Hunpars: mondattani elemző alkalmazás

Babarczy Anna[1], Gábor Bálint[1], Hamp Gábor[2], Kárpáti András[3], Rung András[4], Szakadát István[2]

[1] Kognitív Tudományi Tanszék, BME, 1111 Budapest, Stoczek u. 2.
{babarczy, bgabor}@cogsci.bme.hu
[2] Szociológia és Kommunikáció Tanszék, BME, 1111 Budapest, Stoczek u. 2.
hampg@eik.bme.hu, syi@axelero.hu
[3] Klasszika-filológia Tanszék, PTE, 7624 Pécs, Ifjúság útja 6.
karpati.andras@t-online.co.hu
[4]Nyelvtudományi Intézet, MTA-ELTE, 1068 Budapest, Benczúr Gy. u. 33.
runga@artitude.hu

Kivonat: A Hunpars-projekt célja egy nyílt forráskódú elemző alkalmazás létrehozása, amely automatikusan végzi el bármilyen értelmezhető magyar mondat szintaktikai elemzését, konkrétan a mondatot alkotó szócsoportok és azok egymáshoz való viszonyának azonosítását. Az elemzőt egy többkomponensű rendszer részeként képzeljük el: a fejlesztés alatt álló modul bemenete egy előzőleg tokenizált mondat, amelyben a szavak morfológiai jegyeikkel felcímkézve szerepelnek. A szintaktikai elemzés szabályalapú: elsősorban egy szintaktikai kategóriákra épülő frázis-struktúra nyelvtan és kiegészítésként különböző lexikális tárak felhasználásával valósul meg. Az alkalmazást irodalmi, jogi, tudományos-ismeretterjesztő és sajtószövegből származó, kvázi-véletlenszerűen kiemelt mondatokon teszteltük. A tesztmondatok 72%-ára helyes elemzést kaptunk, további 11% elemzésének hibája szótári hiányosságra vezethető vissza.

1 Bevezetés

A következőkben egy magyar nyelvre alkalmazható, mondattani elemző működését mutatjuk be. Az alkalmazás fejlesztése 2003-ban kezdődött, és eredetileg egy kérdés-megválaszoló rendszer[1] egyik moduljának készült. Később az elemző fejlesztése függetlenedett az eredeti projekttől, így ma már nem csak egyszerű kérdő mondatok, hanem bármilyen magyar nyelvű mondat elemzésére is használható. Azt a cél tűztük ki tehát, hogy létrehozzunk egy olyan nyílt forráskódú alkalmazást, amely magyar nyelvű természetes mondatok frázisainak és a frázisok közötti viszonyoknak az azonosítását végzi el automatikusan, kézi beavatkozás nélkül.

A mondattani elemzés elsődleges feladata, hogy egy mondatban ne csak az összetartozó szócsoportokat azonosítsa, hanem meghatározza a viszonyokat az egyes szavak és a szavakból alkotott szerkezetek közt is. Ez a cél megvalósítható szigorúan

[1] *Szavak hálójában*, Budapest Műszaki és Gazdaságtudományi Egyetem és Axelero Rt. (jelenleg T-Online Rt.) projekt NKFP és OM támogatással.

lexikalista alapon is [5], amikor a szintaktikai szerkezetet a szavak közötti kapcsolatok határozzák meg. Erre a megközelítésre példa a magyar GeLexi-projekt, ahol a szintaktikai elemzés alapja a gazdag szóleírásokat tartalmazó lexikon, amely megadja az egyes szavak kapcsolódási lehetőségeit [1].

Egy másik lehetséges – és általunk is választott – megközelítésben az elemző a frázisstruktúra nyelvtanokhoz hasonlóan a szavakat hierarchikus szerkezetekbe, frázisokba szervezi, és ezt követően a viszonyokat már ezek között a hierarchikus szerkezetek között határozza meg (a módszer áttekintésére lásd [2]). Például az (1) mondat szerkezetét a (2)-es zárójelezett változat jeleníti meg.

1. Az előadás után meglehetősen leverten álltam a lepusztult mozi előtt.

2. [[Az előadás] után] [meglehetősen [leverten]] [álltam] [a [lepusztult [mozi]]] előtt].

