A Magyar WordNet felhasználhatósága lexikális jelentés-egyértelműsítésben*

Kuti Judit[1], Darja Fišer[2]

[1]MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
1068, Bp., Benczúr u. 33.
kutij@nytud.hu
[2]University of Ljubljana, Faculty of Arts, Department of Translation Studies
Aškerčeva 2, SI - 1000 Ljubljana
darja.fiser@guest.arnes.si

Kivonat: Tanulmányunkban bemutatunk egy vizsgálatot, amellyel a magyar WordNet használhatóságát teszteltük a gépi fordítás során alkalmazható lexikális jelentés-egyértelműsítésben. A vizsgálat során lefordítottuk angolra egy magyar szöveg tartalmas szavait a MetaMorpho gépi fordítórendszerrel, valamint a Magyar WordNettel való jelentés-egyértelműsítésen keresztül megfeleltettük ezeket a Princeton WordNet synseteinek. Egy angol nyelvű referenciafordításhoz képest automatikusan értékeltük ki a kapott fordításokat. A MetaMorpho gépi fordítórendszer magyar–angol nyelvpárra jelenlegi állapotában jobb fordításokat ad, mint a WordNet által javasolt lexikális fordítások; teljesítményét tehát a jelen vizsgálat alapján úgy tűnik, a HuWN nem javítaná.

1 Bevezető

A jelentés-egyértelműsítés mint a nyelvtechnológia egyik központi feladata számos alkalmazásban kap fontos szerepet: a legfontosabbak ezek közül a gépi fordítás, az információkivonatolás, illetve az információkinyerés. A feladat, komplexitásából adódóan, egyelőre nem tekinthető megoldottnak – sem magyarra, sem más nyelvekre (l. [1]). Általánosan a jelentés-egyértelműsítés folyamatát két alapvető lépésre bontjuk: (i) valamilyen jelentéstár kiválasztása, illetve létrehozása, valamint (ii) a jelentéstárban szereplő jelentések hozzárendelése a kívánt szóalakokhoz valamilyen algoritmus segítségével. Nemzetközi szinten az egyik legelterjedtebb jelentéstár a WordNet (PWN – l. [5]).[1] A WordNet mint általános adatbázistípus az angol nyelvű Princeton

* Jelen tanulmány a "A Magyar és Szlovén WordNet összehasonlító kiértékelése gépi fordításban" c. projekt keretén belül született, melyet a Magyar-Szlovén Kormányközi TéT együttműködés támogat 2009-2010-ben.

[1] A legkülönbözőbb jelentés-egyértelműsítő részfeladatokhoz (célszó jelentés-egyértelműsítése, automatikus kulcsszókinyerés, szemantikai szerepek címkézése stb.) nagy százalékban ennek az adatbázisnak különböző verzióit használják mint jelentéstárat. A Senseval versenyeken használt jelentés-egyértelműsített korpuszok több mint fele valamilyen WordNet-típussal lett annotálva.

WordNetre épülő lexikális hálót takar, amelynek alapegysége a fogalom / jelentés (szakszóval *synset*), nem pedig a tradicionális szótárak alapegysége, a szó / lexéma. A wordnetek egy adott nyelv lexikalizálódott jelentéseit az egymáshoz való jelentéstani viszonyaik által alkotott hálóban helyezik el, a viszonyokat a háló éleiként, a jelentéseket ezen élek találkozási pontjaiként, csomópontjaiként szemléltetve.

A WordNetek közismert gyengesége a poliszém szavak jelentéseinek túl finom megkülönböztetése, ami a gyakorlatban nagyon megnehezíti egy átlagos beszélő számára egy szóalak valamely WordNet-beli fogalomnak való megfeleltetését. A WordNet-beli fogalmak elkülönítése utólag már gyakran nem tűnik motiváltnak, a jelentésegységek sokszor átfedésben vannak egymással. Párhuzamos WordNetekben – mint pl. a HuWN és a PWN, ahol a magyar és angol nyelvű fogalmak egyedi azonosítójukon keresztül meg vannak feleltetve egymásnak – gyakori az a jelenség, hogy egy mindkét nyelven poliszém szó más-más aljelentéssel több párhuzamosított synsetet is "elfoglal" önmagában, anélkül, hogy a jelentések közötti különbségtétel oka nyilvánvaló volna. Így látszólag mindkét WordNetben duplázva, triplázva szerepel egy adott szót tartalmazó synset – fordítási szempontból mindenképp redundáns módon. Az alábbiak jól példázzák ezt az esetet: a Princeton Wordnet 2.0-s verziójában a *give* ige több mint 40 synsetben szerepel (nem kollokációban, hanem önálló igeként), s ebből több mint 20 esetben egymagában szerepel a synsetben, egyéb szinonima nélkül. Ezek közül a synsetek közül többnek is olyan magyar megfelelője van, amelyben az *ad* ige szintén önmagában, szinonima nélkül szerepel. Két ilyen synsetpár jelentését alább idézzük (a definíciót illetve egy példamondatot emelünk ki):

