“Mazsola” – eszköz a magyar igék bővítményszerkezetének vizsgálatára

Sass Bálint

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Osztály
PPKE ITK, Multidiszciplináris Műszaki Tudományok Doktori Iskola
Nyelvtechnológia Doktori Program
joker@nytud.hu

Kivonat A gépi megértés hagyományos megközelítésének fontos lépése az igei vonzatkeretek kezelése. Az igei bővítmények statisztikai feldolgozása által viszont közvetlenül juthatunk az igék szemantikájáról szóló információhoz. A magyar egyszerű mondat egy modelljét egyfajta speciális kollokációként felfogva lehetővé vált a lényeges bővítmények megragadása az ún. *salience* mérték segítségével. Ezen módszer gyakorlati alkalmazásaként létrejött a *Mazsola*, egy internetes felületen hozzáférhető nyelvészeti kutatóeszköz, melynek segítségével a magyar igék bővítményszerkezetét lehet kvantitatívan tanulmányozni. Bemutatom a rendszer használatát és további felhasználási ötleteket vetek fel.

1 Bevezetés: a gépi megértés két útja

A számítógépes nyelvészet mint tudományág utópisztikus célkitűzése a tudományos fantasztikus művekből ismert *beszélő gép* létrehozása. Olyan gépről van szó, amely képes az emberi nyelvet adekvátan alkalmazni bemenetként és kimenetként is, azaz megérteni és használni is tudja (Lee 2004). Az uralkodó generatív nyelvészeti irányzattal párhuzamosan élő hagyomány szerint a nyelvről lényeges dolgokat tudhatunk meg statisztikai úton. John R. Firth szerint egy szó megismerhető azáltal, hogy tudjuk milyen más szavakkal szokott együtt előfordulni. („You shall know a word by the company it keeps.”) (Firth 1957). A '80-as években a számítógépek kapacitásának növekedésével a statisztikai irányzat újból fellendült, a nyelvészet számos területén kezdték sikerrel alkalmazni a statisztikai módszereket, kialakult a modern korpusznyelvészet. Lényegében ez az a módszertan, melyben a nyelvről szóló ismeretek alapvető forrása a nagy mennyiségű, statisztikailag feldolgozott valós nyelvi adat.

Hogyan közelíthetjük meg a megértést számítógépes eszközökkel? Az általánosan elfogadott hagyományos generatív keret szerint: a megértéshez először az egyes szavak jelentését kell egyértelműen meghatároznunk, majd ez alapján a nagyobb szerkezeti egységek, mondatok szerkezeti felépítését figyelembe véve juthatunk el azok jelentéséhez. Szükséges a szöveg szintaktikai elemzése, az igei vonzatkeretek felderítése valamint az egyes szemantikai összetevők azonosítása, az argumentumszerkezet feltérképezése.

I. Alknyelvdok Konferencia kötet. Szerk.: Váradi Tamás
MTA Nyelvtudományi Intézet, Budapest, 2007.
ISBN 978-963-9074-44-6

A korpusznyelvészet elveit és küldetését összegző cikkében Wolfgang Teubert a jelentésnek a fentitől gyökeresen eltérő megközelítését fogalmazza meg: „A jelentés körülírás." („Meaning is paraphrase.") E felfogás szerint adott jelentéssel bíró egység („unit of meaning") jelentését az egység körülírásai, átfogalmazásai adják, másképpen azon megnyilatkozásoknak az összessége, melyekben az adott egységről szó esik. („The meaning of the unit lemon is everything that has been said about lemons.") Teubert két dolgot mond tehát: adott jelentéssel bíró egység jelentését (1) az egység átfogalmazásai adják; (2) azon megnyilatkozásoknak az összessége adja, melyekben az adott egységről szó esik. Itt a hagyományossal ellentétes irány rajzolódik ki: mintegy a mondatokból származtatjuk a szavak jelentését (Teubert 2005).

Doktori munkám lényege összefoglalva: a teuberti álláspont korpuszalapú tesztelése, azaz, hogy (1) a parafrázisok összegyűjtése által valóban a megnyilatkozások jelentését tudjuk-e megfogni; (2) illetve hogy származik-e ebből egy implementálható gépi módszer a jelentések megközelítéséhez. A hagyományos úton elindulva, az volt a szándékom, hogy a hagyományos módon kidolgozott szemantikai annotáció hasonlóságából vezessem le a parafrázisokat. Az igei vonzatkeretek, igei bővítmények vizsgálata során tapasztaltam, hogy azok statisztikai feldolgozása által az igék szemantikájáról lehet valamit közvetlenül mondani. Ennek a megfigyelésnek a kapcsán jött létre ez a tanulmány.

A továbbiakban vázolom a magyar mondatnak azt az egyszerű modelljét, ami alapján a vizsgálatokban a mondat igéjének és a bővítményeknek a viszonyát ábrázoltam (2. rész). Bemutatok egy hasznos mértéket – az ún. *salience*-t –, mellyel a lényeges bővítmények megfoghatók. Ismertetem azt a módot, ahogyan ezt a két szó kollokacionalitásának vizsgálatára kifejlesztett mértéket az igékre és bővítményeikre alkalmaztam (3. rész). Szó lesz egy ezt a módszert megvalósító új kutatóeszközről, mely magyar igei bővítményszerkezet tanulmányozására szolgál (4. rész). Végül az eszköz alkalmazásáról fogok beszámolni, illetve javaslatokat teszek a további felhasználásra(5. rész).

