Igék szemantikai klaszterezése bővítménykereteik alapján

Gábor Kata[1], Héja Enikő[1]
[1] MTA Nyelvtudományi Intézet, Korpusznyelvészeti osztály, Postafiók 701/518, H-1399 Budapest, Magyarország

{gkata,eheja}@nytud.hu

Kivonat: A bemutatott kísérlet célja a magyar igék bővítménykereteik alapján történő szemantikai csoportosítása nem felügyelt tanulási módszerrel. A bővítménykereteket a Szeged Treebankből nyertük ki. Az igéket hierarchikus klaszterezési eljárással csoportosítottuk. A cikkben bemutatjuk az eljárást, az eredményeket, valamint kitérünk a szemantikai osztályok kiértékelésének nehézségeire, és javaslatot teszünk egy új értékelési módszerre.

1 Bevezetés

Az utóbbi évtizedben a nyelvtechnológiai kutatás fontos célkitűzésévé vált a nagy lefedettségű, robusztus, több nyelvi szintet lefedő lexikonok létrehozása. Ennek egyik oka, hogy a többszintű struktúra megkönnyíti a karbantartást és lehetővé teszi az automatikus bővítést, mivel a strukturált szerkezetnek köszönhetően a műveletek egyedi tételek helyett szóosztályokra alkalmazhatók, vagyis lehetővé teszik a nyelvészeti általánosításokat. Másfelől pedig a lexikális információ automatikus kinyerése kevésbé idő- és munkaigényes, mint az adatbázisok kézi felépítése. A korai automatikus lexikonépítési kísérletekben nem számítógépes célokra készült, kétnyelvű vagy értelmező szótárak elektronikus változatát használták nyersanyagként. Ez a megközelítés rendelkezik azzal az előnnyel, hogy a nyersanyagából a zajt már emberi munkaerővel kiszűrték, ám a módszerben rejlő lehetőség épp a nyersanyag korlátolt mennyisége miatt viszonylag hamar kimerült, a későbbi automatikus frissítésre pedig nem ad lehetőséget. A szótár használatánál robusztusabb megközelítést jelent az igei vonzatkeret-információ automatikus kinyerése nagyméretű korpuszokból. Napjainkban már a legtöbb európai nyelvre, így a magyarra is rendelkezésre állnak olyan erőforrások (morfológiailag elemzett és egyértelműsített korpusz: [15], szintaktikailag annotált treebank: [4]), melyek lehetővé teszik gépi tanulási módszerek alkalmazását a vonzatkeret-információ és a szemantikai tulajdonságok kinyerése céljából. A lexikai tuljdonságok gépi tanulásával foglalkozó kutatások többsége az igei szubkategorizációs keret korpuszból való kinyerését tűzte ki célul ([12], [3], magyarra: [Sass, 2007]). Ugyanakkor a lexikai szemantika-szintaxis interfész elméleti kutatásával párhuzamosan történtek

