Magyar jelentés-egyértelműsített korpusz

Szarvas György[1], Hatvani Csaba[2], Szauter Dóra[1],
Almási Attila[1], Vincze Veronika[1] és Csirik János[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{szarvas, csirik}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
szauter.dora@freemail.hu, vizipal@gmail.com,
{vinczev, hacso}@inf.u-szeged.hu

Kivonat: Az első magyar WSD korpusz elkészítéséhez 39 olyan szóalakot választottunk ki, melyek jó mintapéldák a jelentés-egyértelműsítés feladatának vizsgálatára. A kiválasztásnál a kritériumok között szerepelt, hogy az adott szóalak legyen gyakori a magyar nyelvben (ennek mérésére a *Magyar Nemzeti Szövegtár (MNSZ)* [8] gyakorisági adatait használtuk), illetve, hogy legyen több, használatában gyakorinak tekinthető jelentése. A korpusz szövegeit is az MNSZ-ből, annak Heti Világgazdaság (HVG) számaiból összeállított részkorpuszából válogattuk. Így minden egyes példához rendelkezésre áll a vizsgálat szempontjából releváns kontextus (teljes HVG-cikk), illetve automatikus tokenizálás, szófaji kódolás, szótőre vonatkozó információ.

1 Jelentés-egyértelműsítés

A jelentés-egyértelműsítés (Word Sense Disambiguation, WSD) problémája alatt a szövegekben előforduló többértelműségek (homonímia, illetve poliszémia) feloldásának feladatát értjük. A többértelműség feloldásának problémája egyidős a gépi szövegfeldolgozással, és a legtöbb nyelvtechnológiai alkalmazás (pl. szövegmegértés, ember-gép párbeszéd, gépi fordítás, információ-visszakeresés, illetve -kinyerés) számára fontos köztes feladat.

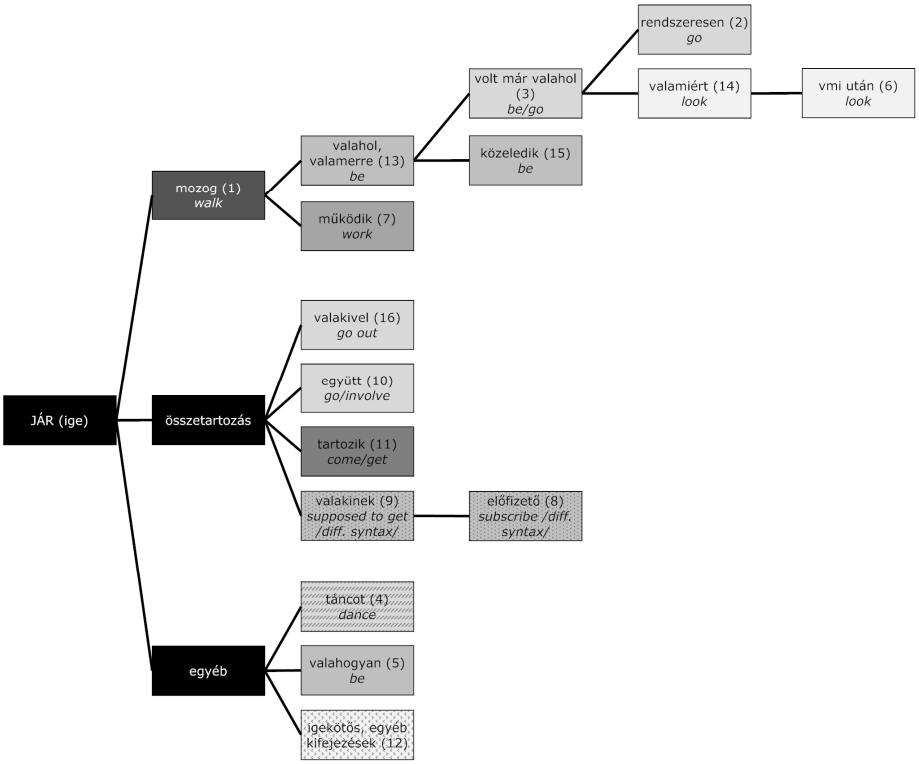
1.1 Kapcsolódó eredmények, áttekintés

Jelentés-egyértelműsítési kutatások más nyelveken
A kezdetben angolra, majd a későbbiekben más nyelvekre folytatott jelentés-egyértelműsítési kutatások nagyrészt kapcsolódtak az ACL-SIGLex által szervezett SensEval [4], [5] workshopokhoz. A 2006-ban megjelent Word Sense Disambiguation [1] című könyv, valamint a SensEval sorozat folytatásaként 2007-ben megrendezett SemEval workshop kiadványa [2] részletes áttekintést ad az eddigi eredményekről.