2 A magyar egyszerű mondat modellje

Korábban kidolgoztam egy módszert, melynek segítségével korpuszból nyerhetünk ki új vonzatkereteket. A *vonzatok: nem kompozicionális bővítmények* elvére alapozva az igék és a mellettük álló NP-k alkotta keretjelöltek idiómaságát mértem az ún. *distributed frequency* idiómasági mértékkel. Eszerint a mérték szerint az a keret az idiomatikusabb, melynek bővítményei az adott formában kevés (szélső esetben egyetlen) igével fordulnak elő (pl. *fittyet hány*). A magasabb idiómasági érték azt jelenti, hogy a keretjelölt valódi vonzatkeret (Sass 2006a).

A vonzatok és a szabad határozók elkülönítésében, mely sok esetben elméletileg is nehezen kezelhető kérdés, ez a módszer nem bizonyult elég megbízhatónak. Létezik azonban a kereteknek egy, a valódi vonzatkereteknél valószínűleg bővebb halmaza mely az ige jelentéskincse szempontjából valamint a gépi fordítás szemszögéből is *lényeges*nek mondható, ide tartozik például: *hajat vág*, *fésüli a haját*, *választ ad valamire*, *meggyőződésének/véleményének ad hangot*. Figyelmem a valódi vonzatkeretek helyett az igék bővítményszerkezete, a lényeges bővítmények és a

lényeges igei keretek felé fordult. Azaz a továbbiakban nem mérlegelem, hogy mi vonzat és mi szabad határozó, csak azzal foglalkozom, hogy melyik bővítmény lényeges.

Az ige *bővítményszerkezete* alatt az igének azt a tulajdonságát értem, hogy milyen bővítményekkel, főnévi csoportokkal szokott általában együtt előfordulni. Ez implicit módon tartalmazza a gyakorisági szempontot is, azaz a gyakoribb *bővítménykeret* fontosabb lesz. Például a *forog* bővítményszerkezetének négy legfontosabb bővítménykerete a következő: *forog valami*, *szeme vérben forog*, *élete valami körül forog*, *kockán forog valami*. A *kockán forog valami* típusú szerkezeteket, amelyeknek az alapigétől eltérő jelentése van *összetett igé*nek fogom nevezni. Ide tartoznak az igemódosítós igék, és általában véve azok a szerkezetek, amikor egy vagy több főnévi csoport szervesen hozzátartozik az ige jelentéséhez, mint például a *górcső alá vesz valamit* esetében a *górcső alá*. Az összetett igék a legtöbb esetben önálló, az alapigétől független bővítményszerkezettel bírnak, ezért is önálló igéknek, az igék egy csoportjának tekintem őket.

A magyar egyszerű mondat szórendjéről elmondhatjuk, hogy az ige és az egyes főnévi csoportok szinte bármilyen sorrendben elhelyezkedhetnek a mondatban. A funkcionális elemzésre emlékeztető módon az egyszerű mondatot felfoghatjuk egy halmaznak, amiben egy ige és valahány főnévi csoport van. A főnévi csoportokat reprezentálhatjuk a csoport fejét adó szótővel és a fej esetragjával/névutójával. Ennek megfelelően a magyar egyszerű mondat illetve egy bővítménykeret reprezentációja a következő lehet:

ige + NP(szótő+eset)–lista

A keretekre a továbbiakban a `'kér bocsánat-t -tÓl'` formában fogok hivatkozni. Első helyen mindig az igét tüntetem fel, utána következnek a bővítmények: esetleges szótő majd kötőjel után az esetrag vagy a névutó. Az esetragban az illeszkedő magánhangzó helyén mindig a hátulképzett változat nagybetűs alakja jelenik meg. A példa tehát a „bocsánatot kér valakitől” keret szabványos formája. Egy bővítménykeret egy *pozíció*ja alatt nem fizikai, sorrendi pozíciót értek, hanem egyfajta morfoszintaktikai-funkcionális pozíciót, például tárgyi, datívuszi, -bAn-i stb. pozíciót. Egy pozíciót *szabad*nak fogok nevezni, ha nincsen meghatározva a hozzá tartozó névszói fej.

A számítógépes nyelvészetben bevett fogalom az *n-gram*, amely egyszerűen *n* darab egymást követő szót jelent. Ezt a fogalmat terjeszthetjük ki úgy – ezt nevezik *concgram*nek –, hogy a szavak között egyéb közbeékelődő szót is megengedünk, valamint a szavak sorrendjét sem kötjük meg (Cheng 2006). Egy bővítménykeret elemei a mondatban tetszőleges sorrendben fordulhatnak elő, és mellettük további bővítmények is megjelenhetnek, így a magyar egyszerű mondatot egy olyan concgramnek foghatjuk fel, melyben az egyes pozíciókat frázisok képviselik, illetve az ezeket reprezentáló névszói fejek.