kisérletek az igék automatikus szemantikai csoportosítására is kvantifikálható szintaktikai jellemzőik alapján. ([5], [13], [14], [11]). Ezen kutatások közös elméleti kiindulópontját a szintaxis-szemantika interfésszel foglalkozó elméletek közös feltevése képezi, mely szerint az igei argumentumok vonzatokként való megvalósulásáról az ige jelentéséből kiindulva adhatunk számot. Ez a széles körben felhasznált elméleti előfeltevés (*Semantic Base Hypothesis,* [8], [10]) azon a megfigyelésen alapul, hogy a szemantikailag hasonló igék hasonló szintaktikai környezetben fordulnak elő, azaz hasonló vonzatkeret-mintázatokat mutatnak. Az igék automatikus szemantikai csoportosítását célzó munkák vagy a [10]-ben meghatározott igeosztályokat kívánják igazolni korpusz-adatok segítségével, vagy olyan algoritmusok fejlesztésén dolgoznak, amelyek lehetővé teszik új igék szemantikai kategorizációját. A fentiekkel szemben jelen kutatás célja azoknak a lexikai-szemantikai tulajdonságokkal meghatározható igeosztályoknak az azonosítása, amelyek relevánsak a magyar igei argumentumrealizáció leírása szempontjából. Mivel nem tudjuk előre, milyen szemantikai osztályokba szeretnénk besorolni az igéket, nem felügyelt tanulási módszerhez kell folyamodnunk. Így [13] nem felügyelt módszerét követtük. A Szeged Treebank ([4]) 150 leggyakoribb igéjét soroltuk csoportokba hierarchikus agglomeratív klaszterezési eljárással, szintaktikai bővítményekereteik alapján. A kísérlet kettős célt szolgált: egyrészt magát a tanulási módszert akartuk tesztelni, vagyis arra kerestük a választ, hogy bővítménykeret-információ alapján kinyerhetők-e szemantikailag koherens osztályok a korpuszból. Amennyiben ige, úgy a kísérlet megerősíti a *Semantic Base Hypothesis*-t, hiszen alátámasztja, hogy a szemantikailag hasonló igék szintaktikailag is hasonló viselkedést mutatnak. Másrészt azt akartuk megtudni, hogy melyek azok a szemantikai jelentéskomponensek, melyek köré az alapvető igeosztályok szerveződnek. Előfeltevésünk szerint a leggyakoribb igékből előűlló csoportok tükrözni fogják a legalapvetőbb igei jelentéskomponenseket.

A következő fejezetekben bemutatjuk a felhasznált jegykészletet [2] és a klaszterezési eljárást [3], majd ismertetjük az eredményeket [4]. A kiértékelés nehézségeire az [5] részben térünk ki. Végül felvázoljuk a kutatás további lehetséges irányait [6], majd összegezük az elmondottakat [7].

2 A jegykészlet

Mivel a jelenleg rendelkezésre álló magyar szintaktikai elemzők (Babarczy et al. 2005, Gábor és Héja 2005) nem végeznek teljes elemzést, nem terjednek ki az igei vonzatkeret automatikus felismerésére. Ezért úgy döntöttünk, hogy az első kísérlethez eleve szintaktikailag elemzett korpuszt, a Szeged Treebank-et használjuk. A korpuszt alkotó szövegek különböző forrásból származnak: üzleti hírek, napi sajtó, szépirodalom, jogi szövegek és iskolások fogalmazásai alkotják. A kísérlethez valamennyi részkorpuszt felhasználtuk. A treebank mérete 1.2 millió szó. Az igei bővítménykeret annotálációjában nem szerepel a vonzat és a szabad határoz kategóriája, az ige és bővítménye közti relációt a bővítmény esetragja definiálja.

A klasszifikációs (felügyelt tanulási) és klaszterezési (nem felügyelt) eljárások alkalmazásakor a siker szempontjából alapvető fontosságú kérdés, hogy milyen, a korpuszból kinyer-

hető, kvantifikálható tulajdonságok tükrözik leginkább azokat a lexikális tulajdonságokat, melyek köré a nyelvészetileg releváns osztályok szerveződnek. [2] reguláris mintákat használ. [13] és [14], [3] szintaktikailag elemzett korpuszból kinyert szubkategorizációs kereteket használnak. [13] kísérletképpen a szemantikai szelekciós információval is kiegészítette a kereteket, ám az összehasonlításból kiderült, hogy a pusztán szintaktikailag jellemzett keretek használata pontosabb eredményre vezet[1]. [7] az igeosztályokra jellemző szintaktikai alternációkat nyelvész szakértők által definiált jegykészletekkel közelítik. Módszerük előnye, hogy csak POS-taggelést igényel, ám hátránya, hogy az igeosztályok halmazát előre definiálni kell, az osztályokra jellemző jegykészlet kiválasztásához szakértői munkára van szükség, és időigényes. Ennek a hátránynak a kiküszöbölésére [11] létrehozott egy általános jegykészletet, mely tetszőleges osztályba tartozó angol igék felügyelt csoportosítására használható.

