NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével

Miháltz Márton[1], Naszódi Mátyás[1], Vajda Péter[2], Varasdi Károly[2]

[1] MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5,
{mihaltz, naszodim}@morphologic.hu
[2] MTA Nyelvtudományi Intézet, 1399 Budapest, Benczúr u. 33.
{vajda, varasdi}@nytud.hu

Kivonat: A cikkben bemutatunk egy tudásalapú anafora-feloldó rendszert, mely személyes névmások, zéró névmások, határozott névelős köznevek, valamint tulajdonnevek közötti koreferencia-viszonyok azonosítását végzi. A rendszer mély szintaktikai elemzésre, szintakszis-elmélet tételekre, pszicholingvisztikai kutatásokra, valamint a Magyar WordNet ontológiában tárolt nyelvi tudásra támaszkodik. A bemutatott módszerek a különböző típusú jelenségeket átlagosan 70%-os pontossággal képesek kezelni.

1 Bevezetés

Természetes nyelvű szövegekben az NP-koreferenciák feloldása egy adott dokumentumban eltérő pontokon megjelenő, de azonos entitásra referáló főnévi csoportok (NP-k) közötti viszonyok azonosítását jelenti. A feladat megoldása fontos olyan nyelvtechnológiai alkalmazások számára, mint a gépi fordítás, az információkivonatolás és egyéb szövegfeldolgozó alkalmazások ([7]).

A cikkben bemutatott, jelenleg még folyamatban lévő munka az alábbi koreferencia-jelenségek kezelésére – a visszautaló elem (anafora) és a szövegben korábban előforduló, vele koreferens NP (antecedens) közötti kapcsolat azonosítására – tesz kísérletet magyar nyelvű szövegekben ([10] alapján):

1. Táblázat: a vizsgált koreferencia-típusok, példák (az egymással koreferens NP-k félkövérrel kiemelve)

Típus	Példa
Ismétlés	Tegnap találkoztam **egy ismerősömmel**. **Az ismerősöm** nagyon sietett, mindössze pár percet beszéltünk.
Tulajdonnév-variáns	**Kovács Jakab** tegnap sajtótájékoztatót tartott. Az eseményen **Kovács úr** bejelentette az új termékeket.
Szinonima	Tamás kapott **egy biciklit**. Én is láttam **a kerékpárt**.
Hiper-/hiponima	Bejött **egy puli**. **Az állat** fáradtnak tűnt.
Névmás	Beszéltem **Julival**. Megadtam **neki** a számodat.
Zérónévmás	**Viktor** ismeri **Ferit**, de (**ő**) nem kedveli (**őt**) túlságosan.

Jelenleg nem foglalkozunk a személyes névmáson és bizonyos mutató névmásokon kívüli egyéb névmás-típusok (visszaható és kölcsönös névmások, vonatkozó névmások stb.) feloldásával. Főnévi csoportok alatt a mondatban előforduló maximális NP-ket értjük, melyek jellemzően a mondat főigéjének vonzatai, illetve főnévi eredetű szabad határozói. Jelenleg nem foglalkozunk a komplex, hierarchikus szerkezetű főnévi csoportok (pl. birtokos szerkezetek, koordinált NP-k stb.) összetevőivel. Így jelenleg nem foglalkozunk a birtokos szerkezetekben a morfológiailag jelölt számú-személyű, de a mondatban a birtoktól akár távolabbra is kerülő, vagy akár hiányzó birtokosnak megfelelő szerkezetek azonosításával. Szintén nem foglalkozunk az anaforát a szövegben követő antecedensű koreferencia-típus (katafora), valamint az epithetonnak nevezett jelenség („***Balázs*** *nem találta a kulcsát.* ***A szerencsétlen*** *nem tudott bejutni a lakásba.*") kezelésével.

A következő részben bemutatjuk a jelenleg működő, szabályalapú megközelítést alkalmazó koreferencia-feloldó rendszerünket, majd a kiértékelésére kialakított, annotált korpuszokat használó környezetet. Az utolsó részben részletesen ismertetjük a további lehetséges fejlesztések irányát.