A kollokáció szokásos két egymás melletti szóra (egy *2-gramra*) vonatkozó definícióját kiterjeszthetjük a concgram struktúrára. Másképp fogalmazva arról van szó, hogy a kollokáció kifejezést itt abban a tág értelemben használom, hogy „együttes előfordulás”. Az egy tagmondaton belüli tetszőleges sorrendű, akár megszakított együttes előfordulásról van itt szó, a bővítmények sorrendje illetve egymás mellettisége nem számít, csak az, hogy az igével egy mondatban vannak.

Ezáltal a bővítménykeretek felfoghatók kollokációknak, és a lényeges kereteket mint lényeges kollokációkat vizsgálhatjuk.

A modellből következik, hogy a feldolgozás során milyen lépéseket volt szükséges megtenni. A Magyar Nemzeti Szövegtár (Váradi 2002) egy részkorpuszának szövegét először heurisztikus eljárással tagmondatokra, illetve egy igét, és ezáltal egy bővítménykeretet tartalmazó egységekre bontottam (Sass 2006b). A részleges szintaktikai elemzés eredménye: a kerethordozó ige és a mellette álló főnévi csoportok azonosítása (Sass 2005). Az elemzőnek része az igekötő- és igetőazonosítás, az elváló igekötőket az igetőhöz kapcsoltam, a gyakori képzőket (illetve egyelőre csak a -hAt képzőt) levágtam, mivel nem befolyásolják a vonzatkeretet. A határozószókat figyelmen kívül hagytam, csakis a főnévi csoport által képviselt bővítményeket vettem tekintetbe. A rendszer tartalmazza azt az egyszerűsítést, hogy a bővítménykeretek minden pozíciójában csak egy darab főnévi csoportot enged meg. Ha egy mondatban több azonos esetű főnévi csoport is szerepelt, akkor azok közül csak a legutolsót vettem tekintetbe. Csak olyan tagmondatokkal foglalkoztam, melyekben található ige vagy főnévi igenév. Utóbbi is elegendő, mert a főnévi igenév hordozza a neki megfelelő ige vonzatkeretét, azaz ilyenkor a főnévi igenév tövét tekintettem a mondat igéjének, és a névszói csoportokat bővítményként hozzárendeltem.

Beth Levin 1993-as művében jelenik meg az igék szemantikájának és viselkedésének kapcsolatáról szóló híres hipotézise, mely kimondja, hogy „az ige viselkedése, különösen az argumentumainak kifejez(őd)ése és értelmezése tekintetében nagy mértékben függ az ige jelentésétől.” (Levin 1993) Ez a fenti modell alapján számunkra leegyszerűsítve úgy fogalmazható meg, hogy hasonló igei jelentés hasonló bővítményszerkezettel jár. Például látható, hogy a jelentésben nagyon szoros hasonlóságot mutató *reagál*, *válaszol*, és *felel* bővítményszerkezetben is hasonló (nagyon gyakran jár -rA ragos bővítménnyel, ami sokszor a *kérdés* szótővel jelenik meg).

A statisztikai módszerrel felfedezett hasonló bővítményszerkezet kijelöli a hasonló jelentésű igéket, ilyen igék csoportjait. Ezek az igék egymás átfogalmazásai, parafrázisai. Ha két mondatunk van, az egyikben az egyik, a másikban pedig a másik ige szerepel, és a két mondat bővítményszerkezete azonos (vagy legalábbis hasonló), akkor kimondhatjuk, hogy a két mondat szemantikailag hasonló, hogy az egyik a másiknak parafrázisa. Ez lehet az egyik módszer parafrázisok gyűjtésére, így az igék és bővítményeik alkotta lényeges kollokációk által közvetlen visszacsatolódás jöhet létre a szemantikai szinthez, azaz közelebb juthatunk az 1. részben megfogalmazott eredeti célhoz.

3 A lényegesség mérése

A valódi vonzatkeretek megragadására nincs megbízható automatikus eszközünk. A lényeges kollokációkat ezzel szemben viszonylag egyszerűen meg tudjuk fogni automatikusan: mégpedig az ún. *salience* (lényegesség, jellegzetesség) mérték segítségével (Kilgarriff 2001). Ezt a két elem együttes előfordulásának vizsgálatára kidolgozott mértéket alkalmazzuk itt azzal a lényegi trükkel, hogy az egyik elem a

vizsgálandó bővítménykeret egyik (kiválasztott) pozíciója lesz, a másik elem pedig az ige az esetlegesen mellette megkövetelt egyéb bővítményekkel együtt. Ez megtehető, hiszen szabadon dönthetünk arról, hogy mit veszünk egy kollokáció egy elemének (Kilgarriff 2001). Így valójában az adott bővítménynek az igei keret többi részéhez viszonyított lényegességét tudjuk mérni.

A tipikus kérdés tehát, amit vizsgálni tudunk: adott ige illetve igei keret melletti adott pozícióban mely szavak fordulnak elő legjellemzőbben. A megkövetelt egyéb bővítmény bármi lehet: igemódosító, vonzat vagy szabad határozó is, a bővítménykeret formális fogalmába mindegyik beletartozik. A kérdésben megadhatunk egy igetövet és valahány főnévi csoportot, függetlenül attól, hogy a főnévi csoportnak adott esetben mi a szerepe, és megnézhetjük, hogy egy további pozíción milyen szavak jelennek meg. Ezáltal vizsgálható az összetett igék önálló bővítményszerkezete, valamint összevethető egy alapige és egy összetett ige bővítményszerkezete is. Példa: *x* = `'ad hang-t'`; *y* = `'meggyőződés-nAk'`, `'vélemény-nAk'` stb. A fix elem az *x*, a vizsgált elem az *y*, a kérdés pedig az, hogy az egyes *y*-ok közül melyek a jellemzőek.