Jelentés-egyértelműsítési kutatások magyarra
Az angol–magyar, illetve magyar–angol fordítórendszerek fejlesztése kapcsán hosszú ideje foglalkoznak a jelentés-egyértelműsítési feladatokkal magyar nyelven, a fordítórendszer eredményének javítása érdekében [6], [7].

1.2 A jelentés-egyértelműsítési feladat

A jelentés-egyértelműsítő eljárások az alkalmazhatóságuk határai alapján és a jelentésmegkülönböztetés foka szerint két-két főbb csoportra oszthatók. Hatókör tekintetében a teljes szókincsre alkalmazható (all-words WSD) és előre megadott szóalakokon működő (lexical sample WSD) módszereket különböztethetünk meg, míg a jelentésmegkülönböztetés részletessége szerint aprólékos vagy finom (fine grained), illetve durva (coarse grained) szinteket különböztethetünk meg. Az alábbi ábra szemlélteti a *jár* ige jelentéseinek különböző felbontási szintjeit – minden címke egy önálló elkülöníthető jelentés, melyek azonban 3 nagyobb csoportba oszthatók, amiket a színekkel illusztráltunk:



A Magyar WordNet projekt során elkészült – itt ismertetett – korpusz kijelölt szóalakokra részletes felbontású jelentésannotációt tartalmaz (fine grained lexical sample corpus).

1.3 Címkézési elvek

A munka első fázisában a megadott 39 szóalak lehetséges jelentéseit adtuk meg. Az egyes jelentések meghatározásában segítségül hívtuk a Magyar Értelmező Kéziszótár papír és elektronikus változatát, valamint nyelvi intuíciónkat is. Két külön jelentésnek vettük azokat az eseteket, amelyek valamelyik szótári definíció szerint határozottan elkülöníthetők[1].

A nemzetközi gyakorlatnak megfelelően a korpusz példáinak címkézését két független annotátor (képzett nyelvész) is elvégezte. Független alatt azt értjük, hogy a munka elvégzése során tilos volt a nyelvészek közötti kommunikáció. A páros jelölés segítségével egyrészt lehetővé vált az adatbázis konzisztenciaszintjének mérése, másrészt az esetleges jelölési hibák konszenzussal javíthatók.

Fontos kitétel volt, hogy a szóalakokat csak egy adott szófaj keretén belül címkéztük. Pl. a *pont* szónak csakis főnévi jelentéseit címkéztük. Határozószóként - *pontosan* jelentésben – nem címkéztük, azaz poliszémiát vettünk figyelembe, homonímiát nem. Ugyanezen okból nem lett felvéve a *század* szó *századrész* jelentése, mivel ez nem főnév, hanem törtszámnév.

2 A korpusz bemutatása

Ebben a fejezetben ismertetjük az elkészült korpusz főbb technikai, illetve tartalmi jellemzőit.

2.1 A korpusz összetétele

A korpusz építésekor minden szóalakra 350-500 példa címkézését tűztük ki célul, ezzel a mérettel az egyes szóalakokra készülő részkorpuszok méretben összevethetők a más nyelvekre elérhető adatokkal. Az elkészült adatbázisban azonban benne hagytuk azokat a példákat is, melyek végül nem kerültek kézi egyértelműsítésre.
A kiválasztott szóalakok a következők:
melléknév: *anyagi, élő, erős, képes, pontos, szociális*
főnév: *család, élet, ház, helyzet, intézmény, iskola, kép, képviselő, kormány, nap, oldal, ország, perc, pont, program, század, személy, szervezet, tanár, világ, víz*
ige: *függ, hat, jár, kap, kerül, marad, rendelkezik, szerepel, tart, tartozik, tud, válik*

A 39 szóalak átlagos jelentésszáma igen magas (átlag 6 jelentés szóalakonként), melyből a korpusz anyagát képező szövegekben átlagosan csak 5 jelentés jelenik meg. Ha azonban nem vettük számításba az elhanyagolható mértékben (1-2%-ban) jelen levő jelentéseket, akkor az átlagosan megjelenő jelentések száma még kevesebb, 3,7 (mérsékeltebb a többértelműség). Külön érdekes a *tanár* szóalak, mely a vizsgált szövegben teljesen egyértelműnek bizonyult annak ellenére, hogy a szóalakok válo-

