treebank. Technical report, Linguistic Data Consortium, Philadelphia (2012) LDC2012T13.
9. Bell, A.: The language of the News Media. Blackwell, Oxford (1991)
10. Vincze, V., Szauder, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
11. Vincze, V., Varga, V., Papp, P.A., Simkó, K.I., Zsibrita, J., Farkas, R.: Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, Szegedi Tudományegyetem (2015) 122–132
12. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002, Mexico City, Mexico (2002) 1–15
13. Nagy T., I., Vincze, V.: VPCTagger: Detecting Verb-Particle Constructions With Syntax-Based Methods. In: Proceedings of the 10th Workshop on Multiword Expressions (MWE), Gothenburg, Sweden, Association for Computational Linguistics (2014) 17–25
14. Tu, Y., Roth, D.: Sorting out the Most Confusing English Phrasal Verbs. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. SemEval '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 65–69
15. Cortes, C., Vapnik, V.: Support-vector networks. Volume 20. Kluwer Academic Publishers (1995)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
17. Farkas, R., Kojedzinszky, T., Zsibrita, J., Wieszner, V.: Yako: egy intelligens üzenetváltó alkalmazás nyelvtchnológiai kihívásai. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2015) 323–325