A kollokációk keresésére használt klasszikus mérték, a *kölcsönös információ* (*mutual information*) a következő képlettel adható meg:

$$MI(x,y) = \log_2 N \left(f(x,y) / f(x) \cdot f(y) \right) \quad (\mathbf{1})$$

ahol *N* a korpusz mérete, *f* az előfordulási szám, *x* a fix elem, *y* pedig a vizsgált elem. Ennek akkor magas az értéke, ha a két elem a vártnál gyakrabban fordul elő együtt. Hátrányos tulajdonsága, hogy a túlzottan kiemeli a ritka elemeket (Sass 2006b). Gondoljuk meg:

Ha *y* hapax és éppen *x*-szel együtt fordul elő, akkor f(y) = 1, f(x,y) = 1, azaz

$$MI(x,y) = \log_2 N \left(1 / f(x) \cdot 1 \right) = \log_2 N \left(1 / f(x) \right) \quad (\mathbf{2})$$

Ha *y* előfordulási száma 500, és ebből 250-szer *x*-szel együtt fordul elő, akkor f(y) = 500, f(x,y) = 250, azaz

$$MI(x,y) = \log_2 N \left(250 / f(x) \cdot 500 \right) = \log_2 N \left(1 / 2f(x) \right) \quad (\mathbf{3})$$

Az előbbi esetben nagyobb értéket kapunk, mert ez a mérték annak tulajdonít nagy jelentőséget, hogy az *összes y*-re igaz, hogy *x*-szel együtt fordult elő, hiába igaz az is, hogy az *y* előfordulási száma mindössze 1.

Egy elfogadott megoldás az, hogy az *MI* értéket korrigáljuk a vizsgált elem (*y*) előfordulási számának a logaritmusával, így kapjuk meg a már említett *salience* mértéket (Kilgarriff 2001).

$$S(x,y) = \log_2 f(y) \cdot MI(x,y) \quad (\mathbf{4})$$

A salience szerint rendezett listában valóban a tipikus, lényeges kollokációk kerülnek a lista elejére.

Nézzük meg egy konkrét példán a két mérték különbségét. Az `'ad -t'` keret esetében az *MI* mérték szerinti csökkenő sorrendben a *tanújel, életjel, ízelítő, személyleírás, áldás* szavakat kapjuk. A salience viszont a *hang, lehetőség, válasz, otthon, tájékoztatás* listát szolgáltatja. Az előbbiek ritka, különleges szavak, az utóbbiak a triviálisabbnak tűnők, mégis ezek a lényegesek az ige jelentéstartománya és a a gépi fordítás szemszögéből. Mondhatjuk: az *MI* nem a lényegeset, hanem a

különlegeset mutatja. Az *MI* által mutatott listára az anyanyelvi beszélő is rácsodálkozhat, hogy tényleg ezeket is `'ad -t'` keret „aleseteként” fejezzük ki, de amiket leginkább érdemes tudni, ha meg akarunk érteni egy magyar szöveget, azok a salience által adott listában találhatók. Egyszerűen fogalmazva hasznosabb ha egy gépi fordító rendszer helyesen le tudja fordítani a `'ad válasz-t -rA'` keretet, mintha helyette az `'ad ízelítő-t -bÓl'` keretet kezelné jól.

A modellből és a salience alkalmazási módjából következően a fix elem (*x*, ti. az ige és a mellette megkövetelt nem vizsgált bővítmények) bonyolult és sokféle lehet. A fix elemek gyakoriságát nem kell kiszámolni, nem kell számon tartani a következő összefüggés miatt.

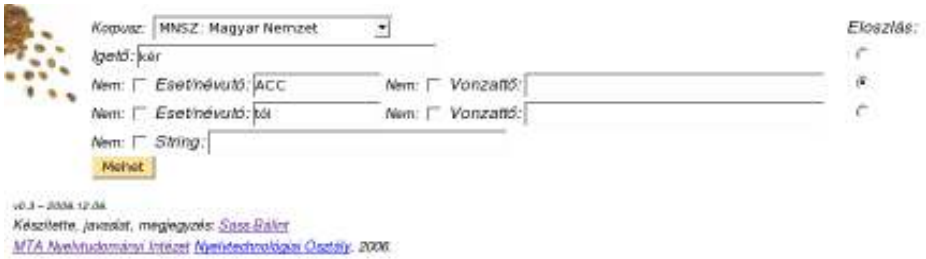
$$\begin{aligned} MI(x,y) &= \log_2 N \left(f(x,y) / f(x) \cdot f(y) \right) \\ &= \log_2 N/f(x) \cdot f(x,y) / f(x) \\ &= \log_2 C \cdot f(x,y) / f(x) \\ &= \log_2 C + \log_2 f(x,y) / f(x) \end{aligned} \qquad (5)$$

A vizsgálat mindig arra irányul, hogy melyek a leglényegesebb bővítmények. Mivel csak *össze akarunk vetni* azonos fix elemet tartalmazó kereteket, a konkrét *MI* (és az abból származtatott salience) értékek nem érdekesek. Két ilyen *MI* érték különbségének kiszámításakor pedig az f(y)-tól nem függő első tag kiesik, azaz az összehasonlításhoz nincs szükség az f(x) értékre.