A magyar igék automatikus csoportosításakor nem állt rendelkezésünkre nyelvészek által meghatározott besorolás, így nem tudtuk előre, milyen szemantikai osztályokat szeretnénk eredményül kapni. Ezért mindenképpen nem felügyelt módszerhez kellett folyamodnunk. Igeosztályok híján nem ismerhettük az osztályokra jellemző alternációk halmazát sem, ami kizárta a magyar alternációkra jellemző közelítő jegyek definiálását. Így [13]-hoz hasonlóan az igéket szubkategorizációs kereteikkel jellemztük.

A Treebank annotációjának megfelelően a vonzatok mellett a szabad határozókat is a szubkategorizációs keret részének tekintettük. Ennek a döntésnek nem csupán gyakorlati oka van. A magyar nyelvre eddig nem született megbízhatóan alkalmazható vonzatteszt, aminek az az oka, hogy a vonzatok és a szabad határozók szintaktikai viselkedése hasonló. A felszíni sorrend alapján nem különböztethetjük meg őket, hiszen az ige és a vonzata közé beékelődhet egy vagy több szabad határozó is (szemben például az angollal). A szintaktikai funkciót jelölő esetragok különböző igék mellett különböző szerepeket kódolnak. Emmellett azt gondoljuk, hogy az ige melltt megjelenő adjunktumok legalább annyira jellemzőek az ige jelentésére, mint a vonzatok. Az adjunktumok sem produktívak abban az értelemben, hogy tetszőleges ige mellett használhatnánk őket: olyan igék mellett tehetők ki, melyek jelentése kompatibilis az adjunktum esetragja által kódolt szemantikai szereppel. Ezen megfontolásokból úgy döntöttünk, hogy az adjunktumokat is figyelmbe vesszük az igék szintaktikai környezetének jellemzésénel.

A treebankből kinyert szubkategorizációs kereteket azok az esetragos NP-k és főnévi igenevek alkotják, melyek az ige alá tartozó csomópontokban helyezkednek el. Az annotáció jellegéből fakadóan e gyermek-csomópontok kinyerése egy nem rendezett listát eredményez, melynek elemei az ige szintaktikai bővítményei. A szubkategorizációs keretek hosszának, vagyis a keretben szereplő bővítmények számának nem szabtunk felső határt. A kerettípusokat úgy állítottuk elő az egyedi keretekből, hogy elhagytuk a bővítmények lemmáját, morfoszintaktikai elemzésükből pedig csak a szófaj taget és az esetragot tartottuk meg. A bővítmények sorrendjét nem vettük figyelembe. Az általánosítás eredényeként 839 féle keretet kaptunk, melyeket mind megtartottuk.

[1] Az angol igék csoportosításának teszteléséhez Schulte im Walde Levin igeosztályait használta.

3 A klaszterezési eljárás

Mivel a kísérlet eredményeként a magyar igékre jellemző szemantikai igeosztályokra voltunk kíváncsiak, a klaszterezési eljárást a 150 leggyakoribb magyar igére alkalmaztuk. Az igék reprezentálásánál és a klaszterezési módszer kiválasztásánál [13]-ban leírt módszerre támaszkodtunk.Az igék korpuszbeli előfordulásait az ige és a különböző szubkategorizációs keretek együttes előfordulásainak maximum likelihood becslése adja:

$$p(t|v) = f(v,t) / f(v)$$

ahol *f(v)* az ige gyakorisága, *f(v,t)* pedig az ige és a keret együttes gyakorisága. Az értékeket a 150 igére és mind a 839 szubkategorizációs keretre kiszámoltuk.