[1] A későbbiekben felmerült két újabb kritérium is, miszerint ha az adott jelentéseknek két külön szó feleltethető meg egy másik nyelvben, illetve ha a szóalak más-más vonzatkerettel fordul elő, akkor is külön jelentésnek vesszük fel.

gatásánál követelmény volt, hogy legyen több, a nyelvben gyakran használt jelentésük. Némely szó azonban a homogén nyelvhasználat ellenére is igen sokféle formában megjelenik, mint a *jár* ige, melynek 16 jelentéséből 14 előfordul a szövegben.

2.2 A korpusz formátuma

A korpusz építésekor a *SensEval/SemEval* (Association for Computational Linguistics által szervezett) nemzetközi konferencia workshopokon WSD-feladatokhoz készített korpuszok formátumát követtük. Ezzel a választással egyrészt egy meglévő XML-formalizmust vettünk át, így nem kellett az adatformátumot tervezni, másrészt a szabvány adattárolás remélhetőleg megkönnyíti a korpusz terjesztését is.

Egy példa a korpuszból:

```
<instance id="jár.V.mnsz.01" docsrc="press-hvg.1">
        <answer instance="jár.V.mnsz.01" senseid="jar_v_5_valahogyan"/>
        <context>
        Ez azonban a dolognak már csak a technikája és nem a tartalma volt . Az üzenet , amely az első
        forduló után az akkor még csak kvázigyőztesek szájából megfogalmazódott , és amely a jelek
        szerint a választáson részt vevők többségénél meghallgatásra talált , körülbelül így szólt : az
        MSZP az elmúlt négy évben meghatározó pozícióból folytatott kormányzati politikájával betöl-
        tötte a társadalmi-gazdasági átmenet időszakában neki osztott szerepet .
        Elvégezte azt az egyébként hálátlan feladatot , a stabilizációt , ami nélkül egyetlen kelet-közép-
        európai országnak sincs esélye a felemelkedésre . A polgári átalakulás további vezényletét vi-
        szont az abban érintettek többsége a legalábbis külsőre frissebbnek , dinamikusabbnak tűnő és a
        múlt relikviáitól , a már egyszer eldobott maníroktól nem ( vagy legalábbis kevésbé ) terhes poli-
        tikai erőkre és személyekre kívánja bízni .
        Ezért is tűnik e pillanatban irrelevánsnak annak felvetése , hogy nem <head>jártak</head> vol-
        na -e jobban a szocialisták , ha Horn Gyulát időben katapultálják a pártelnökségből , de leg-
        alábbis a miniszterelnök-jelöltségből , és mondjuk a külügyminiszteri tevékenységével és szemé-
        lyes karakterével , amellett fiatalabb korával a választóközönség számára esetleg vonzóbb Ko-
        vács László vezényletével vágnak bele a választási kampányba .
        Az MSZP a jelek szerint így is a lehetséges maximumot hozta ki magából , legalábbis ami az el-
        ső forduló eredményét illeti .
        </context>
</instance>
```

2.3 A korpusz főbb statisztikai jellemzői

Az alábbi táblázatban foglaltuk össze a korpusz főbb statisztikai jellemzőit. Részletesebb elemzések az interneten érhetők el.

1. Táblázat: A korpusz főbb adatai

	Szóalakok	Duplán annotált példa	Szimplán annotált példa	Annotáció nélkül
Melléknév	6	2087	462	688
Főnév	21	6853	2714	11459
Ige	12	3537	1898	13501
Összesen	39	12477	5074	25648

2.4 Elérhetőség

A korpusz első változata a *Magyar ontológia építése és alkalmazása információkinyerő rendszerekben* [3] projekt keretében készült el. A korpusz – kutatási és oktatási célokra – szabadon hozzáférhető, letölthető a www.inf.u-szeged.hu/hlt oldalról.

3 A korpusz értékelése

Ebben a fejezetben kiértékeljük a korpusz jelentés-annotációinak konzisztenciáját, ismertetjük a két párhuzamos jelölés eltéréseinek egységesítése során követett protokollt, valamint egy egyszerű vektortérmodellen alapuló osztályozó segítségével megmutatjuk, hogy a jelentés-egyértelműsítés kihagyása esetén az alkalmazások által használható leggyakoribb jelentésnél jobb eredmények érhetők el.