4 Kutatóeszköz

A fenti módszerek gyakorlati alkalmazására létrejött a *Mazsola*, egy internetes felületen hozzáférhető nyelvészeti kutatóeszköz, melynek segítségével a magyar igék bővítményszerkezetét lehet kvantitatívan tanulmányozni. Az elnevezés onnan ered, hogy reményeim szerint izgalmas nyelvi tényeket mazsolázhatunk ki vele a korpuszokból.

4.1 Használata



1. ábra A *Mazsola* felülete

Az 1. ábrán látható felület a `http://corpus.nytud.hu/mazsola` címen érhető el. Az első sorban a korpuszt választhatjuk ki. Jelenleg a 2. ábrán látható három

kisebb korpuszban végezhetünk vizsgálatokat. Mindhárom a Magyar Nemzeti Szövegtár egy részkorpusza. Az első az MNSZ speciális „egyszerű” mondatainak gyűjteménye: a 3-10 szavas írásjelet nem tartalmazó és ezáltal jó eséllyel eleve egy keretet tartalmazó mondatait tartalmazza, itt tehát a tagmondatra bontó lépés elmaradhatott. Ebben a korpuszban az eredetileg többször előforduló mondatok csak egyszer vannak benne. A másik kettő (amelyek az elsőt értelemszerűen átfedik), az MNSZ két nagyon elütő stílusrétegét képviseli.

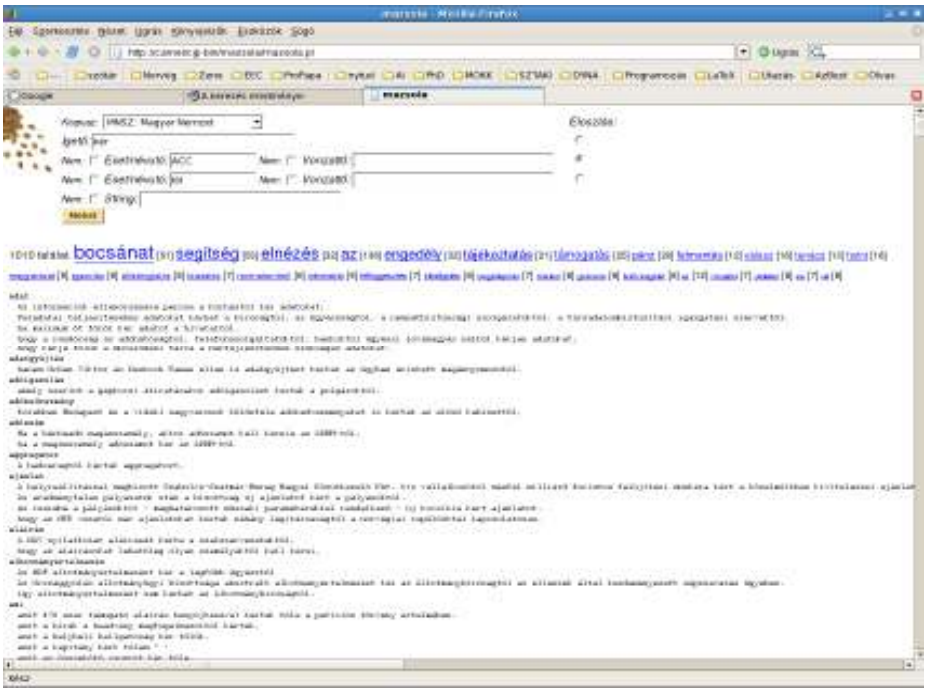
3-10 szavas mondatok	8 millió szó
Magyar Nemzet napilap anyaga	12 millió szó
Index fórum anyaga	18 millió szó

2. ábra A *Mazsola* közzétett korpuszai

A felületen megadhatjuk az igetövet, alatta pedig egy vagy két bővítményt specifikálhatunk. A modellnek megfelelően a bővítmény fejének esetragját vagy névutóját, illetve a bővítmény fejét adó névszó szótövét (utóbbit a *Vonzattő* mezőben) lehet megadni. A szótőnél használhatunk szóközzel elválasztott szótőlistát is. Az esetet a szokásos latin elnevezés hárombetűs kódja (ACC, DAT, ILL stb.) mellett többféle kiírt formában is elfogadja a felület, helyes megadás például: t, tárgy, -bA, babe, -ba / -be stb.

A sor végén az *Eloszlás* alatt kell megjelölni, hogy melyik az a pozíció, amelyet vizsgálni akarunk, azaz hogy melyik bővítményi pozícióban megjelenő szótövek salience szerinti listáját kérjük. A tagadás értelemszerűen a *Nem* jelölőnégyzet megjelölésével történik. Kétféleképpen használható: vagy kizárjuk bizonyos esetű bővítmény jelenlétét (sor elején álló *Nem*), vagy pedig amellett, hogy megköveteljük bizonyos esetű bővítmény jelenlétét, kizárunk bizonyos szótöveket (sor közepén lévő *Nem*). A legalsó sorban, egy szintén tagadható mezőben szabadszavas kereséssel szűkíthető a vizsgálat. Itt szóközzel elválasztva több szót is megadhatunk, illetve tetszőleges kiterjesztett reguláris kifejezést használhatunk. Az ábrán látható példában arra kérdezünk rá, hogy a `'kér -t -tÓl'` keret tárgyi pozícióját milyen lényeges szavak adják.