2 A koreferencia-feloldó algoritmus

A koreferencia-feloldás kutatásának irodalmában az utóbbi időben megfigyelhető hangsúlyeltolódás a tudásalapú megközelítésektől az adatalapú, gépi tanulásra támaszkodó, a szabályalapú rendszerek teljesítményével vetekedő megközelítések felé ([8]). Számunkra azonban a munka kezdetekor nem állt rendelkezésre magyar nyelvre az adatalapú megközelítésekhez elengedhetetlen, jellemzően több ezer, kézzel annotált példából álló tanítókorpusz, így kénytelenek voltunk a tudásalapú megközelítések felől indítani.

Rendszerünk többféle tudásra támaszkodik. A legfontosabb inputot a MetaMorpho fordítóprogram-projektben fejlesztett magyar mondatelemző elemzéseiben kapott morfológiai, szintaktikai és szemantikai jegyek, nyelvtani szerepek, mélyszerkezeti elemzési struktúrák stb. jelentik. Ezekre támaszkodnak a kötéselmélet ([4]) és a magyar mondatmegértés kutatásainak ([9], [10]) eredményeire támaszkodó szabályaink. További, világismereti tudásra alapuló szabályok forrásaként a magyar WordNet ontológiát ([3]) használjuk. Végül a tulajdonnevek közötti referencia-azonosság felismeréséhez karakteralapú megközelítéseket alkalmazunk.

A feldolgozandó dokumentumokban balról jobbra haladva vizsgáljuk az egyes NP-ket. Minden, anaforikusnak feltételezett NP-hez legfeljebb egyetlen, korábbi NP antecedenst – a szövegben hozzá legközelebb esőt – rendelünk, a megközelítésünkben így a visszautalások láncokba szerveződhetnek (szemben a mindig a szövegben legelső antecedensre visszautaló annotálási megközelítésekkel.) Így a névmások, zéró névmások antecedensei lehetnek korábbi névmások, zérónévmások is.

A koreferencia-feloldás a teljes input dokumentum nyelvi elemzésével kezdődik. A bekezdésekre tagolt szöveg mondatainak mindegyikéhez a MetaMorpho elemzővel előállított szintakszisfák egyszerűsített változatát rendeljük, melyek a gyökércsomópont alatt csak a tagmondatoknak (CP), maximális igei frázisoknak (VP) és a főnévi csoportoknak (NP) megfelelő csomópontokat tartalmazzák. A szintaktikai elemző

gyakran (főként a hosszabb, összetett mondatok esetében) nem képes teljes, a mondat minden szavát lefedő elemzési fát előállítani, ilyenkor a rendelkezésre álló részelemzéseket használjuk fel (VP-k, NP-k, illetve főnévi eredetű határozói csoportok (ADVP)). Az azonosított főnévi csoportokban 25 jegy reprezentálja a MetaMorpho segítségével meghatározott pozicionális, lexikai, morfológia, szintaktikai és szemantikai tulajdonságokat.

A nyelvi előfeldolgozást követi a koreferencia-viszonyok feldolgozása, mely az antecedens-jelölteket szűrő megszorítások és a fennmaradó jelöltek közül választó preferenciák módszerén alapul ([7]). A módszer minden lépése a feloldandó anaforikus elem típusától (tulajdonnév, határozott névelős köznév vagy (zéró)névmás) függő szabályokat tartalmaz. Az általános algoritmus a következő 4 lépésben működik:

1. *Előszűrés*: az anaforikusnak feltételezett, tovább feldolgozandó NP-ket azonosítjuk. A jelenleg nem kezelt visszautaló elemek mellett próbáljuk felismerni és kizárni azokat a formailag visszautaló, azonban a szövegből kiutaló, tehát szövegbeli előzménnyel nem rendelkező NP-ket is, melyek további feldolgozása zajként jelentkezne ([12]). Ebbe a lépésbe beépítettünk 5 olyan heurisztikát is, melyek célja a nyelvi elemző által nagy valószínűséggel hibásan elemzett, így a koreferencia-feloldásban is szükségképpen hibát okozó NP-k felismerése és kizárása, pl. töredék-elemzésekben 2 szónál többet nem lefedő, névszói állítmány VP alá eső zérónévmások kizárása.

2. *Az antecedens-jelöltek listájának előállítása:* ebben a lépésben az anafora típusától függően a szövegben megadott távolságtól visszakeresve kijelöljük azokat a korábbi, az anaforának megfelelő típusú NP-ket, amelyek antecedensként szóba jöhetnek. A kötéselmélettel összhangban az anaforához legközelebbi antecedens-jelölt sem eshet az anaforával egy VP alá (mivel jelenleg nem kezeljük a visszaható és kölcsönös névmásokat.)