- give:27 def.: estimate the duration or outcome of something
  He gave the patient three months to live.
  ad:16 def.: Időtartamot megbecsül.
  Az orvosok három hónapot adtak a betegnek.
- give:40 def.: allow to have or take
  I give you two minutes to respond.
  ad:12 def.: Valamennyi időt kiszab, illetve engedélyez vmire.
  Öt percet adok neked arra, hogy elkészülj.

A WordNetnek ez a tulajdonsága megnehezíti az egyébként is köztudottan nehéz jelentésannotációs feladatot[2], és rontja annak az esélyét, hogy a jelentésannotációs feladatot végző humán annotátorok közötti egyetértés megüssön egy, a gépi jelentésegyértelműsítéshez elfogadható referenciamértéket.[3]

[2] "Wordsense tagging is one of the hardest annotation tasks." ([3])

[3] Ilyen jellegű kísérletre a tavalyi évben került sor, amikor megvizsgáltuk (l. [7] és [8]), hogy – többek között – a Magyar WordNet igei része mennyire alkalmas arra, hogy egy szövegben a legpoliszémebb igék előfordulásait humán annotátorok egyértelműen beannotálják jelentésekkel. A vizsgálat kiértékelése az annotátorok közötti egyetértés mértékét vette figyelembe. A Magyar WordNet (csakúgy, mint a többi vizsgált adatbázis) ezen a vizsgálaton gyenge eredményt ért el, azaz az annotátorok közötti egyetértés mértéke nem ütötte meg azt a szintet, amelyet általánosan az egyértelmű jelentésannotáció feltételeként szabni szoktak.

Nem minden jelentés-egyértelműsítési feladathoz szükséges azonban, hogy a kiértékelést egy humán annotátorok által jelentésannotált tesztkorpuszhoz képest végezzük. A fordítás szempontjából ugyanis legitimálható módon nincs jelentősége, hogy a jelentéstár "megfelelő" vagy "nem megfelelő" jelentése vezetett-e a helyes fordításhoz. A fenti esetet példáként véve mindegy, hogy az *ad* igét tartalmazó két synset közül melyiken keresztül érjük el a *give* angol megfelelőt. A gépi fordítás olyan speciális, jelentés-egyértelműsítést igénylő feladat, ahol a forrásnyelvi szó célnyelvi fordításának megfelelő volta elegendő kiértékelési szempont. A kiértékelés itt tehát hasonlítható a rossz matematikatanuló módszeréhez: mindegy, hogy hogyan jutunk el a végeredményhez, csak helyes legyen. Tanulmányunkban tehát nem közvetlenül a HuWN-nel végzett jelentés-egyértelműsítés kiértékelése zajlik egy humán annotátorok által jelentésannotált tesztkorpuszhoz képest, hanem annak kiértékelése, hogy a párhuzamos WordNetek alapján kapott lexikális szintű fordítás hogy viszonyul a gépi fordítás során kapott fordítási eredményhez.

2 Célkitűzés és módszertan

Esettanulmányunkban azt vizsgáltuk, hogy a Magyar WordNet mint jelentés-egyértelműsítő rendszer és a vele összekötött Princeton WordNet mint angol szótár tud-e javítani lexikális szinten egy adott a magyar–angol irányú gépi fordítórendszer teljesítményén, illetve hogy szófajonként van-e releváns különbség az eredményekben.