A valószínűségi eloszlások összehasonlításához különbözőségi mértékként a relatív entrópiát használtuk:

$$D(x||y) = \sum_{i=1}^{n} x_i \cdot log\ (x_i / y_i)$$

A szubkategorizációs keretek nagy száma miatt az igék valószínűségi eloszlásában sokszor szerepel nulla érték. A relatív entrópia választása azzal jár, hogy ezeket az értékeket simítással korrigálni kell. Másfelől viszont nem akartuk elveszíteni azt az információt, amit a nulla számú előfordulás hordoz – vagyis az ige és a keretben megjelenő esetrag(ok) szemantikai inkompatibilitását. Mivel a leggyakoribb igékkel dolgoztunk, feltételeztük, hogy ezek a hiányok nem véletlenszerűek, és olyan simítási módszert akartunk választani, mely megfelel annak az elképzelésnek, hogy az együttes előfordulás hiánya egy gyakori ige esetében nagyobb eséllyel jelez inkompatibilitást, mint egy kevésbé gyakori lemma esetében (ahol inkább beszélhetünk véletlenről). Ezért az alábbi módszerrel simítottuk az adatokat:

$$f_e = 0{,}001 / f(v) \quad ha \quad f_c(t,v)=0$$

ahol f_e a becsült, f_c pedig a korpuszban mért gyakoriság.

Ezután két hierarchikus agglomeratív klaszterezési eljárást alkalmaztunk az adatokra:

1) Először kiindulásként a 150 ige mindegyikét egyelemű klaszternek tekintettük. Az interáció minden lépésénél kiszámoltuk az összes klaszter közti távolságot, azaz a relatív entrópiákjukat, és minden lépésben összevontuk a két leghasonlóbb klasztert. A klaszterek közötti távolságot minden lépésben újra kiszámoltuk. Ahogy Schulte im Walde is megjegyzi, ennek a módszernek az

a hátulütője, hogy néhány iteráció után az igék a legnépesebb klaszterek köré csoportosulnak, így kevés, nagy elemszámú csoportot eredményez. Hogy elkerüljük a problémát, meghatároztuk klaszterek maximális elemszámát: a negyedik elem után több igét nem olvasztottunk a klaszterbe. Az összevonást addig folytattuk, míg az intuitív értékelés alapján hasznosnak találtuk. Ez a módszer, valamint a korpusz mérete a 150 igéből 120 besorolását tette lehetővé, az összevonások folytatása ezután csökkentette volna a csoportok belső koherenciáját. Mindazonáltal így is szembesültünk a lánceffektussal (*chaining effect*), ami azt jelenti, hogy az eredményül kapott csoportok némelyikének legkevésbé hasonló tagjai közötti távolság túl nagy volt.

2) A második kísérletben a lánceffektus elkerülése céljából az elemszám helyett a klaszterek legkevésbé hasonló elemei közötti távolság maximális értékét határoztuk meg. Csak akkor soroltunk be egy új igét egy klaszterbe, ha a klaszter elemei közül a tőle legtávolabb eső igétől mért távolsági értéke kisebb volt, mint a – tesztfutások alapján meghatározott – maximális érték. Ezzel a módszerrel 71 igét sikerült 23 osztályba sorolnunk. Az első módszerrel szemben a második előnye, hogy képes nagy elemszámú, mégis koherens csoportok kialakítására, ami esetünkben különösen fontos, mivel a legalapvetőbb magyar szemantikai igeosztályokra vagyunk kíváncsiak. Ugyanakkor éppen ebből az okból későbbi terveink között szerepel nem agglomeratív, top-down módszerek kipróbálása is, melyek alkalmasabbak az adatok szerkezetének áttekintésére.

4 Eredmények

Mindkét fent ismertetett módszernél azt tapasztaltuk, hogy az igék egy része kis számú, népes csoportokba szerveződik, míg a maradékuk általában egy közeli szinonímájával (pl.: zár - végez) vagy ellentétpárjával (pl.: ül - áll) alkot egy klasztert. Természetesen az 1) módszer, vagyis a csoporton belüli igék számának korlátozása kevesebb klasztert ereményez és értékesebb eredményeket ad a kevésbé gyakori igék esetében. Ezzel szemben a második módszer, vagyis az egy csoportba sorolt igepárok közti maximális távolság korlátozása hatékonyabb az alapvető, nagy elemszámú igeosztályok meghatározására. Mivel az a célunk, hogy Levinéhez hasonló igeosztályokat találjunk a magyar nyelvben, a következő lépés annak megvizsgálása, hogy az igeosztályok koherensek-e szemantikailag, és ha igen, elemeik milyen jelentéskomponensekben osztoznak. Elsőként a három legnagyobb osztályt vizsgáltuk meg, mert a leggyakoribb igék közül sokat tartalmaznak, mégis – a módszer jellegzetességéből adódóan –belső koherencia jellemzi őket. A 71 kategorizált ige egyharmada a három