3.1 Annotátorok egyetértési rátája (konzisztencia-ellenőrzés)

A korpusz készítésekor első lépésben a megkülönböztetni kívánt jelentések halmazát definiáltuk minden szóalakra, melyeket rövid szöveges leírással (definícióval) láttunk el. A Magyar WordNet állományát is bővítettük a korpuszban használt, a HuWN állományából még hiányzó jelentések synsetjeivel. A nemzetközi gyakorlatnak megfelelően a korpusz példáinak címkézését két képzett nyelvész is elvégezte, egymástól függetlenül.

Az annotátorok egyetértési rátája, azaz a címkézési feladat szakértők általi elvégzésének pontossága azokban az esetekben alacsonyabb, amikor a leggyakoribb jelentés részaránya nem túl magas. Ezekben az esetekben az egyértelműsítés igen nehéz feladat (hisz az egyetértést képzett nyelvészek közt mértük). Másrészt a nyelvtechnológiai alkalmazások, mint a gépi fordítórendszerek, információkinyerő rendszerek stb. éppen ezekben az esetekben profitálnának leginkább egy hatékony WSD-megoldásból, a leggyakoribb jelentés választása helyett. Az annotátori egyetértési ráta a teljes korpuszra nézve 84,78%-os volt.

Az egyik legnehezebb feladat az volt, hogy az annotátor következetességét a szóalak címkézése során végig megőrizze. Ha egy kérdéses esetben az annotátor egy adott jelentés mellett döntött, akkor ugyanazon címkével lássa el a szóalakot egy későbbi, hasonló kontextusban történő előfordulásakor is. Pl. ha X annotátor a BUX tőzsdeindex *pontját* címkézte a *pont_2: az értékelés egysége* jelentéssel első előfordulásakor, akkor ugyanígy kellett eljárnia az összes esetben, még akkor is, ha az egyes esetek egymástól „nagy távolságra” helyezkedtek el[2].

A gyakorlatban a jelentések nem mindig lettek a legprecízebbek, nem mindig tükrözték az elméleti vagy szótári jelentés-megkülönböztetést. Sokszor túl finom, nehezen megkülönböztethető jelentésárnyalatok is fel lettek véve, ezzel tovább romlott az

[2] Az annotátor még önmagán „belül” sem mindig konzisztens, nemhogy másokkal összehasonlítva. Épp ezért bizonyos esetekben következetlenségek előfordulhatnak a korpuszban. Az egyes távoli esetek fejben tartása a kézi annotálás egyik nagy nehézsége.

egyetértési mutató, és az annotátor saját következetessége is csorbát szenvedhetett. A nagyobb következetesség elérése céljából a rendszer tovább finomítható.

Bizonyos esetekben a morfológiai elemző nem megfelelően kategorizálta a szóalakot. Pl. a *vált* jelen idejű igealakot a *válik* múlt idejének elemezte, így felkínálta címkézésre a *válik* jelentései közé. Ebben az esetben a szóalakot nem címkéztük.

A szövegek tematikájából adódóan bizonyos jelentések sokkal gyakrabban fordultak elő a többinél. Mivel a korpusz HVG-szövegekre épül, például a *kormány* szó előfordulásainak túlnyomó többsége a 'politikai kormány' jelentést hordozza. Ha azonban a korpuszban lennének például autókról szóló szövegek, az 'irányító szerkezet' jelentés aránya rögtön megnőne.

Külön problémát jelentettek a kollokációk és a szólások, közmondások, mert sok esetben lehet tudni, hogy a kifejezésen belül milyen jelentésben szerepel az adott szó. Például a *sok víz lefolyik a Dunán addig* szólásban a víz jelentése pontosan azonosítható: *víz_2: a föld felszínének valamely részét borító folyadéktömeg*, kérdés azonban, hogy ezt a címkét kapja-e, vagy pedig *egyéb* címkével lássuk el, mivel kifejezés része.

3.2 A korpusz címkézésének véglegesítése

A javítási munkaszakaszban egy harmadik független annotátor nézte át azokat az eseteket, amikor a két annotátor címkézésében eltérés mutatkozott, és véglegesítette a problémás esetek címkéit. Így a korpusz azon részének címkézése, melyre duplán rendelkezésre állt jelentésannotáció, a lehetőségekhez mérten konzisztens. Azon példákat, ahol csak egy címkézés készült el, nem ellenőríztük.