A megjelenő (3. ábrán látható) válaszképernyő három részből áll. A már ismert lekérdezőfelület alatt a kívánt pozíciót kitöltő szótövek listáját látjuk, salience-érték szerinti csökkenő sorrendben, azaz az első helyen a leglényegesebb szótő áll. Alul az összes releváns korpuszpéldát is megkapjuk a vizsgált pozíciót kitöltő szótő szerint csoportosítva. A lényeges szavak „tartalomjegyzékében” csak az 5-nél gyakoribb szótövek szerepelnek, nagyobb betűméret jelzi a nagyobb salience-értéket, szögletes zárójelben tájékoztatásképpen az előfordulási szám van feltüntetve, a szavakra kattintva azonnal elérhetők a hozzájuk tartozó korpuszpéldák. Az ábrán látható példában azt kaptuk, hogy a `'kér -t -tÓl'` keret tárgyi pozícióját legjellemzőbben a ezek a szavak adják: *bocsánat, segítség, elnézés, engedély, tájékoztatás, támogatás, pénz* stb.



3. ábra A *Mazsola* válaszképernyője

Ha alaposabban megnézzük az egyes korpuszpéldákat, látszik hogy számos esetben valamilyen hiba folytán helytelen eredményre jut a rendszer, azaz helytelenül állapítja meg az igét és/vagy a bővítményeket. A hibák forrásának egy része a Magyar Nemzeti Szövegtár morfoszintaktikai elemzésének hibájából adódik, ezen felül a tagmondatrabontás és a részleges szintaktikai elemzés sem tökéletes. Bár az egyes mondatról sok esetben hibás specifikus megállapítást tesz a rendszer, ettől még igaz az, hogy a bővítmények lényegességéről szóló általános állítások megfogalmazásához biztos alapot ad. Megerősíthetjük azt az ismert tényt, hogy a statisztikai alapú általános állítások igazságára az alkalmazott eljárásban előforduló ritka hibák nincsenek számottevő hatással (Teubert 2005, Kilgarriff 2004).

4.2 Példák

A közzétett korpuszok például lehetőséget adnak a különféle stílusrétegű szövegek bővítményszerkezetének összehasonlítására. Látni fogjuk, hogy a különböző stílusrégetű szövegek bővítményszerkezetükben is különböznek. Alább a Magyar Nemzet és az Index fórum korpuszból nyert, az `'ad -t'` keretre vonatkozó adatokat elemzem. A 4. ábrán láthatók a tárgyi pozíciót betöltő lényeges szavak, itt a közös elemeket jelöltem meg. Ezek azok a szavak, keretek, melyek stílusrétegtől függetlenül lényegesek. Az 5. ábrán látható ugyanez az összehasonlítás, de itt az eltérő szavak vannak kiemelve. Valóban, az `'ad otthon-t -nAk'` és az `'ad hír-t -`

`rÓl'` sajtónyelvbe illő keretek, szemben az `'ad igaz-t'` és az `'ad tipp-t'` kollokviális, hétköznapi, beszélt nyelvi jellegével. A 6. ábrán az `'ad hang-t -nAk'` keret szabad datívuszi pozíciójának jellemző szavait látjuk. Itt is találunk hasonlóságot és különbséget is. Ez a keret sokkal gyakoribb a formálisabb nyelvben, vannak közös elemek (*aggodalom*, *vélemény*) és specifikus, csak az egyik stílusrétegre jellemző szavak (*meggyőződés* illetve *felháborodás*).

4. ábra `'ad -t'` keret a Magyar Nemzet (fent) és Index fórum (lent) korpuszban: közös elemek

5. ábra `'ad -t'` keret a Magyar Nemzet (fent) és Index fórum (lent) korpuszban: eltérések

6. ábra `'ad hang-t -nAk'` keret
a Magyar Nemzet (fent) és Index fórum (lent) korpuszban

Láttuk, hogy az `'ad -t'` keretben megjelenő lényeges tárgyi bővítmények is sok esetben állandósult szókapcsolatot, összetett igét alkotnak az alapigével. Nézzünk most erre még néhány példát a 7. ábrán, mely azt is illusztrálja, hogy az esetragokat és a névutókat valóban teljesen egyenrangúan kezeli a rendszer. Ha egy keret egy pozícióján igemódosítók, vagy igemódosítók *is* vannak, a salience ezeket hozza elő. Ez lehetőséget ad összetett igék felfedezésére illetve az összetett igék saját, önálló, az alapigétől legtöbb esetben független bővítményszerkezetének vizsgálatára. Érdekes

jelenség, hogy egy anyanyelvi beszélő a szabad keretből nehezen találja ki a jellemző bővítményi fejet, ugyanakkor a fordított irányú asszociáció (bővítményhez az igét) azonnali. Ilyen aszimmetrikus asszociációs viszonnyal van dolgunk a 7. ábrán található összes esetben.