legnagyobb osztály valamelyikébe lett besorolva. Az alábbiakban ismertetjük az osztályokat[2].

C-1: létigék: *marad, van, lesz, nincs*

C-2: modálisok: *megpróbál, próbál, szokik, szeret, akar, elkezd, fog, kíván, kell*

C-3: mozgásigék: *indul, jön, elindul, megy, kimegy, elmegy*

Míg a C-1 és a C-3 osztályok erős szemantikai koherenciát mutatnak, a C-2 osztály leginkább a szintaktikai *modális* leírással jellemezhető. C-2 egy alosztályára laklamazható az a leírás, hogy valamilyen cselekvés elvégzésével kapcsolatos mentális attitűdöt fejeznek ki (*szeret, akar, kíván*), de ay osztály többi tagja esetében nehéz közös szemantikai tulajdonságot találni.
Általánosságban elmondható az eredményül kapott igeosztályokról, hogy lehorgonyozhatók valamilyen szemantikai jelentéskomponenshez, vagy jól jellemezhetők valamely argumentumuk közös szemantikai szerepével. Például: állapotávltozást jelentő igék: *erősödik, gyengül, emelkedik*; beneficiens argumentummal rendelkező igék: *biztosít, ad, nyújt, készít*; képességet jelentő igék: *tud, lehet, sikerül*. Néhány csoportot a fentieknél specifikusabb metapredikátummal jellemezhetünk – például külön csoportot alkotnak a megjelenést vagy az ítélkezést jelentő igék. Más esetekben viszont a szemantikai reláció sokkal kevésbé szoros, mint például az „ágenses, folytonos cselekvést jelentő igék" címkével leírható csoport esetében: *ül, áll, lakik, dolgozik*. A skála másik végén elhelyezkedő klaszterek olyan igéket tartalmaznak, melyek szemlátomást nem osztoznak közös jelentéskomponensben, pusztán „véletlenül" ugyanolyan esetraggal járó vonzatuk miatt kerültek egy csoportba: *foglalkozik, találkozik, rendelkezik.*

5 Értékelés

A szemantikai igeosztályok kiértékelésére még nincs bevett módszer. A létező eljárások két csoportba sorolhatók. Az első csoportba tartozó módszerek a csoporton belüli koherenciát vizsgálják. Másképpen szólva azt ellenőrzik, hogy egy független távolsági mérték használatával is azt az eredményt kapjuk-e, hogy a csoporton belüli elemek közelebb vannak egymáshoz, mint más csoportok elemeihez. Ezzel az eljárással azonban nem sokat tudunk meg a csoportok szemantikai koherenciájáról. A másik megoldást valamilyen kézileg előállított csoportosítással való összevetés jelenti. Az angol kísérletekben Levin osztályozását használják, ami a magyarra nem áll rendelkezésünkre. Kézenfekvő megoldás lehet a magyar WordNettel ([9]) való összehasonlítás is, azonban ezzel szemben is kifogások merülnek fel. Az osztályokat akkor tudjuk

[2] Mivel nem voltak előzetes osztályaink, az elnevzéseket utólag adtuk, az intuitív értékelés megkönnyítése céljából.