Minden frázisnak (zárójelezett egységnek) van feje, amely egy olyan szó, amely meghatározza a frázis viselkedését a mondatbeli hierarchia következő szintjén. A frázisstruktúra nyelvtanok kiegészíthetők lexikális függőségi információval [8]. A mondat szerkezetének helyes elemzéséhez szükségünk van az adott nyelv módosítóinak nyelvtanára is (például hogy az *előtt* névutó jelölhet egy szabadon előforduló hely- vagy időhatározót), és a régensek szubkategorizációs követelményeire, azaz hogy milyen argumentumai lehetnek egy adott régensnek, amelyek jelenlétében az adott szerkezet jól formált lesz. Egy ilyen típusú komplex nyelvtant meghatározhatunk lexikális és általánosított konstrukciós minták halmazával (Kálmán et al. 2003) vagy frázisstruktúrákat létrehozó szabályok egymás után való rendezésével és lexikális függőségi adattárak alkalmazásával.

A Hunpars alkalmazás az utóbbi eljárást használja kisebb módosításokkal, mint azt a következőkben részletezzük. A megközelítésünkhöz hasonló, de csak szócsoportok azonosítására fókuszáló kutatások folynak a Nyelvtudományi Intézetben [9], [10], illetve ide sorolható a szintén szabályokat létrehozó, de automatikus módszereket alkalmazó HumorEsk [6] és Hócza András kutatásai [3].

2 A Hunpars felépítése

Projektünk nem törekszik arra, hogy egy adott elméletet minél hívebben adaptálva hozzon létre egy mondattani elemzőt. A elemző tervezésénél a nagy lefedettség elérése volt az elsődleges cél, azaz hogy egy viszonylag egyszerű szabályrendszer és néhány jól megválasztott algoritmus rugalmas kombinációjával minél többféle természetes nyelvi mondatot tudjunk elemezni, beleértve az esetleg egyedien formált, pontatlan vagy nem teljes mondatokat is.

A Hunpars alkalmazás tehát három pilléren nyugszik:

- Frázis-struktúra nyelvtan

- Lexikális adattárak
- Elemzési algoritmusok

2.1 A nyelvtan

A nyelvtant kifejezetten a magyar nyelvre fejlesztettük, szem előtt tartva a nyelv gazdag morfológiai rendszerét és az ezzel összefüggésbe hozható variálható szó- és konsztituenssorrendet. A Hunparsnak szüksége van az elemezni kívánt mondat szavainak morfológiai elemzésére, ehhez a Hunmorph morfológiai elemzőt [7] használjuk. A Hunmorph által adott elemzés az egyes szavak szófaji besorolása mellett megadja a szóalakok teljes morfológiai jegyhalmazát is. A *dolgoknak* szóalakhoz így például a következő elemzés tartozik: `dolog/NOUN<PLUR><CAS<DAT>>`. Az elemző nyelvtana tehát elsősorban a morfológiai elemzés kimenetére és másodsorban a szavak lineáris elhelyezkedésére épül.

A frázisok fejének megválasztásakor a lexikalista hagyományokat követjük. A mondattani elemzés során a zárójelezett frázisok alapértelmezésben öröklik a fej jegyeit, illetve egyes esetekben a frázis más alkotószavainak jegyei is öröklődhetnek. Így például *a lepusztult mozi* konstrukció nemcsak a tartalmi *mozi* fej jegyeit, hanem a névelő jegyeit is hordozza.

Ha a mondatban szereplő szavak bármelyike morfológiailag többértelmű, akkor az adott mondatnak ennek megfelelően újabb változatait hozzuk létre, és ezek mindegyikére lefut az elemzés. Azaz, ha egy mondatban három kétértelmű és egy háromértelmű szó található, akkor az adott mondatnak akár $2^3 \times 3$, azaz 24 különböző elemzése is lehet. A többértelműségek nagy része egy statiszikai egyértelműsítő modullal kiszűrhető. Bár ilyen modul jelenleg nem áll rendelkezésünkre, néhány egyszerű előszűrő szabállyal is jelentősen sikerült csökkentenünk a többértelműségek számát. Az előszűrést követően megmaradt mondatváltozatokat a nyelvtan tovább szűri: a szabályrendszert ki nem elégítő változatokat elvetjük.