2.1 A felhasznált erőforrások

Kísérletünkhöz a magyar–angol nyelvpárra elérhető gépi fordító szolgáltatások közül a legjobb teljesítményt nyújtó[4] MetaMorpho rendszert választottuk. A MetaMorpho gépi fordítórendszer szabályalapú rendszer, de a transzfer és közvetítőnyelves módszerekkel szemben kizárólag direkt megfogalmazásokból áll. Ezek a direkt megfeleltetések azonban nem direkt módon, hanem az elkülönülő generáló fázisban érvényesülnek. A minták egységesen szolgálnak a nyelvtan és szótár leírására is. A fordító gerincét igei vonzatkeretminták adják, amelyeknek az illesztése kulcsfontosságú a fordítandó mondat szintaktikai elemzése szempontjából. A fordítórendszerben a szófaji egyértelműsítést a forrásnyelvi mondat névszói szerkezeteinek és az igei vonzatkeretminták illesztése biztosítja. Az igei vonzatkeretminták illesztése egyben poliszém igék jelentései közötti jelentés-egyértelműsítést is végez.

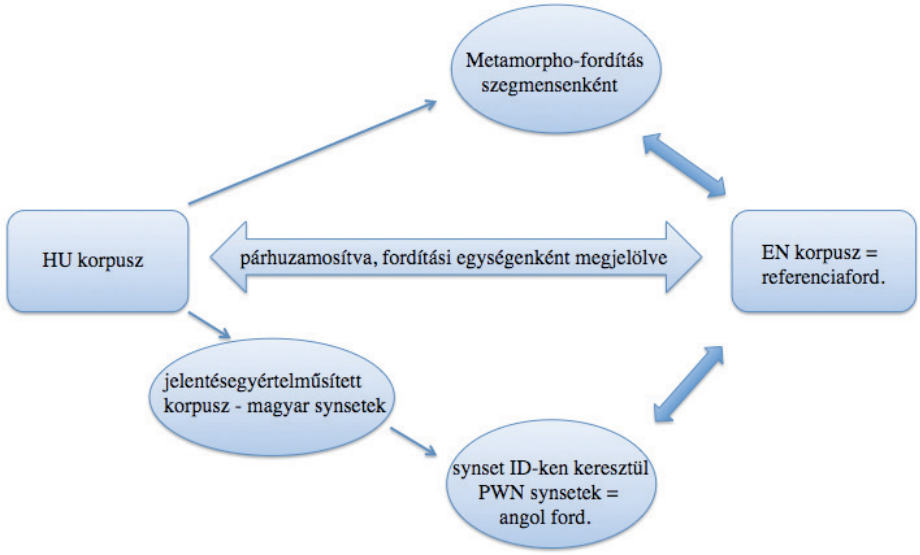
A Magyar WordNet (HuWN) jelenlegi állapotában kb. 37.000 fogalmat tartalmaz, melyeknek nagy része főnév (mintegy 28.500 fogalom), a maradék 8.500 fogalom szófaji megoszlása pedig a következő: 4100 melléknév, 3400 igei, 1000 határozószó. A Magyar WordNet csaknem minden synsete vagy egyedi azonosítóján keresztül, vagy egy ún. nyelvközi relációval meg van feleltetve az angol nyelvű Princeton WordNet 2.0-s verziójának.

[4] A teljesítményt előzetes felmérések alapján becsültük meg.

A vizsgálathoz egy EU-s rövidhíreket tartalmazó párhuzamos korpuszt használtunk, amely mind magyarul, mind angolul kb. 50.000 szövegszó hosszúságú. A korpusz a http://ec.europa.eu/news/ weboldalról származik.[5] Egy-egy rövidhír átlagosan 10 mondatot tartalmaz. A korpusz öt domént ölel fel (mezőgazdaság, pénzügy, kultúra, gazdaság, munkaügy), azaz az általános szókincset hivatott lefedni.

2.2 A munkafolyamat

A vizsgálat a következő fő lépésekből áll: (1) a magyar nyelvű korpusz főneveinek, igéinek és mellékneveinek jelentés-egyértelműsítése a HuWN felhasználásával, majd a beazonosított jelentések (synsetek) angol megfelelőinek kikeresése az angol nyelvű WordNetből, (2) a magyar nyelvű korpusz lefordítása a MetaMorpho fordítóval, majd a fordítás lemmatizálása, (3) az 1. és 2. lépésekből nyert angol fordítások helyességének automatikus kiértékelése a párhuzamos korpusz angol mondataihoz képest, lemmaszinten.[6] A vizsgálat főbb lépéseit az alábbi ábra szemlélteti:



1. ábra. A vizsgálat főbb lépései.