3. *A jelöltek szűrése:* ebben a lépésben megpróbálunk kizárni minél többet az antecedens-jelöltek közül (a konkrét módszer az anafora tipusától függ, ld. később), illetve a jelöltekre is alkalmazzuk az 1. lépésben ismertetett elemzési hiba-felismerő heurisztikákat.

4. *Antecedens kiválasztása a fennmaradó jelöltek közül:* az anafora típusától függő módszer szerint. Bizonyos típusú anaforák esetében az algoritmusnak kötelező kiválasztani egy jelöltet, mások esetében nem (ld. később.)

Az alábbiakban ismertetjük az algoritmus konkrét lépéseit a különböző anafora-típusok esetében.

2.1 Tulajdonnevek

Előszűrés: jelenleg nincs előszűrés (minden, a szövegben előforduló tulajdonnevet feldolgozunk).

Jelöltek: a jelöltek listázásának hatóköre a teljes megelőző dokumentum, az összes tulajdonnév NP-t hozzáadjuk a listához az anaforát tartalmazó VP kezdetéig.

Szűrés: jelenleg nincs szűrés.

Antecedens kiválasztása: az anafora és az antecedens-jelölt normalizálása (a kezdő determinánsok elhagyása, a fej tövesítése) után kiszámítjuk közöttük a Levenshtein-távolságot ([11]), melyet a hosszabbik string hosszával normalizálunk. Az algoritmusnak nem kötelező az antecedens-jelöltek közül választania, így az anaforához legjobban hasonlító (a legkisebb Levnshtein-távolságot mutató) antecedens-jelöltet csak azok közül a jelöltek közül választjuk ki, amelyek egy paraméterben meghatározott küszöbérték alatti hasonlóságot mutatnak (amennyiben a lista nem üres.)

2.2 Határozott névelős köznevek

Előszűrés: a „szemantikus NP"-knek ([12]) nevezett, közös világismeretből azonosítható unikus obejktumokra referáló, tehát a szövegben antecedenssel nem rendelkező határozott névelős közneveket próbáljuk meg felismerni és kizárni a feloldás alól (pl. *„az amerikai elnök"*). Ehhez jelenleg egy külön, előre összeállított listát használunk.

Jelöltek: tulajdonnevek és köznevek (determináns típusától függetlenül) az anafora teljes megelőző bekezdésében, az anafora VP-jéig.

Szűrés: jelenleg nincs.

Antecedens kiválasztása: a jelöltek közül meghatározzuk az anaforához legközelebb eső, vele azonos fejű NP-t (ismétlés), vagy szinonimát, vagy hipo-/hipernimát.

A szinonimitás vizsgálatához mind az anafora, mind az antecedens-jelölt lehetséges jelentéseit kikeressük a Magyar WordNetben, és ha van olyan synsetet, ami mindkettőt tartalmazza, szinonimáknak tekintjük őket. Mivel nincs jelentés-egyértelműsítés, a módszer nyilvánvalóan nem lesz minden esetben helyes.

Az anafora és az antecedens-jelölt közötti hipernima-viszony meghatározására a Leacock-Chodorow szemantikai hasonlósági formulát alkalmazzuk ([5]), amely a visszautaló és a jelölt összes WordNet-beli megfelelőit összekötő, hipernima-reláció szerinti útvonalak közül a legrövidebb alapján számít ki egy, az útvonal hosszától függő pontértéket. Hipernima/hiponima jelölteket csak az anaforát megelőző mondatban fogadunk el akkor, ha a Leacock-Chodorow hasonlósági képlet meghaladt egy paraméter küszöbértéket, és csak akkor, ha nem találtunk azonos fejű, vagy szinonim antecedenst a bekezdésben. Jelentés-egyértelműsítés hiányában a lexikális többértelműségek nyilvánvalóan itt is fognak hibákat okozni.

2.3 Névmások

Előszűrés: csak a zérónévmásokkal, személyes névmásokkal, valamint az „az" mutatónévmással foglalkozunk, feltéve, hogy utóbbi a VP-jében alanyi szerepben áll, és nem egy alárendelt tagmondatra utal. Nem foglalkozunk az első, illetve második személyű, ún. deixikus névmásokkal és zérónévmásokkal ([9]).

Jelöltek: az anafora mondata előtti második mondattól kezdve (amennyiben az létezik a bekezdésben) választjuk ki az összes NP-t, az anaforát tartalmazó tagmondat határáig.