'hány -t'	→ fitty
'hány -rA'	→ szem
'kerül alá'	→ víz, kalapács, fennhatóság
'rejt alá'	→ véka
'hoz alá'	→ tető
'helyez alá'	→ vád
'vesz alá'	→ górcső, kalap, tűz

7. ábra Lényeges bővítmények – állandósult szókapcsolatok

5 Alkalmazás

A COBUILD szótár óta ismert, hogy a korpuszok fontos segédeszközt jelentenek a lexikográfiai munkában. A korpuszból származó adatok, konkordanciák elemzése a hagyományosnál objektívebb munkát tesz lehetővé, eredményeképpen a szótár anyaga teljesebb lehet. A szótáríró elszakadhat idiolektusától, szembenézhet a valós nyelvhasználattal, és egyes szavaknak olyan jelentésére, használatára bukkanhat, melyek a korábbi szótárakban nem szerepelnek.

A manapság elérhető nagyméretű korpuszok elég pontos képet adnak a nyelvről, de méretükből adódóan képtelenség az összes releváns adat manuális feldolgozása. Szükség van egy olyan eszközre, mely egy bizonyos nyelvi jelenségről valamiképpen összegzi a korpuszból leszűrhető tényeket. Az első ilyen eszköz az ún. *Word Sketch* volt, amit a 2002-ben megjelent Macmillan English Dictionary készítésekor alkalmaztak (Kilgarriff 2004). A *Mazsola* is ilyenfajta eszköz, ennek megfelelően hasznos lehet különféle lexikográfiai feladatok során.

A *Mazsolát* aktívan használjuk egy magyar-angol gépi fordító projektben az igei keretek egyes szabad pozícióinak különféle fix szótövekkel való leköthetőségének vizsgálatakor, az így kapott különféle keretek elemzésekor, osztályozásakor. A magyar WordNet kialakítását célzó projektben is alkalmazzuk, mikor arról kell dönteni, hogy adott igének van-e egyéb jelentése illetve, hogy milyen feltételek mellett, milyen bővítmények jelenléte esetén van egyéb jelentése és ezek alapján melyik másik ige szinonímájának tekinthető. Tervek vannak arról, hogy a jövőben egy német-magyar vonzatszótár munkálatait fogjuk segíteni vele. A dolgozat hátralévő részében ötleteket szeretnék adni a további felhasználáshoz, új kutatási lehetőségeket szeretnék felvillantani.

5.1 Igék szemantikai osztályozása

Levin 2. rész végén idézett hipotézise szerint hasonló jelentés hasonló bővítményszerkezettel jár. Ez azt is jelenti, hogy ha két hasonló bővítményszerkezetű

igét találunk, akkor azt a következtetést vonhatjuk le, hogy a két ige szemantikailag is hasonló. Ebből előállhat egy automatikus számítógépes módszer, mellyel magyar szemantikai igeosztályokat tudunk kialakítani a bővítményszerkezet hasonlósága alapján (Schulte im Walde 2006).

nő	***növekedik***
alany: szám arány ár	*alany:* szám arány érdeklődés
-bAn: mérték év	*-bAn:* mérték év
-rA: forint százalék dupla	*-rA:* év
-vAl: százalék forint	*-vAl:* százalék forint
emelkedik	***drágul***
alany: szám ár árfolyam	*alany:* részvény TVK kenyér
-bAn: mérték év	*-bAn:* mérték forgalom
-rA: forint jogerő százalék	*-rA:* forint
-vAl: százalék forint	*-vAl:* százalék forint

8. ábra Az egy szemantikai igeosztályba tartozó igék bővítményszerkezetének hasonlósága. Az egyes bővítményi pozíciókat kitöltő szavak salience-érték szerinti csökkenő sorrendben szerepelnek.

A 8. ábrán látható esettanulmány ad képet erről az elgondolásról. A *nő, növekedik, emelkedik, drágul* igék intuitíve egy szemantikai csoportba tartoznak: mindegyik valamiféle többé válást fejez ki. Ha összevetjük a mellettük megjelenő lényeges bővítményeket, akkor nagymértékű hasonlóságot látunk. Érezzük azt is, hogy a *drágul* lóg ki legjobban ebből a csoportból, ennek megfelelően a bővítményszerkezete is kissé elüt a többitől. Azt mondhatjuk, hogy a bővítményszerkezet hasonlósága "arányos" az igék szemantikai hasonlóságával. Ugyanilyen eredményre jutunk egyéb hasonló jelentésű igék vizsgálatakor, például a *nyújt, megad, kínál* vagy a *reagál, válaszol, felel* osztály esetében.

Ezt a módszert kidolgozva automatikusan magyar igei szinonímaszótárt építhetünk, esetlegesen olyan funkcióval, ami megmondja, hogy jelentés szempontjából milyen távol van egymástól két ige. Távlatilag pedig közelebb juthatunk a gépi megértéshez, amint ezt a 2. rész végén olvasható megközelítés ígéri.