[5] Ezúton szeretnék köszönetet mondani Héja Enikőnek az általa gyűjtött korpusz felhasználhatóvá tételéért, valamint a jelen tanulmányhoz fűzött hasznos megjegyzéseiért.

[6] A vizsgálat elvégzéséhez a következő előfeldolgozó lépésekre volt szükség: (1) a Magyar WordNet XML formátumából a jelentés-egyértelműsítő szoftvernek megfelelő bemeneti fájlokat készíteni, (2) a magyar nyelvű korpuszt a jelentés-egyértelműsítő szoftver által megkövetelt bemeneti formátumra hozni, (3) a rendelkezésünkre álló magyar és angol korpuszt lemmatizálni, (4) a korpuszokat párhuzamosítani (szószintű megfeleltetésre nem volt szükség).

Az ábrán vastag ferde nyilak jelölik a kiértékelés automatikusan végezhető részét: (a fenti felsorolásban a (3)): mind a MetaMorpho gépi fordítórendszer által nyújtott lexikális fordításokat, mind a magyar és angol párhuzamos WordNeteken keresztül kapott lexikális fordításokat összevetettük fordítási egységenként (szegmensenként) a referenciafordításként használt, lemmatizált angol korpusszal.

A magyar korpusz szavainak jelentés-egyértelműsítését a HuWN gráf felhasználásával a Baszk Egyetem által fejlesztett, ingyenesen elérhető, nyelvfüggetlen, UKB nevű eszközzel végeztük (l. [2]), amely a PageRank algoritmust [4] használja fel a jelentés-egyértelműsítés során. Az UKB az általa használt tudásbázist (jelen esetben a WordNetet) gráfként kezeli, melyben a csomópontokat a relatív szerkezeti jelentőségükhöz mérten súlyozza. Az egyes csomópontok súlya a hozzájuk vezető relációktól függ: ha *i* és *j* csomópont között van kapcsolat, akkor *j* súlya *i* súlyának arányában megnő. A program kimenete az adott szóra a szóhoz tartozó csomópontok (itt WordNet synsetek) közül a legnagyobb súlyú. Agirre és Soroa [2] újítása többek között abban áll, hogy az egész tudásbázis gráfját felhasználják, nem csak egy algráfot vonnak be a jelentés-egyértelműsítésbe.

A magyar nyelvű jelentés-egyértelműsítés során a korpusz főneveit, igéit és mellékneveit egyértelműsítettük, azaz az UKB minden ilyen szóhoz hozzárendelt egy-egy WordNet-synsetet. A WordNet-synsetek azonosítóján keresztül eljutottunk a PWN megfelelő fogalmához, azaz a kiindulási szavunk WordNet által javasolt angol nyelvű fordításához. A korpuszt a MetaMorpho fordítóval is lefordítottuk. A kapott gépi fordítást lemmatizáltuk, és kiválogattuk belőle a főneveket, mellékneveket és igéket.

A MetaMorpho fordítót a teljes korpuszszövegen futtattuk le, nem pedig lemmatizált alakokon, mert a fordító mintaillesztése szempontjából szükséges, hogy teljes igei szerkezeteket felismerhessen, és ne csak lemmákat fordító szótárként funkcionáljon a program (pl. egyben felismerjen olyan kollokációkat, mint az *érvénybe lép*). Ilyen szintű kollokációfelismerésre a WordNet a jelen egyértelműsítő algoritmus használata mellett nem volt alkalmas (bár pl. az említett kollokációt tartalmazza), mert az UKB által megkívánt korpuszbemenet lemmatizált alakokat kívánt meg, azaz az esetleges többszavas kifejezéseket eleve elválasztva kezelte, és próbálta egyértelműsíteni.

A hírkorpusz szövegében előforduló főneveknek jelentős hányada tulajdonnév (Named Entity). Ezek, a hírek aktuálpolitikai jellegéből adódóan túlnyomó többséggel olyan esetleges, szótárban nem szereplő nevek, amelyek, ha a kiértékelésnél figyelembe vennénk őket, torzítanák a két fordítási módszer összevethetőségét.[7] Ezért ezeket mind a WordNet, mind a MetaMorpho fordításaiból kiszűrtük.