Szűrés: az anafora és az antecedens-jelölt számának, személyének és két szemantikai jegyének (+/-élő, +/-ember) egyezését vizsgáljuk. Utóbbiak értéke lehet alulspecifikált (zérónévmások, illetve az elemző szótárában többértelmű főnevek esetében), ezek minden lehetséges értékkel kompatibilisek.

Kizárjuk továbbá azokat a lehetséges antecedenseket is, amelyekre már koreferenciát állapítottunk meg a vizsgált anaforával egy tagmondatban szereplő valamelyik másik névmási vagy zérónévmási anaforára nézve (ld. kötéselmélet).

Antecedens kiválasztása: egy mondatban mindig először az alanyi szerepű névmási anaforát oldjuk fel, és utána a többit (ha van). Így az előbb említett, már kötött antecedensek kizárásának segítségével kizárásos alapon is sok nem alanyi szerepű névmási anafora feloldható.

A (tag)mondatában alanyi szerepű névmási vagy zérónévmási visszautaló antecedensének meghatározásában Pléh Csaba és munkatársainak a magyar mondatmegértés pszicholingvisztikája körében végzett kutatási eredményeire támaszkodtunk ([10]). A heurisztika a szerkezeti párhuzamosság feltételezéséből indul ki, mely szerint az alanyi helyzetű anafora az előzménymondat alanyára utal vissza. Ezt felülbírálhatja az alanyi szerepben álló „az" mutatónévmás, ami alanyváltást jelöl:

(2a) **Hugó**$_j$ felhívta **Amáliát**$_k$. **(Ő**$_j$**)** elmondta **neki**$_k$ a történetet.
(1b) **Hugó**$_j$ felhívta **Amáliát**$_k$. **Az**$_k$ elmondta **neki**$_j$ a történetet.

Alanyváltást egyéb jelenségek is előidézhetnek (pl. a második mondat predikátuma szemantikailag inkább a nem alanyi vonzatot preferálja stb.), ezekkel jelenleg nem foglalkozunk. Amennyiben a megelőző tagmondatban a szűrés után nem maradt rendelkezésre álló alany, az algoritmus a jelöltek listájában továbblép az azt megelőző tagmondat alanyára (amennyiben nem megy túl a bekezdés határán.)

„Az" formájú alany esetén, amennyiben az előzménymondatban több, nem alanyi szerepű antecedens-jelölt NP is található, az alábbi szabályok alapján választunk:

1. Hozzáférhetőség: az oblikuszi hierarchiában (tárgyi vonzat < egyéb vonzat < szabad határozó) magasabb helyen álló NP-t választjuk.

2. Távolság: a mondatában az anaforához közelebb eső NP-t preferáljuk (az oblikuszi hierarchiában azonos szinten álló NP-k közül).

Nem alanyi pozícióban álló névmások, zérónévmások esetén több, az alannyal nem koreferens antecedens-jelölt közül szintén a fenti két szabály alkalmazásával választunk.

A koreferencia-feloldást először minden mondatban a tulajdonnevekre, határozott névelős köznevekre végezzük el, ez után következik a mondat névmási, zérónévmási anaforáinak feldolgozása. Reményeink szerint ezzel további segítséget adunk a névmási anaforák feloldásának a szűrési feltételekben leírt szabály alkalmazásával.

3 Kiértékelés

A koreferencia-feloldó modul pontosságának kiértékelése jelenleg is folyamatban van, így csak részleges eredményekről tudunk az alábbiakban beszámolni. A kiértékeléshez létrehoztunk egy kézzel annotált kiértékelő-korpuszt, amely 5 darab, általános iskolai történelemkönyvekből kiemelt szöveget tartalmaz (2. Táblázat). A szövegekben a MetaMorpho segítségével azonosítottuk a maximális NP-ket, majd annotáltuk közöttük a koreferencia-viszonyokat. Az automatikus annotációhoz hasonlóan a koreferencia-láncokban mindig az anaforához legközelebbi antecedenseket jelöltük be. A munkát egyetlen annotátor végezte.

Mivel a nyelvtani elemző nem minden NP-t ismert fel, illetve egy részüket hibásan, csak a jól felismert NP-ket tudtuk annotálni (és azokat is csak akkor, ha az antecedensük is helyesen volt bejelölve), így fedés(recall) kiértékelésére a korpusz jelenleg nem alkalmas.