rávetíteni a hálóra, ha az osztály igéi közötti szemantikai kapcsolatot sikerül leképezni a WordNet-hierarchiára. Ha azonban az igei jelentések közti kapcsolatot a háló csomópontjaiban mérve fogalmazzuk meg, problémát jelent, hogy a WordNet csomópontok közti távolság nem egyenletes.
Mivel végsősoron Levin-típusú osztályzást akarunk kidolgozni a magyarra, az osztályokat kiértékelhetjük úgy is, hogy megpróbálunk az osztály igéire jellemző szintaktikai alternációkat keresni. Ehhez azonban figyelembe kell vennünk a magyar szintaxis jellegzetességeit, melyek megnehezítik az alternációk leírását. A lehetséges szubkategorizációs keretek nagy száma és a vonzatok nagy részének elhagyhatósága miatt túl sok lehetséges alternációval kell számolnunk. Ezért úgy döntöttünk, hogy a kezdetekben megpróbáljuk leszűkíteni vizsgálódásunk tárgyát. Kiindulásként megpróbáltuk meghatározni az egy csoportba sorolt igék közös jelentéskomponenseit. Ezután megvizsgáljuk, milyen szemantikai szerepeket kódoló bővítmények megjelenését engedélyezik ezek a jelentéskomponensek. Ha az egy osztályba sorolt igék ugyanazokkal a szemantikai szerepekkel kompatibilisek, és megegyeznek abban is, hogy milyen esetrag kódolja a szerepet, akkor az osztályt koherensnek tekintjük. Precízebb megfogalmazásban ez azt jelenti, hogy az igeosztályokat mátrixokkal ábrázoljuk, melyek oszlopait a főnévi esetragok, sorait az igei lemmák töltik ki, a cellákban pedig az a szemantikai szerep szeerepel, melyet az adott esetraggal megjelenő bővítmény az ige mellett betölt. Az osztály akkor koherens, ha a hozzá tartozó mátrix megfelel az alábbi két követelménynek:

1) A mátrix specifikus az adott igeosztályra.
2) Az egy oszlopba tartozó cellák ugyanazt a szerepet tartalmazzák.

A következő táblázatok koherens és osztályspecifikus mátrixokat tartalmaznak.

1. Táblázat: A C-3 klaszter igéi és a hozzájuk tartozó szemantikai szerepek

	ACC	**INS**	**ABL**	**ELA**
indul	-	eszköz - társ	forrás	Forrás
jön	-	eszköz - társ	forrás	Forrás
elindul	-	eszköz - társ	forrás	Forrás
megy	-	eszköz - társ	forrás	Forrás
elmegy	-	eszköz - társ	forrás	Forrás
kimegy	-	eszköz - társ	forrás	Forrás

2. Táblázat: A C-1 klaszter igéi és a hozzájuk tartozó szemantikai szerepek

	ACC	**INS**	**ABL**	**ELA**
marad	-	társ	ok	Összetevő
van	-	társ	ok	Összetevő
lesz	-	társ	ok	Összetevő
nincs	-	társ	ok	Összetevő

Amint az 1. táblázat mutatja, a C-3 csoport igéi mellett ablatívusz vagy elatívusz rag is jelölheti a forrás szerepű összetevőt. Az esetragok közti választás a főnévi csoporton múlik. A C-1 csoport igéi mellett ezek az esetragok más szemantikai szerepeket kódolnak.
Fontos megjegyezni, hogy nem rendelkezünk a szemantikai szerepek egy előre meghatározott halmazával. A megbízható kiértékeléshez szükséges, hogy az oszlopok kitöltését egymástól függetlenül több ember végezze.

