Vonzatkeretek a Magyar Nemzeti Szövegtárban

Sass Bálint

MTA Nyelvtudományi Intézet, Korpusznyelvészeti Osztály
1068 Budapest, Benczúr u. 33.
joker@nytud.hu

Kivonat: Jelen munkálat célja a Magyar Nemzeti Szövegtár vonzatkereteinek felismerése, az MNSZ vonzatkeret-gyakorisági szótárának előállítása a rendelkezésre álló vonzatkeret-táblázat felhasználásával. Manuális vizsgálatok alapján megállapítható, hogy a készülő korpuszfeldolgozó eszköz egyszerű nyelvtanokkal is képes a vonzatkeretek felismerésére, elkülönítésére, gyakorisági viszonyaik hozzávetőleges megállapítására. A gyakorisági szótár visszahathat a vonzatkeret-táblázat fejlesztésére illetve a szótárírásban és szintaktikai elemzők készítésekor is felhasználható.

1 Bevezetés

Jelen munkálat távlati célja a Magyar Nemzeti Szövegtárban (továbbiakban MNSZ)található igei vonzatkeretek felismerése, azonosítása és az MNSZ *vonzatkeret-gyakorisági szótárának* elkészítése.

A vonzatkeretek felismerése az első lépés lehet a gépi szövegmegértés felé. Általa a korpuszból részleges szemantikai információt nyerhetünk. Lehetőség nyílik a vonzatkeret-táblázat empirikus adatokra támaszkodó továbbfejlesztésére. Hasonlóan, empirikus gyakorisági adatokra támaszkodó szótárak jöhetnek létre: segítségével a szócikkek gyakorisági alapú megválogatásán túl lehetőség nyílik a vonzatkeretek megválogatására gyakori keretek felvétele, ritka keretek elhagyása által. Ezen kívül a gyakoriság a szócikkbeli jelentéssorrend megállapításának is egyik szempontja lehet.

A 2005 nyarán indult munkálat kezdeti eredményeiről számolok be.

2 Eszközök

Három nyelvi erőforrás kerül felhasználásra: (1) az MNSZ-ben ismeri fel az alább ismertetett (2) vonzatkereteket a projekt keretében elkészülő (3) korpuszfeldolgozó program.

2.1 A vonzatkeret-táblázat

A Nyelvtudományi Intézetben 2001 és 2004 között készült lexikai adatbázis [2] a magyar nyelv alapszókincsét alkotó szavak szintaktikai és alapvető szemantikai tulajdonságait kódolja szófajonkénti elosztásban. A kiindulópont az igei argumentumszerkezetek kódolása volt (az adatbázis a vonzatkeret-táblázaton kívül a főnevek és melléknevek szemantikai tulajdonságait kódoló táblázatokból áll). Igevonzatként csak az olyan összetevők szerepelnak, ahol az igével szintaktikailag vagy szemantikailag nem teljesen kompozicionális szerkezetet áll elő. Az ezt a követelményt nem teljesítő gyakori igei kontextusok nem szerepelnek a táblázatban. Az adatbázis kezdetben az MNSZ leggyakoribb 20000 szavából indult ki, több lépésben bővült, jelenleg a Szeged Korpusz összes vonzatkeretét is tartalmazza.

Jelen munka során az igei vonzatkeretek forrása ennek az adatbázisnak gazdag igei vonzatkeret-táblázata, mely 9000 ige 18000 vonzatkeretét tartalmazza. A feldolgozás alapja az eredetileg Excel táblázatos forma XML-re konvertált egységesített változata.

2.2 Szintaktikai elemzés és vonzatkeret-illesztés

A vonzatkeretek felismerésének menete két részre tagolódik: a részleges szintaktikai elemzést és annotációt követi a vonzatkeretek tényleges illesztése.

Mindkét lépést a projekt keretében elkészülő korpuszfeldolgozó eszköz végzi. Az eszköz a vonzatkeretek felismerésére készül, de a későbbiekben egyéb, általános célú feldolgozó modulokkal kényelmesen kiegészíthető. A nyelvtant alkotó szabályokban távlatilag, szükség szerint több rendelkezésre álló tagmondatra bontó, tulajdonnévfelismerő [3] illetve frázisfelismerő [6-7] eljárást és eszközt fel kívánok használni. Ezeket egyesítve, ezekre építve alakítom ki a szabályrendszert.