Az eltérések nagy többsége abból adódott, hogy az annotátorok máshogy értelmeztek bizonyos, egymást részben fedő jelentéseket (ez egyben utalás is arra, hogy e jelentések szétválasztása talán nem teljesen indokolt). Például: *jár_6: valami után jár, megszerzésén fárad* és *jár_14: valamiért több helyre is elmegy*. A legtipikusabb eltérés amikor az egyik annotátor egy adott szóalakot egy adott kontextusban *egyéb* címkével látott el, míg a másik annotátor úgy érezte, hogy a szó adott előfordulása még belefér egy pontosabban meghatározott jelentéstartományba. Például a *kap* igénél gyakori volt, hogy az egyik annotátor a *jogdíjat kap* kifejezésben (és hasonló esetekben) a *kap_1: valamit adnak neki* címkét jelölte meg, míg a másik az *egyéb* címkét választotta.

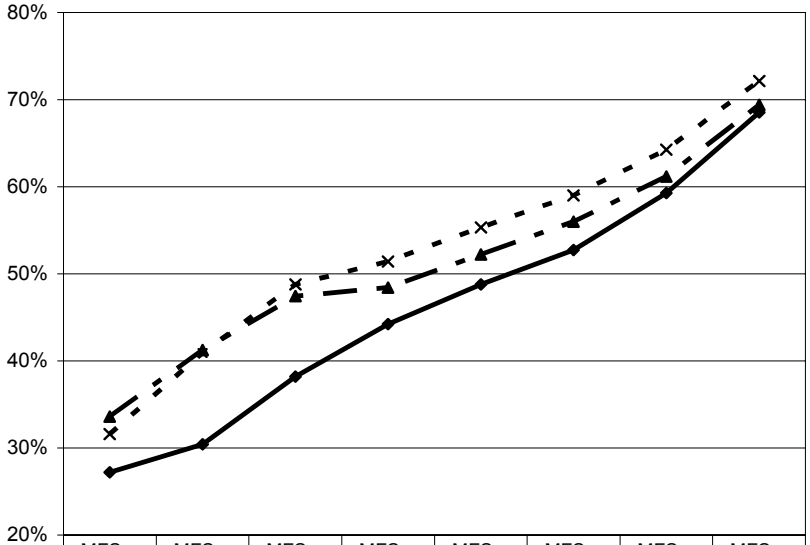
Előfordult néhány olyan eset is, amikor a harmadik annotátor nem értett egyet a két meglévő címkézés egyikével sem. Ilyenkor – a lehetőségekhez mérten – a három annotátor együttes megbeszélése és egyeztetése alapján alakult ki a végleges címke. Egy példa: *a víz felforralása vagy – erre szolgáló filtereken való – átszűrése* kontextusban a két annotátor a *víz_1: élethez nélkülözhetetlen folyadék*, illetve a *víz_2: a föld felszínének valamely részét borító folyadéktömeg* jelentéseket adták meg, a harmadik annotátor viszont a *víz_3: ivóvíz, fürdővíz* jelentésre szavazott. (A végső döntés a *víz_3* címke lett.)

Az eltérések jelentősen kisebb hányada nyilvánvaló tévesztésből fakadt: az egyik annotátor véletlenül a szomszédos címkére kattintott, vagy kifelejtette (azaz jelöletlenül hagyta) az adott szóalakot.

3.3 Baseline mérések, C4.5, naiv Bayes osztályozók

A jelentés-egyértelműsítő eljárások értékelésekor baseline pontosságként a leggyakoribb jelentés részarányát célszerű tekinteni (továbbiakban MFS), hiszen ez a triviálisan elérhető pontosság. Egy eljárás által adott címkézés (egyértelműsített szóelőfordulások) akkor tekinthető értékelhetőnek, ha a leggyakoribb jelentés részarányánál nagyobb hányadban rendeli a szóalakokhoz a megfelelő jelentést.

A felügyelt tanulási modellek építéséhez szükséges a feladat példáinak a tanuló algoritmus számára kezelhető formátumra való konvertálása. Kísérleteink során a példák leírására csak a címkézett szóalak közvetlen környezetét (egyetlen bekezdés) használtuk fel, illetve ábrázolására a jól ismert *vektortérmodell*t használtuk. A jelentés-egyértelműsítési feladat esetében ez a reprezentáció nyilvánvalóan túlzott egyszerűsítés, hiszen a célszó közvetlen környezete, a mondattani szerepek sokszor kiemelten fontosak az egyértelműsítésben, és ez az adat itt elvész. Eredményeinket a korpusz terjesztéséhez mint összehasonlítási alap szántuk.