2.2 Lexikális adatbázisok

A morfológiai jegyeken túl nyelvtanunk lexikális adatbázisokban található információkat is használ. Például melléknevek (*egy fiára büszke anya* , de **fiának büszke* **fiával büszke*), névutók (*a házzal szembe*, **a házra szembe*, **a házba szembe*) és a későbbiekben majd igék bővítményeire vonatkozó megszorításokat. A Hunpars használja továbbá az igék és igekötők lehetséges kombinációira vonatkozó információs tárat.

2.3 Elemzési algoritmusok

Első lépésben az összetett mondatokat az elemző tagmondatokra bontja, melyeket külön elemez a továbbiakban, és a folyamat végén ezeket a részelemzéseket egyesíti. Az elemző algoritmus sorrendezett elemzési fázisokból áll. Mindegyik fázis szabályil-

lesztések egy sorozatát és/vagy egyéb algoritmikus lépéseket foglal magában, melyek egy adott frázistipus zárójelezését végzik. A szabályillesztések során a mondatokban jobbról-balra vagy balról-jobbra keresünk olyan szót, mely morfológiai jegyei alapján lehet a keresett frázis feje a mindenkori szabálynak megfelelően. A fej azonosítása után a nyelvtani szabály által meghatározott (kötelező vagy opcionális) egyéb elemeket a fejhez csatoljuk. Ha egy frázis elemeit megtaláltuk, az elemző tovább halad a tagmondatban.

Miután az elemző egy fázison belül zárójelezte a lehetséges frázisokat, továbblép a következő fázisra. Az elemzés későbbi lépéseinél az előzőleg lezárt frázisokat egy egységnek kezeljük.

A keresés irányát a nyelvtan határozza meg, ez a fázisok szabálycsoportjaiban különbözhet. A keresés irányának meghatározó szerepe van. Az irányváltoztatásnak bizonyos rekurzív tulajdonságokat mutató szerkezetek elemzésénél (pl. birtokos szerkezet) van kiemelt szerepe, amelyekben így elkerülhető volt, hogy vermet igénylő rekurzív szabályillesztéseket használjunk.

Az elemző másik fontos eszköze beágyazott szerkezetek kezelésére a lezáratlan frázis funkció. A lezáratlan frázisok elemei az elemzés későbbi lépései során (speciális szabályok segítségével) még bővíthetők.

A nyelvtan szükség esetén – mint láttuk – a lexikai adatbázisokhoz fordul.

A Hunpars a következő elemzési fázisokat tartalmazza.

- Előfeldolgozás: morfológiai egyértelműsítés
- Előfeldolgozás: tagmondatokra bontás
- Az igei frázis és a tagmondat régensének felismerése
- Határozószói frázisok elemzése
- Számnévi frázisok elemzése
- Melléknévi frázisok elemzése
- Főnévi frázisok elemzése
- Névutói frázisok elemzése

Bizonyos fázisokat követően egy mellérendelői szerkezeteket azonosító fázis is lefut az adott szinten lévő mellérendelések azonosítására. Ez a fázis akkor ismer fel egy mellérendelést, amikor a kötőszó előtt és után álló mellérendelő viszonyban frázisok már elemzésre kerültek. Például mellérendelés a *piros labda és kék szalag* kifejezésben csak a főnévi frázisok azonosítása után jöhet létre, míg a *piros és kék labda* esetében már a melléknévi szakasz után azonosítható.

A fázisok lefutása során az azonosított fejekkel egy frázisba kerülnek módosítóik és bővítményeik egy hierarchikusan szervezett frázisstruktúrát alkotva.

A fázisok jellegének szemléltetéséhez az alábbiakban bemutatjuk a szabályok leírására alkalmas formalizmus egy rövid kivonatát és példaként a melléknévi fázis leírását:

$A ::= B_1\ B_2\ \ldots\ B_n$

B_1 B_2 … B_n egymás után alkosson A frázist.

`A ::= { B1 B2 … Bn }`

B_1 B_2 … B_n bárhogy elhelyezkedve alkosson A frázist.

`Phase(<name>):`

A következő szabályok a <name> nevű elemzési fázishoz tartoznak.