A gépi fordítórendszer, illetve a jelentés-egyértelműsítő algoritmus jellegét tekintetbe véve előzetes elvárásunk az, hogy az igék esetében a MetaMorpho rendszer fog jobb fordítást nyújtani – lévén, hogy a fentiekben említett igei kollokációfelismerésre a WordNet számára a jelen vizsgálatban nem volt esély. Főnevek esetében feltételezhető, hogy a WordNeten keresztül kapott fordítások nagyobb arányban lesznek jobbak, mint a MetaMorpho rendszeréi, mivel az UKB egy adott szó egyértelműsítésekor a szó szövegkörnyezetében előforduló, vele valamilyen szemantikai viszonyban lévő

[7] A MetaMorpho rendszer ugyanis, ha nem ismer fel egy szót, visszadja azt kimenetként – azaz tulajdonnevek esetében automatikusan, valódi fordítási eljárás nélkül helyes kimenetet produkál.

szavakat vizsgálja, és veti össze a HuWN gráffal, és a HuWN főnévi gráfja kellően nagy ahhoz, hogy sikeresnek tételezzük fel ezt a műveletet. A melléknevek esetében kb. egyforma teljesítményre számítunk a két fordítórendszer részéről.

3 Eredmények és kiértékelés

Az automatikus kiértékelő lépésben a két módszerrel kapott fordításokat összevetettük a lemmatizált angol korpusszal, amit emberi fordításként, referenciafordításként kezeltünk. Mind a két gépi fordítási kimenetben (itt tág értelemben véve a WordNet által kínált lexikális fordításokat is gépi fordításnak nevezzük) megőriztük azokat a szegmenshatárokat, amelyek a magyar és angol korpusz párhuzamosításakor mint fordításiegység-határok születtek. Így az összehasonlítás szegmensenként történt. Amennyiben egy adott szegmens egy lexikális egységének volt megfelelője az angol korpusz megfelelő szegmensében, azt automatikusan jó fordításnak könyveltük el.[8] Az ilyen találat hiánya azonban nem tekinthető automatikusan rossz fordítás indikátorának – hiszen egy szövegnek többfajta helyes fordítása is létezhet –, pusztán kézi kiértékelést tesz szükségessé.

A vizsgálatot az a kérdés vezérelte, hogy a HuWN felhasználása a lexikális jelentés-egyértelműsítésben javíthat-e lexikális szinten a magyar-angol irányú gépi fordítás minőségén. A fordítás eredményét pontosság (precision) alapján mértük, azaz a jól fordított szavak és az összes lefordított szó arányára voltunk kíváncsiak, s ezt hasonlítottuk össze a két módszerrel kapott fordítások esetében. Jelen vizsgálat során a fedés (recall) mérése nem volt alkalmazható, a következők miatt. A WordNet által kapott fordítások esetében meg lehet határozni, hogy a korpusz összes főnévi, melléknévi és igei szavának mekkora százalékára készült fordítás. Ugyanez azonban nem állapítható meg a MetaMorphóval készült fordítás esetében. A MetaMorpho rendszer ugyanis nem szavakat, hanem mondatokat fordított, s ily módon nem feleltethetők meg egymásnak egyértelműen a magyar szöveg és az angol fordítás szavai. A MetaMorphóval készült angol fordítás összesen jóval több főnevet, melléknevet, igét tartalmaz, mint a magyar korpusz – míg a WordNettel készült fordításra ez nem igaz, hiszen ott lexikális elemeket fordítottunk. Az arányok összehasonlítása tehát értelmetlen lenne a két esetben.

A kiértékeléskor tokenek arányát vettünk figyelembe, bár ez implikálja, hogy egy gyakoribb előfordulású szó nagyobb súllyal szerepel a kapott pontossági értékben. Ennek ellenére, mivel szövegek fordítási minőségéről van szó, ennek a súlyozásnak van létjogosultsága.

A WordNeten keresztül kapott fordítások esetében az esetleges többszavas kifejezések helyességének automatikus ellenőrzésére két út is kínálkozott: a referenciafordítással való összehasonlításkor egyrészt tekinthettük illeszkedésnek, ha az adott többszavas tagjai közvetlenül egymás mellett jelentek meg, de megengedőbb esetben

[8] E mögött az az előfeltevés húzódott meg, hogy amennyiben ugyanaz a szó szerepel az emberi fordítás és egy „gépi" fordítás ugyanazon fordítási egységében, jogunk van feltételezni, hogy ugyanannak a szónak a fordításáról van szó, nem pedig puszta véletlenről.

azt is, ha a tagok bárhol szerepeltek a megfelelő referenciaszegmensben.[9] Alább bemutatjuk mindkét úton kapott eredményeket.