A feldolgozás az MNSZ formátumának megfelelő, morfoszintaktikailag a Morphologic *Humor* elemzőjével [4] elemzett, egyértelműsített korpuszból indul ki. A kidolgozott morfológiai reprezentáció részletekbe menő lekérdezéseket tesz lehetővé, az elemzési lépésekben kihasználhatjuk a magyar nyelv morfológiája adta lehetőségeket. Az eszköz implementálja a többszintű reguláris nyelvtan (*cascaded regular grammar*) technológiát [1]: a nyelvtanokat egymásra épülő, tokenek feletti reguláris kifejezésekből alakíthatjuk ki.

Néhány hasznos kiegészítő funkció:

A szabályok megfogalmazásakor pozíciót is meg lehet adni, hivatkozhatunk például a mondat első szavára.

Tagadást is használhatunk, ezáltal könnyen megjelölhetjük például a nem-alanyesetű névszókat.

Öröklődés: a jelenlegi egyszerű formában az összes szerkezet automatikusan az utolsó tokenjének tulajdonságait örökli.

A többszintű annotációs tagek segítségével például a szabály számát bele lehet kódolni a tagbe hibakeresés céljából (pl. `X:1` és `X:2`), ugyanakkor a továbbiakban egységesen lehet hivatkozni a tagre `X`-ként; vagy az `NP:pred` annotációra hivatkozhatok ebben a konkrét formában, de általánosságban `NP`-ként is. (A szinteket kettőspont választja el egymástól.)

A szükségtelenné vált annotációkat törölhetjük. Ha adott szabály elől „el akarunk fedni” egy szerkezetet, ideiglenes cimkével annotáljuk, amit a szabály alkalmazása után törlünk.

A keretek illesztése során a program egyenként megnézi, hogy a mondat és a vonzatkeret egyes szerkezetei megfelelnek-e egymásnak, ha a vonzatkeret összes elemének talál megfelelőt, akkor a keret illeszkedik. Több illeszkedő keret esetén a legspecifikusabb keretet választja. Az illesztést nem befolyásolja, hogy esetleg a szintaktikai elemzés nem tudott teljesen lefedő elemzést nyújtani, csak a szükséges vonzatok megléte számít. Igementes mondatban természetesen nem lehet illeszkedő keretet találni.

3 Jelenlegi állapot

A jelenlegi rendszer a felismerési folyamat valamennyi lépését magában foglalja, a legtöbbet többé-kevésbé egyszerűsített formában. Az elemzett szövegtől eljutunk a nyers vonzatkeret-gyakorisági szótárig.

A nyelvtan egyszerű tulajdonnév felismerője lényegében nagybetűs szavak sorozatait keresi meg, kiegészítve azzal, hogy a mondatkezdő nagybetűs szót legtöbb esetben nem tekinti igazi nagybetűs szónak. Erre épül az főnévi csoportot felismerő, valamint az alany, az igei állítmány, a tárgy és a határozók azonosítására szolgáló nyelvtan.

A tagmondatra bontás problémáját egyelőre kikerülve, a tesztkorpuszba rövid, írásjelet nem tartalmazó, így jó eséllyel egy tagmondatból álló mondatokat választottam ki az MNSZ-ből. A tesztkorpuszt egész pontosan az MNSZ összesen 131682 darab 9-szavas mondata alkotta. Az irodalmi és a hivatalos nyelvben kissé ritkábbak a 9-szavas mondatok (átlagban 1500 szóra jut egy), mint a korpusz többi részében (ott 1200 szóra jut egy), azért nagyjából egyenetesnek tekinthető a megjelenésük. A tesztkorpusz így az MNSZ „nyelvét” képviseli, ami a sajtó többsége miatt leginkább a sajtónyelvhez hasonlítható.

A szövegben fellelhető hiányosságokkal, hibákkal (ismeretlen szavak, rossz mondatszegmentálás, elírások, szószéttöredezés, ékezetmentes részek, stb.) nem foglalkoztam.

Első lépésben csak a legegyszerűbb kereteket dolgoztam fel. Az egy tagmondat – egy vonzatkeret munkahipotézis alapján a vonzatkeretek közül elhagytam azokat, melyekben vonzatként tagmondat szerepelt. A szemantikai jegyeket nem vettem figyelembe, ezek kezelése a névszói táblázatok részletes feldolgozását kívánta volna meg. Figyelmen kívül hagytam az ige szintaktikai jegyeit és a főnévi igenévi vonzatokat is. Így az általam használt egyszerűsített vonzatkeretek lényegében a következőképpen épültek fel: adott igetőhöz névszói alany, tárgy és vonzatok tartoznak és minden egység esetében meg lehet adni, hogy milyen szófajú legyen, milyen esetet kíván illetve, hogy konkrétan mely szóalak, vagy lemma képviselje az adott pozíciót (pl. *részt vesz*, vagy *semmibe vesz*). A honnan? és hová? kérdésre felelő lokatívuszi vonzatok a táblázatban egybevonva szerepelnek, ezeknek a kezelése a megfelelő esetekkel illetve névutós kifejezésekkel történik. A keretekben meglévő opcionalitást úgy kezeltem, hogy önálló, csak kötelező elemeket tartalmazó alkeretet hoztam létre. Az aktuálisan feldolgozandó kategóriába így 16300 vonzatkeret (a 18000 keret 90%-a) került be.