	MFS < 30%	MFS < 40%	MFS < 50%	MFS < 60%	MFS < 70%	MFS < 80%	MFS < 90%	MFS < 100%
MFS	27,20%	30,42%	38,21%	44,23%	48,80%	52,72%	59,28%	68,55%
BAYES	33,60%	41,23%	47,44%	48,40%	52,21%	55,98%	61,18%	69,40%
C4.5	31,60%	40,91%	48,78%	51,42%	55,31%	59,00%	64,27%	72,14%

1. ábra: Az MFS, C4.5 és naív Bayes osztályozók pontosságai a leggyakoribb jelentés részarányának függvényében.

Tapasztalataink azt mutatják, hogy a statisztikai modellek jobban szerepelnek az alkalmazások által jelenleg használt *leggyakoribb jelentés* heurisztikánál (lásd 1. ábra). A pontosságbeli különbség jelentős a korpuszon a bonyolultabb célszavak esetén, ahol a *leggyakoribb jelentés* jelentősen elmarad az elvi maximális pontosságtól, melyet az annotátorok egyetértési rátájával mértünk.

4 Javítási és finomítási lehetőségek

A nagyobb következetesség elérése céljából a rendszer tovább finomítható az azonos jelentések csoportos átnézésével:

- Végig kell nézni egy adott jelentés címkéjével ellátott összes szóalakot és összevetni őket egymással, hogy valóban következetes-e a címkézés az adott jelentéstartományon belül.
- Ezt minden jelentésnél érdemes elvégezni!
- Ami kilóg a sorból, azt újra kell címkézni!

A jelentésárnyalatok finomságát, a különféle jelentések megkülönböztetését érdemes lehet újragondolni (l. a *jár* példája). Más témájú szövegek annotálása is hasznos lehet a jelentések gyakoriságának meghatározásában.

A teljes szókincsre alkalmazható (all-words) WSD létrehozása nehéz feladat, óriási energiabefektetést igényel, hiszen a teljes magyar szókincsre ki kellene dolgozni a lehetséges jelentéseket. Tovább nehezíti a feladatot, hogy időnként a szókapcsolat együtt hordoz bizonyos jelentést a kontextusban, és a szóalakról önmagában nehéz eldönteni, hogy hordozza-e az adott jelentést – például az *A védőoltások növekedése terén pedig erős képzelőerő kell a tényleges kormányzati cselekvési mező megtalálásához.* mondatban a kontextus egésze hordozza a túlzó, ironikus jelentést, nem kizárólag az *erős* szóalak.

Bibliográfia

1. Agirre, E., Edmonds, P.: Word Sense Disambiguation – Algorithms and Applications, In Ide, N. and Véronis J., editors: Text, Speech and Language Technology Series, Volume 33, Springer, Dordrecht, The Netherlands (2006)
2. Agirre, E., Márquez, L., Wicentowski, R.: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic (2007)
3. Hatvani, Cs., Kuti, J., Miháltz, M., Szarvas, Gy.: GVOP 3.1.1-2004-05-0191/3.0 – Magyar ontológia építése és alkalmazása információkinyerő rendszerekben, projektzáró összefoglaló jelentés, Technical Report, Szeged, Hungary (2007)
4. Kilgariff, A.: Proceedings of Senseval 2: Second International Workshop on the Evaluating Word Sense Disambiguation Systems, Association for Computational Linguistics, Toulouse, France (2001)
5. Mihalcea, R., Edmonds, P.: Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelon, Spain (2004)
6. Miháltz, M.: Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation. In Proceedings Modern Approaches in Translation Technologies Workshop at RANLP-2005, Borovets, Bulgaria (2005)
7. Miháltz, M., Póhl, G.: Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból. In Proceedings of III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary (2006)
8. Váradi, T.: Szótár, Korpusz – Magyar Nemzeti Szövegtár. In Gecső, T., editor: Lexikális jelentés, aktuális jelentés. Segédkönyvek a nyelvészet tanulmányozásához IV. Tinta Kiadó, Budapest, Hungary (2000)