1. táblázat: Az automatikus kiértékelés eredményei: a két fordítási módszer pontossága.

	HuWN pontossága a többszavas kif-ek pontos illesztésével	HuWN pontossága a többszavas kif-ek laza illesztésével	MetaMorpho pontossága
főnév	31,69%	31,81%	32,24%
melléknév	28,13%	28,27%	32,96%
ige	15,22%	15,28%	20,12%
össz.:	**28,12%**	**28,22%**	**28,97%**

A fentiek fényében elmondhatjuk, hogy az eddig elvégzett automatikus kiértékelés alapján nagyságrendileg mindkét fordítási módszer hasonló eredményt nyújtott. A MetaMorpho rendszer mind összesítésben, mind szófajokra bontva jobban teljesített, mint a két párhuzamos WordNet mint lexikális fordító. Előzetes elvárásunk, miszerint a MetaMorpho az igei jelentés-egyértelműsítő mechanizmusának köszönhetően az igék esetében jobb fordítást nyújt majd, mint a WordNeteken keresztül kapott fordítások, annyiban is beigazolódott, hogy a két rendszer pontossága közötti nagyságrendi különbség az igék esetében a legnagyobb.[10] Azon hipotézisünk, miszerint a főnevek esetében a WordNeteken keresztül kapott fordítások bizonyulhatnak jobbnak, nem igazolódott be, bár nyilvánvaló, hogy nagyságrendileg a főnevek esetében közelíti meg egymást leginkább a két módszer pontossága.

4 További munkálatok

Ahhoz, hogy a jelenlegi kiértékelésnél megbízhatóbb eredményt kapjunk, természetesen szükséges az automatikusan nem kiértékelhető fordítások (legalább egy részének) kézi kiértékelése. Érdekes lenne azt is megvizsgálni, hogy a két fordítórendszer hibaeseteiben van-e felismerhető tendencia: hasonló esetekben adnak-e rossz fordítást, vagy komplementer esetekben. További kutatás tárgya lehetne egyrészt az ellentétes irányú nyelvpáron (angol–magyar) lefuttatott hasonló kísérlet ugyanezzel a két fordí-

[9] A MetaMorpho által nyújtott fordítást lemmatizálás után tudtuk csak összevetni a referenciafordítással, így ebben a lépésben sajnos mindenképp külön tagokra bontódtak fel az esetleges többszavas kifejezések. Tagjaikat azonban külön-külön természetesen meg lehetett találni az angol korpusz megfelelő mondatában.

[10] A Szlovén WordNet és egy szlovén-angol gépi fordító program (Presis) viszonylatában Fišer és Vintar [6] hasonló kísérletet végzett, amelyen valamivel jobb eredményt el a WordNet által nyújtott fordítás. Ez valószínűleg annak tudható be, hogy a Presis fordítórendszerben semmiféle jelentésegyértelműsítés nincs beépítve.

tórendszerrel, valamint más, magyar–angol / angol–magyar nyelvpárra elérhető fordítók és a magyar–angol párhuzamosított WordNet teljesítményének összehasonlítása.

Bibliográfia

1. Agirre, E., Edmonds, Ph.: Word sense disambiguation. Algorithms and Applications. (Text, Speech and Language Technology). Springer-Verlag New York, Inc., Secaucus, NJ (2007)
2. Agirre, E., Soroa, A.: Personalizing PageRank for Word Sense Disambiguation. In: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece (2009)
3. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics Vol. 34 No. 4 (2008) 555–596
4. Brin S., Page L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems Vol. 30 No. 1-7 (1998)
5. Fellbaum, C.: WordNet An Electronic Lexical Database. MIT Press (1998)
6. Fišer D., Vintar, Š.: Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajanju. In: Proceedings of the 13th International Multiconference Information Society - IS 2010 (2010)
7. Héja, E., Kuti, J., Sass, B.: Jelentésegyértelműsítés - egyértelmű jelentésítés? In: Tanács A., Szauter D., Vincze V. (szerk.): MSZNY2009, VI. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2009) 348–352
8. Kuti, J., Héja, E., Sass, B.: Sense disambiguation - "Ambiguous sensation"? Evaluating sense inventories for verbal WSD in Hungarian. In: Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages (2010)