2. Táblázat: a kiértékelő korpusz jellemzői

Szövegek száma	5
Bekezdések száma	79
Mondatok száma	652
NP-k száma	3115
Antecedenssel annotált NP-k száma	338

A koreferencia-feloldó algoritmust ezután lefuttattuk a korpusz szövegein, majd összevetettük az automatikus annotáció eredményeit a kézivel. A rendszer 145 NP-re adott eredményt, ezek közül 101 egyezett meg az annotációval (69 %-os átlagos pontosság.) Ezután megvizsgáltuk a különböző visszautalási típusok felismerésének pontosságát külön-külön is (3. Táblázat.) A kiértékeléskor nem állt rendelkezésre a tulajdonnevek koreferenciájának felismerése, így az erre vonatkozó adatok hiányoznak a táblázatból.

Szembeötlő a különbség a hipernimán alapuló és a többi módszer között. A hipernima-kereső módszer nélkül az algoritmus átlagos pontossága 80%-os lenne.

3. Táblázat: a különböző koreferencia-feloldó módszerek pontossága

Visszautalási típus	Automatikusan annotált NP	Helyesen annotált NP	Pontosság (%)
Ismétlés	21	20	97%
Szinonima	8	6	75%
Névmás	110	75	68%
Hipernima	6	0	0%

Kíváncsiak voltunk arra, hogy mennyiben befolyásolja a nyelvi elemző a névmási anafora-feloldás teljesítményét, ezért egy másik szövegen részletesen megvizsgáltuk a névmási anafora-feloldás hibáit. A szöveg 109 mondatot tartalmazott, a rendszer ezekben 521 db NP-t jelölt be, melyek közül 34 db névmási NP-hez azonosított antecedenseket. Az automatikusan azonosított antecedensek mindegyikét kézzel helyes vagy hibás eredményként értékeltük, és a hibás automatikus annotáció alábbi három esetét különítettük el

a) a hiba a helytelen nyelvi elemzés következménye (rosszul elemzett anafora és/vagy rosszul/nem elemzett antecedens) (*jelölés: KO_parser*)
b) az anaforának nem volt a szövegben antecedense (a program nem ismerte fel, hogy az elem nem utal vissza) (*jelölés: KO_noant*)
c) az anaforának volt a szövegben antecedense, de a program helytelenül azonosította (*jelölés: KO_cr*)

A helyes elemzések és a különböző hibatípusok arányait a 4. Táblázatban foglaltuk össze.

4. Táblázat: az anafora-feloldás hibatípusainak aránya

Eredmény típusa	előfordulása	%
OK	24	67%
KO_parser	7	20%
KO_noant	0	0%
KO_cr	3	8%

A táblázatból látható, hogy az automatikus annotáció hibájának nagy százaléka a nyelvi elemző hibájának következménye. Hibák nélküli szintaktikai elemzésre támaszkodva a névmási anafora-feloldás pontossága a jelenlegi algoritmussal akár 83%-os pontosságot is elérhetne.

A névmási anafora-feloldás az általunk vizsgált mintában nem azonosított antecedenseket olyan NP-khez, melyeknek nincs szövegbeli előzménye. Ennek oka, hogy a névmási visszautalások általunk kezelt fajtái mindig létező, szövegbeli antecedensre utalnak, hibát csak a helytelen nyelvi elemzésből származó, invalid NP-k okozhatnak, azonban az ezeket kiszűrő heurisztikák jól működtek.

4 További munka

Elsőként szeretnénk folytatni a jelenlegi rendszer teljesítményének kiértékelését. Ehhez a tulajdonnevek feloldásának kiértékelése, másrészt a fedés vizsgálati módszerének kidolgozása szükséges. Ezután természetesen a hibák részletes kategorizálása és elemzése következik, különös tekintettel a hiponimák azonosításának módszerére, amely lényegesen rosszabbul teljesített a többi módszerhez képest.

Szeretnénk egy baseline megoldásnak megfelelő algoritmust implementálni, amelyhez képest meghatározható a rendszerünk teljesítménye. Ehhez a Centering elméletre alapuló, a szakirodalomban jól ismert BFP-algoritmust ([1]) szeretnénk kipróbálni, melyet vizsgáltak már magyar szövegekkel is ([6]).