6 Konklúzió és további teendők

A 150 leggyakoribb magyar ige szemantikai osztályokba sorolása egy kezdeti lépés az igei szintaxist meghatározó szemantikai tulajdonságok feltérképezése felé. Mivel nem volt előfeltételezésünk az eredményként vár igeosztályokról, és nem voltak az értékeléshez használható, kézzel kialakított csoportjaink sem, az eredmények kiértékeléséhez nyelvészeti elemzésre van szükség. Mindazonáltal a 4 részben bemutatott intuitív értékelés alapján azt mondhatjuk, hogy a kapott osztályok meglepően erős szemantikai koherenciát mutatnak. Figyelembe kell vennünk továbbá, hogy ezek a biztató eredmények rendkívül kis korpusz használatával születtek, ami megerősíti, hogy a *Semantic Base Hypothesis* jó eredménnyel használható szemantikai osztályok automatikus kinyeréséhez.
A feladat és az eredmények tükrében a további teendők két részterületre oszthatók. Egyfelől tervezzük további klaszterezési eljárások (például a [13] –ban leírt top-down eljárás) kipróbálását, valamint a jegykészlet finomítását, egyrészt a zajos jegyek kiszűrésével, másrészt újabb, releváns morfoszintaktikai jegyek bevonásával. Hosszabb távon szükséges a vizsgálódást kiterjeszteni a Magyar Nemzeti Szövegtár adataira, ehhez azonban szükség lesz a korpusz szövegének legalább részleges automatikus szintaktikai elemzésére. Másik fontos feladatunk a jövőben az igeosztályok nyelvészeti elemzése, mely az eredmények kiértékelésével szorosan összefügg. A kiértékeléshez használt mátrixok előállítását az igékhez társított példamondatok segítségével képzeljük megvalósítani.

Bibliográfia

1. Babarczy Anna, Gábor Bálint, Hamp Gábor, Kárpáti András, Rung András, Szakadát István.: *Hunpars. Mondattani elemző alkalmazás.* In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 20-28 .
2. Brent, M.: From Grammar to Lexicon: unsupervised learning of lexical syntax.. *Computational Linguistics*, 19(2): 243-262. MIT Press, 1993, Cambridge, MA, USA
3. Briscoe, T., Carrol, J.: Automatic Extraction of Subcategorization from Corpora. In: *Proceedings of the 5th Conference on Applied Natural Language Processing,* 1997, pp. 356-373, Washington DC, USA
4. Csendes Dóra, Csirik János, Gyimóthy Tibor, Kocsor András: The Szeged Treebank. *LNSC series Vol. 3658,* pp. 123-131

5. Dorr, Bonnie J., Jones, Doug: Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In: *Proceedings of the 14th International Conference on Computational Linguistics (COLING-96)*, 1996, pp. 322–327., Kopenhagen, Denmark.
6. Gábor, K., Héja, E.:.Vonzatok és szabad határozók szabályalapú kezelése. In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 245-257 .
7. Joanis, E., Stevenson, S.: A general feature space for automatic verb classification. In: *Proceedings of the 10th Conference of the EACL,* 2003, pp. 163-170, 200. Budapest, Hungary.
8. Koenig, J-P, Mauner, G., Bienvenue. B..: Arguments for Adjuncts. *Cognition,* 2003*,* 89, pp.67-103.
9. Kuti J., Vajda P., Varadsi K.: Javaslat a magyar igei WordNet kialakítására. In: *Alexin Z., Csendes D. (szerk): A Harmadik Magyar Számítógépes Nyelvészeti Konferencia Előadásainak kötete,* Szeged Egyetemi nyomda, 2005. Szeged, pp. 79-87.
10. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. Int. J. Digit. Libr. 1 (1997) 108–121
11. Merlo, P., Stevenson, S.: Automatic Verb Classification Based on Statistical Distributions of Argument Structure In: *Computational Linguistics. 27:3*, 2001, pp. 373-408.
12. Pereira, C. N. Fernando, Tishby, N., Lee, L.: Distributional Clustering of English Words. *31st Annual Meeting of the ACL*, 1993, Columbus, Ohio, USA, pp. 183-190.
13. Schulte im Walde, Sabine: Clustering Verbs Semantically According to their Alternation Behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, 2000, Saarbrücken, Németország, pp. 747–753.
14. Schulte im Walde, Sabine and Brew, Chris: Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, Philadelphia, PA, USA, pp. 223–230.
15. Váradi, T.: The Hungarian National Corpus. Proceedings of the Third International Conference on. Language Resources and Evaluation, 2002. Las Palmas pp.385-389