Az esetlegesen előforduló azonos kereteket összevontam eggyé. Az azonosság oka legtöbbször nyilván az volt, hogy épp olyan tulajdonságokat hagytam el, melyek a keretek közötti különbséget adták, ugyanakkor voltak ténylegesen duplikált sorok illetve az opcionalitás használatának következetlenségéből adódó esetek is.

4 Eredmények

A rendszer a tesztkorpuszon végzett manuális vizsgálatok alapján megfelelően képes felismerni az egyszerű vonzatkereteket, el tudja különíteni adott ige különböző vonzatkereteit. A program jelen változatából látszik, hogy egyszerű nyelvtannal, nagy mennyiségű, nem hibátlan szövegen is képesek lehetünk körülbelüli gyakorisági viszonyok megállapítására. Ehhez ugyanis nem szükséges az összes vonzatkeret pontos felismerése, csak az, hogy a program a felismerésben ne egyoldalúan hibázzon.

Elkészült egy mag program, aminek a továbbfejlesztésével létre lehet hozni egy olyan rendszert, amely képes előállítani a bevezetőben említett gyakorisági szótárat.

4.1 Példák

Néhány jó példa:

egybevet vmit vmivel:
"Az önellenőrzés során a dolgozó egybeveti munkáját a követelményekkel."

utasít vkit vmire:
"A Közgyűlés utasítja a Polgármestert a szükséges intézkedések megtételére."

részt vesz vmiben:
"A Pénztárfelügyelet képviselője a közgyűlésen tanácskozási joggal vesz részt."

A jól felismert vonzatkeretek mellett, számos tanulsággal szolgáló eset is előfordult. Fény derült olyan vonzatkeretekre, melyek esetében a táblázat hiányos vagy nem teljesen helyes. Az *őrizetbe vesz* vonzatkeretet például explicit módon nem tartalmazza a táblázat, a *kisüt* esetén pedig mindig megköveteli a tárgyat. Az alábbi mondatban így a program megtalálta az alanyt, az állítmányt és a két határozót, illeszkedő vonzatkeretet viszont – helyes működéssel – nem talált:

"Az ország nagy részén hosszabb-rövidebb időre kisüt a nap."

Hasonlóan a *demonstrál* igének is csak különféle vonzatos (vmit, vmi mellett, vmi ellen) formái szerepelnek, de a vonzatnélküli változat nem, pedig az a leggyakoribb.

„Nemet mondott a kétfős frakcióra a Parlament Ügyrendi Bizottsága"

A fenti mondatra illeszkedett a *mond vmit vmire* keret, de megfontolandó lenne a *nemet mond vmire* keret felvétele. Utóbbi nem szerepel a Magyar Értelmező Ké-

ziszótárban [5] (továbbiakban ÉKSz.), viszont az MNSZ-ben hétszer több az előfordulása, mint az ÉKSz-ben szereplő *rosszat mond vkire* keretnek.

”Fizetéskiegészítést kapnak az év végén az ügyeletes egészségügyi dolgozók.”

A fenti mondatra – helytelenül – illeszkedett a *kap vmit vmin* (pl. lopáson) keret. A problémát az okozza, hogy a szabad határozó esetragja egybeesik a szükséges vonzat esetragjával.

4.2 Esettanulmányok

Két olyan igét választottam ki, melynek húsznál többféle vonzatkerete van, az illeszkedéseket megvizsgáltam, az alábbiakban foglalom össze a tapasztalataimat.

A *vág* szótő 90 mondatban fordul elő. Helyesen ismerte fel a program a *pofát vág*, *vág vmibe* (arcába, szavába, témába) és *vág vmit* kereteket. Utóbbi esetén azonban jól elkülöníthető volt három altípus: legtöbbször a specifikusabb *vág vmit vmire* (szeletekre, darabokra, karikákra stb.) keret fordult elő, (ez a táblázatban nem szerepel, csak a még konkrétabb *zsebre vág*), ritkábban pedig az igemódosítós *fát vág*, *grimaszt vág* illetve a *vág vmit rajta/belőle* keret.

A *vág vmit vmibe* legtöbb esetben a *nagy fába vágta a fejszéjét* specifikus keretként jelent meg, a kissé kétes értelmű *vág vminek* pedig a *neki* szó helytelen névmási elemzése miatt a *nekivág vminek* kerettel bíró mondatokat ismerte fel.