`group:`

Egy szabálycsoport szabályai következnek, a következő `group:`-ig, vagy a következő fázis kezdetéig. Egy szabálycsoporton belüli szabályokat egy keresés során illeszthetünk, melynek iránya alapértelmezésben jobbról balra. Ha egy szabálycsoporton belül több szabálynak is ugyanaz a feje, és a mondat egy összetevője fejként mind a két szabályillesztést lehetővé tenné, akkor azt az illesztést kell végrehajtani, amelyik hosszabb frázist eredményez. A group kulcsszó után zárójelben a szabályok fejének kereséséről adhatunk meg paramétereket:

`group(left2right):` a keresés balról jobbra haladjon

`group(unfinal):` a fejet lezáratlan frázisokban keresse

Szögletes zárójelben (`[]`) adhatunk meg a szabály használatára vonatkozó feltételeket, ezek lehetnek tokenszintűek vagy szabályszintűek. Tokenszintű esetén a token után rögtön szögletes zárójelben szerepel a rá vonatkozó feltétel, szabályszintű esetén a szabály után található egy több tokenre vonatkozó feltétel. A szabályszintű feltételben úgy hivatkozhatunk a tokenekre, hogy azokat egy / jellel és egy számmal megindexeljük.

Pl.:

tokenszintű: `AdjP ::= Adv[canModify(Adj)] Adj`

szabályszintű: `AdjP ::= Noun/1 Adj/2 [arg(1,2)]`

Alapértelmezésben a létrejövő frázis feje a jobb oldalon a legszélen lévő elem, egyéb esetben a fejet `/HEAD`-del jelölhetjük meg.

A tokenek számának meghatározására a reguláris kifejezésekben megszokott jelöléseket használjuk.

A szabály végén kettőspont után utalhatunk a szabályhasználat miként jére vagy következményeire . Ha nem szerepel semmi, akkor a szabályt nem illeszthetjük újra a létrejövő frázisra, ha `:repeat` szerepel, akkor rekurzívan többször is illeszthetjük, ha `:break`, akkor ezt és más a fázishoz vagy szabálycsoporthoz tartozó szabályt sem illeszthetünk többet, ha `:error` akkor az elemzésünk hibajelzéssel leáll. Ha `:rel(<reláció>)` szerepel a szabály után, akkor a megadott relációt a szabály alkalmazhatósága esetén igazra kell állítanunk. `:call(<eljárás>)` esetén a megadott a szabály alkalmazása után végre kell hajtanunk a megadott eljárást, `:split` esetén a szabály illesztésekor két elemzési variánst kell létrehozni: egyikben illesztjük a szabályt, másikban nem.

Ha a ::= jel bal oldalán nincs semmi, akkor illesztés esetén nem jön létre új frázis, de a szabályhasználat megadott következményeit végre kell hajtanunk.

A melléknévi fázis rövid leírása:

```
Phase('adj'):

group:

AdjP ::= Adj
```

Az -ú/-ű végű melléknév előtt nem -ú/-ű végű melléknév állhat:

```
AdjP ::= Adj[adjType!=_U] Adj[adjType==_U] :repeat
```

Nem -ú/-ű végű előtt bármilyen melléknév állhat:

```
AdjP ::= Adj Adj[adjType!=_U] :repeat
```

Melléknevet módosító határozószó lehet melléknév előtt, de ez nem ismételhető szabály:

```
AdjP ::= Adv[canModify(Adj)] Adj
```

Ha a melléknév igenév, akkor bármilyen határozószó állhat előtte:

```
AdjP ::= Adv Adj[adjType==PART]
```

Főnév vagy melléknév vonzatkerettár ellenőrzéssel szerepelhet, lezáratlan frázist hoz létre, és ilyenkor más szabályt már nem alkalmazhatunk:

```
AdjP/1 ::= Noun/2 Adj/3 [argCheck(2,3)] :unfinal(1)
:break

AdjP/1 ::= Num/2 Adj/3 [argCheck(2,3)] :unfinal(1)
:break
```