Szeretnénk létrehozni egy olyan, koreferenciával annotált kiértékelő korpuszt, ami mások számára is hozzáférhető, így a rendszerünk teljesítménye más hasonló rendszerekkel is összevethető lesz. Ehhez legalkalmasabbnak a frázis-annotációkat tartalmazó Szeged Treebank 2.0-s változata tűnik ([2]). A Szeged Treebank használatával nyelvi elemzőtől független, nagy pontosságú szintaktikai elemzésekre lehetne koreferencia-feloldó algoritmusokat építeni (ugyanakkor bizonyos jegyek, melyek a MetaMorpho kimenetében azonosíthatók, nem lesznek elérhetők.)

Az anafora-feloldás fedésének növelésére további főnévi anaforikus jelenségek kezelésére lesz szükség: visszaható és kölcsönös névmások, vonatkozó névmások, birtokos névmások valamint a komplex NP-k részegységeinek, a birtokos szerkezeteknek stb. elemzésére és koreferencia-kapcsolataik feltárására.

A pontosság növelése érdekében a tulajdonnevek felismerésére további karakterhasonlóságon alapuló módszereket és normalizációs eljárásokat mutat be [11]. A karakteres és a szemantikai hasonlóság-képletek számára a küszöbértékeket empirikus úton, korpuszpéldák segítségével lenne célszerű optimalizálni.

A tulajdonnevek és köznevek feloldásánál további, felhasználható információ lehet az anafora és az antecedenst egymástól való távolsága. A MetaMorpho lexikonjában tárolt szemantikai jegyek (pl. tulajdonnév-osztályok, szemantikai kategóriák stb.) egyezésének vizsgálata további szűrési feltételeket adhat.

Határozott névelős közneveknél felmerül a kérdés, hogy az azonos fejű, számban egyező, de eltérő módosítókat tartalmazó anafora–antecedens-jelölt párokat hogyan kezeljünk (pl. *a katonák–az út szélén elrejtőzött katonák.*)

A névmási anaforák kezelésénél további heurisztikák alapja lehet a Centering Elmélet, a diskurzustopik változásának figyelése ([1]), illetve [10] által leírt egyéb jelenségek modellezése (pl. a predikátum által preferált vonzatok korpuszstatisztikai vizsgálata.)

További lehetőség a zajt okozó, feloldást nem igénylő NP-k azonosítása további módszerekkel. Az egyik a szükségszerű/valószínű rész viszony, pl. „*Tegnap szerelőhöz vittem a* ***biciklim****, mert eltört a* ***pedál****.*". Ha rendelkezésre állna megfelelő adatbázis, az esetkeretből levezethető entitásokat is fel lehetne ismerni, pl „***A konferencia*** *véget ért.* ***A résztvevők*** *elégedetten távoztak.*"

Bibliográfia

1. Brennan, Susan E., Marilyn W. Friedman, Carl J. Pollard. A centering approach to pronouns. In Proceedings of the 25th Meeting of the Association for Computational Linguistics (1987), pp. 155-162.
2. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005) kiadványa, Szeged, december 8-9., pp. 409-412 (2005)
3. Hatvani Cs., Kocsor A., Miháltz M., Szarvas Gy., Szécsi K.: Főnevek a Magyar WordNetben. IV: Magyar Számítógépes Nyelvészet Konferencia, Szeged (2006), pp. 109–116.
4. Kenesei István: Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális Magyar Nyelvtan, I. kötet, Mondattan. Akadémiai Kiadó, Budapest (1992)
5. Leacock, C., M. Chodorow: Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998), pp. 265–285
6. Lejtovicz Katalin, Kardkovács Zsolt: Anaforafeloldás magyar nyelvű szövegekben. IV. Magyar Számítógépes Nyelvészet Konferencia, Szeged (2006)
7. Mitkov, Ruslan: Anaphora Resolution: The State of The Art. Working Paper, University of Wolverhampton, 1999.
8. Ng, Vincent: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proceeding of the 43rd Annual Meeting of the Associaction for Computational Linguistics (1995)
9. Pléh Csaba, Radics Katalin: „Hiányos mondat", pronominalizáció és a szöveg. In Általános Nyelvészeti Tanulmányok, XI, 261-277 (1976).
10. Pléh Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998)
11. Uryupina, Olga: Evaluating Name-Matching for Coreference Resolution. In Proceedings of the 4th International Conference on Language Resources and Evaluation (2004)
12. Varasdi Károly: Koreferenciák feloldása. Projektdokumentum (2005)