5.2 Bővítmények szemantikai osztályozása

Levin feltételezésének az az értelmezése is helytálló, miszerint az igék szemantikai megkötéseket tesznek az argumentumaikra. Ebből az következik, hogy adott igéhez tartozó adott pozíciót kitöltő szavak egy szemantikai csoportot alkotnak. Ez nyújt lehetőséget a bővítmények szemantikai osztályozására, melyet a 9. ábrán található példákkal illusztrálok. Az első kettő kivételével a következő példákat az MTA Nyelvtudományi Intézet kézzel szerkesztett vonzatkeret-táblázatából vettem, megnéztem, hogy mennyire felelnek meg az adott pozícióban megkövetelt szemantikus jegynek a tényleges nyelvi adatok. A keret után kettőspontot követően tüntetem fel a szemantikus jegyet (a + a jegy meglétét, a − a jegy hiányát jelöli), a

következő sorban a keretbeli szabad pozíciót kitöltő leglényegesebb szavak *Mazsola* által szolgáltatott listája, majd értelmező megjegyzés következik.

- `'beszélget -vAl': +ember`
  polgármester, ő, aki, vezető, egymás, elnök, igazgató, ember
  Intuitív elvárásunk, hogy a bővítmény itt ember legyen, maradéktalanul teljesül a listára.
- `'alakul alany': +absztrakt`
  sors, dolog, bizottság, kormány, szervezet, helyzet
  Azon túl, hogy az absztraktság teljesül, megállapíthatjuk, hogy legtöbbször valamilyen szervezetről van szó.
- `'ásít alany': +ember`
  A tapasztalat az, hogy nem tekinthetjük kizárólagosnak ezt a jegyet. Korpuszpélda a *Mazsolából*: "A terem üresen ásított."
- `'átél alany': +ember`
  Az eset hasonló az előzőhöz, korpuszpélda: "Az egyezség válságos napokat élt át."
- `'aggat -t': -ember`
  jelző, kép, lámpácska, lufi, plecsni, virágfűzér
  A bővítmények megfelelnek a szemantikai jegynek.
- `'amputál -t': +testrész`
  láb, hüvelykujj, kar, ujj
  A bővítmények megfelelnek a szemantikai jegynek.
- `'kamatoztat -t': +absztrakt`
  tudás, tapasztalat, képesség, tehetség
  A bővítmények megfelelnek a szemantikai jegynek.
- `'alábbhagy alany': -ember`
  érdeklődés, hevesség, kedv, lendület, pánik
  Talán szigoríthatjuk a szemantikai feltételeket: minden esetben valamilyen mentális állapot az alany.
- `'kényszerít -rA': +absztrakt`
  átadás, térd, elhagyás, prostitúció, meghátrálás, távozás
  Szinte mindegyik bővítmény -Ás képzős, esetleg lehetséges, hogy az igének ez egyfajta szintaktikai megkötése a bővítményre.

9. ábra Példák bővítmények szemantikai osztályozására

Látjuk, hogy az utolsó példánál a *térd* nem illik a sorba, a `'kényszerít térd-rA'` nem tartozik ebbe a szemantikai osztályba. Érdemes lehet a fenti két szemantikai osztályozó eljárást – az igékre és a bővítményekre vonatkozót – összekapcsolni: ha látjuk, hogy adott szó nem illik bele a bővítmény szemantikai osztályába akkor felmerül, hogy új összetett igét találtunk. Ez a 8. ábrán a `'emelkedik jogerő-rA'` példáján is jól látható.

6 Befejezés

A magyar igék bővítményszerkezetének, a lényeges bővítmények vizsgálatára szolgáló *Mazsola* kutatóeszköz hozzáférhető a `http://corpus.nytud.hu/mazsola` címen. Az érdeklődők, akik kutatásukhoz fel szeretnék használni, vagy valamilyen alkalmazást szeretnének rá építeni személyes jelszót a `joker@nytud.hu` címen igényelhetnek a teljes hozzáféréshez. Várható, hogy a jövőben további részkorpuszok, illetve a teljes Magyar Nemzeti Szövegtár elérhető lesz ebben a formában.

Irodalom

Cheng, W., Greaves C., Warren M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* Vol. 11. No. 4. 411-433.

Firth, J.R. 1957. *A Synopsis of Linguistic Theory 1930-55.* Studies in Linguistic Analysis 1-32.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. In: *Proceedings of EURALEX*, Lorient, France. 105-116.

Kilgarriff, A., Tugwell D. 2001. Word Sketch: Extraction and Display of Significant Collocations for Lexicography. In: *Proceedings of the 39th Meeting of the Association for Computational Linguistics, Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, Toulouse. 32-38.

Lee, L. 2004. “I’m sorry Dave, I’m afraid I can’t do that”: Linguistics, Statistics, and Natural Language Processing Circa 2001. In: *Computer Science: Reflections on the Field, Reflections from the Field.* The National Academic Press. 111-118.

Levin, B. 1993. *English Verb Classes and Alternations.* The University of Chicago Press

Sass, B. 2006a. Extracting Idiomatic Hungarian Verb Frames. In: *Salakoski T. Et al.* (szerk.) *Advances in Natural Language Processing.* LNCS, Vol. 4139. Berlin Heidelberg New York: Springer-Verlag 303-309.

Sass, B. 2006b. Igei vonzatkeretek az MNSZ tagmondataiban. In: *Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006),* Szeged. 15-21.

Sass, B. 2005. Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: *Alexin Z., Csendes D. (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2005),* Szeged. 257-264.

Schulte im Walde, S. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics* Vol. 32. No. 2. 159-194.

Teubert, W. 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics* Vol. 10. No. 1. 1-13.

Váradi, T. 2002. The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain. 385-389.