3. Teszteredmények és továbblépési lehetőségek

Mivel jelenleg nem áll rendelkezésre különféle szövegműfajokat jól reprezentáló, kézzel elemzett magyar nyelvi korpusz, az elemző tesztelése nem automatizálható, s ezért ezt kézzel kellett elvégeznünk. Erre a célra egymástól független, magyar mondatokból álló korpuszt állítottunk össze kvázi véletlenszerű módon különböző forrásokból. A források között kortárs irodalmi művek, sajtó-, tudományos-ismeretterjesztő és jogi szövegek szerepeltek. Mivel elemzőnk nyelvtana jelenleg nem terjed ki vonatkozói mellékmondatok elemzésére, így az azokat tartalmazó mondatokat eltávolítottuk a korpuszból. Az elemző a teljes korpusz elemzése során a bemeneti mondatokra – illetve ha egy mondathoz több változat is szerepelt, azok mindegyikére – az elemzés eredményét tartalmazó annotált text fájlt hoz létre. Az ellenőrzés megkönnyítésére az text fájlok mellett ezekből létrehozott, grafikusan megjelenített elemzési fákat állítottunk elő; az 1-es ábrán egy ilyen elemzési fa látható. Az ellenőrzést egy nyelvészeti

szaktudással rendelkező szakértő végezte, akinek nem volt alapos ismeretei a nyelvtan részleteiről és a Hunpars algoritmusairól.

Az elemző működésének kvantitatív értékelésékor *sikeresnek* minősítettünk egy elemzést, amikor az elemző az elemzési folyamat során legalább egy mondatváltozatot nem utasított el. Azonban ezen elemzéseknek csak az a része minősült *helyesnek*, amelyet a kézi ellenőrzés során is hibátlannak találtunk.

A teszt kvantitatív kiértékelését az 1-es táblázat mutatja be:

1. Táblázat: Tesztelési eredmények

Bemeneti mondatok száma	309
Sikeres elemzések száma	600
Azon mondatok aránya, amelyeknek legalább egy sikeres elemzése volt	97%
Azon mondatok aránya, amelyeknek legalább egy helyes elemzése volt	72%

Az eredmények hibáinak kiértékelése során a helytelen elemzéseket a következő csoportokba soroltuk:

- Helytelen az elemzések lexikális adatbázisok hiányosságai vagy hibái miatt. Ide tartoznak a morfológiai elemző által fel nem ismert vagy rosszul besorolt szavak vagy olyan bővítmények, amelyek a lexikai adatbázisokban nem szerepeltek, illetve idiómák és tulajdonnevek fel nem ismeréséből fakadó hibák (a bemeneti mondatok 11%-a).
- Helytelen elemzések a nyelvtan hibájából kifolyólag (a bemeneti mondatok 17%-a).
- Helytelen elemzés implementációs hiba miatt (a bemeneti mondatok kevesebb mint 1%-a).

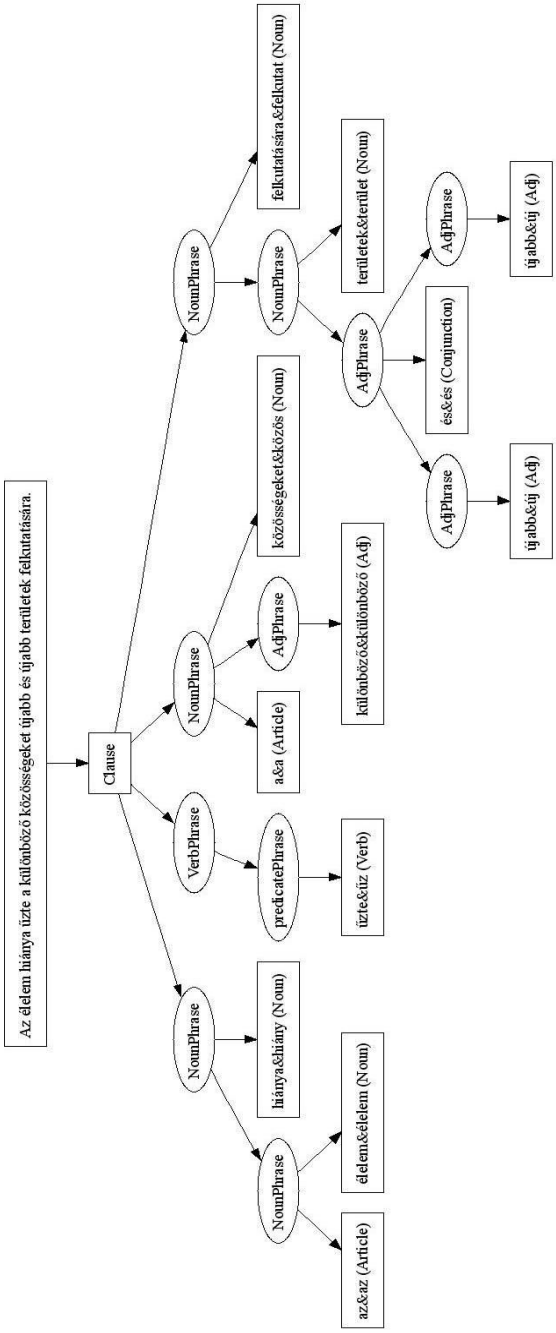
Az elemző hibáinak elemzése azt mutatja, hogy teljesítménye jelentősen javítható lenne a lexikális adatbázisok, különösen a névutók lehetséges bővítménytárának pontosításával. További fejlődést a nyelvtan bővítésétől várunk. A hibák vizsgálata megmutatta, hogy a beágyazott igeneves szerkezetek okozzák a problémák nagy részét. Az alábbi példákban zárójelezéssel emeljük ki a nem helyesen elemzett szerkezeteket:

3. A férfi a szóbeszéd szerint [egy [a felesége telefonjában talált] SMS] miatt kezdett gyanakodni házastársára.
4. A legnagyobb hazai gyorséttermi lánc múlt vasárnaptól [az egyik fizetős wifiszolgáltatóval együttműködve] drótnélküli internettel csalogatja a fizetőképes keresletet.
5. A vipassana meditáció gyakorlása során a meditáló [[[[a testében megjelenő] pszicho-fizikai jelenségek] természetének] helyes megértésére] törekszik.

További problémát okoznak azok a többértelmű mondatok, amelyek esetében az anyanyelvi beszélő számára egyértelmű, hogy a lehetséges helyes elemzések közül melyiket kell kiválasztani. A Hunpars jelenleg nem tud szemantikai információt felhasználni. Erre mutat példát a 6-os mondat, amelyben a Hunpars helytelenül azonosította a megjelölt frázist:

6. [Az országos televízió főműsoridőben] legalább húsz perc, országos rádió legalább tizenöt perc önálló hírműsort köteles egybefüggően szolgáltatni.

Hosszú távú tervünk, hogy az ilyen ilyen jellegű hibákat egy nagyméretű annotált korpuszon tanított, statisztikai módszereket használó komponens segítségével kerüljük el.



1. ábra *Az élelem hiánya űzte a különböző közösségeket újabb és újabb területek felkutatására.* mondat elemzésének grafikus megjelenítése

Bibliográfia

1. Alberti G., Kleiber J., Viszket A.: Főnévi GeLExi projekt: GEneratív LEXIkonon alapuló mondatelemzés. In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) In: MSZNY (2003) 79–846.
2. Appelt, D.E., Israel D.: ANLP-97 Tutorial: Building information extraction systems. (1997). Available as http://www.ai.sri.com/appelt/ie-tutorial.
3. Hócza A.: Teljes mondatszintaxis tanulása és felismerése. In: In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004)MSZNY (2004) 127–135.
4. Kálmán L., Balázs L., Erdélyi Szabó M.: Tudásalapú természetesnyelv-feldolgozás. In: In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) MSZNY (2003) 109–114.
5. Karlsson, F., Voutilainen, A., Heikkila, J., Anttila, A. (1995): *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
6. Kis B., Naszódi M., Prószéky G.: Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. In: In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) In: MSZNY (2003) 145–152.
7. Németh L., Halácsy P., Kornai A., Trón V.: Nyílt forráskódú morfológiai elemző. In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004) 163–171.
8. Sag, I., Wasow, T.: *Syntactic Theory: A Formal Introduction*. Stanford : CSLI Publications, (1999).
9. Váradi T., Gábor K.: A magyar Intex fejlesztésről. In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. december 9–10. SZTE, Szeged. (2004) 3–10.
10. Váradi T.: Főnévi csoportok annotálása a CLaRK rendszerben. In: Magyar számítógépes nyelvészeti konferencia MSZNY2003. Szeged, 2003. december 10–11. Konferenciakötet. SZTE, Szeged (2003) 65–70.