A ritkább keretek sok esetben helytelenül szabad határozót találtak meg. A *vág vkit vmin* általános keret helyett hasznosabb lenne a *pofon vág*-ot kiegészíteni a *kupán vág* kerettel. Fokozottan igaz ez az extrém ritka keretekre: a *vág vmiben* (ti. hónaljban) – minden esetben helytelenül – szabad határozóra illeszkedett.

Az elváló igekötő miatt előkerültek a vág igekötős formái is. Nem szerepel a táblázatban a *levág vmit vmiből* (szeletet, darabot), illetve a *kettévág vmit* (kelbimbót, szelet, uborkát, országot). Utóbbi 4 előfordulással a vág szótövet tartalmazó mini részkorpuszban 4%-ot képvisel! Az *elvág vmit* pedig összevonja az *elvágja a torkát* illetve *elvág vmit vmitől* nagyon különböző jelentésű kereteket.

A *vesz* szótő (1011 mondat) konkrétan lemmát megadó kereteinek a gyakoriságát és az ÉKSz-beli jelentéssorrendet vetettem össze az 1. Táblázatban.

1. Táblázat: A *vesz* lemmát megadó keretei az ÉKSz-ben

vonzatkeret	***db***	***ÉKSz. jelentés***
részt vesz vmi(be)n	210	5. jelentés
tudomásul vesz	18	16. jelentés
fordulatot vesz	8	18. jelentés (sajtó)
semmibe vesz	2	*nincs benne!*
feleségül vesz	1	11. jelentés

Megfigyelhető, hogy a jelentések sorrendje nem teljesen felel meg a gyakoriságnak: bizonyos keretek/jelentések (*tudomásul vesz*, *fordulatot vesz*) hátrébb, bizonyosak (*feleségül vesz*) előrébb vannak sorolva. Egy szótár írásakor megfontolandó lehet, hogy a *részt vesz*, mely 210 előfordulásával az összes(!) vesz igéjű mondat 21%-át adja, a jelentések között valahol legelöl szerepeljen.

Megvizsgáltam még két olyan igét, melyek többféle névutóval előfordulhatnak: a *fut vmi elől* és *fut vmi után* ÉKSz-beli megjelenése megfelel a gyakoriságnak. A *vádat emel vki ellen vmi miatt* (5 előfordulás) keret az ÉKSz-ben nem szerepel.

4.3 A vonzatkeret-gyakorisági szótár első változata

A 131682 mondatos tesztkorpuszon 142 perc alatt futott le a vonzatkeret-illesztés. 111522 mondatban talált kerettel rendelkező igetővet a program. A többi mondatban számos oka lehetett annak, hogy nem sikerült kerettel rendelkező igét azonosítani: eleve nem volt ige; képzett ige szerepelt; az ige minden kerete kiesett az egyszerűsítés során; hibásan vonta össze az igekötőt a program, stb. Végül ezek közül 102184 mondatban (92%) azonosított be a rendszer keretet.

A 2. Táblázatban egy szemelvényt látható a vonzatkeret-gyakorisági szótár első változatából, mely még minden bizonnyal külön féle forrásokból eredő számos hibát tartalmaz.

2. Táblázat: A lemmát is megadó keretek gyakorisági listájának kezdete

#	*vonzatkeret*	*db*
1	részt vesz vmiben	124
2	részt vesz vmin	103
3	kérdést tesz fel	27
4	tudomásul vesz	23
5	győzelmet arat	16
6	szert tesz vmire	16
7	figyelmet fordít vmire	13
8	hatást gyakorol vmire	12
9	világra jön	9
10	letartóztatásba helyez vkit	9

5 Alkalmazás

Mint a fentiekben láttuk, a vonzatkeret-gyakorisági szótár alkalmas a vonzatkeret-táblázat továbbfejlesztésének támogatására. A gépi használatra szánt lexikai adatbázist "kipróbálva" tisztázódnak azok a pontok, ahol talán változtatni érdemes a táblázaton: cél lehet a vonzatkeretek megfelelő finomítása, specifikussá tétele, leginkább a ritka keretek esetén, és főleg azokban az esetekben, ahol nem azokra a mondatokra illeszkedett egy-egy keret, amelyekre a táblázat szerzői gondoltak (pl. *vág vmiben*). Ha a vonzatkeret-táblázat célja valamiféle gépi megértés, akkor egyrészt valamilyen formában jelentéseket kell rendelni az egyes vonzatkeretekhez, másrészt az egymástól határozottan eltérő jelentésű kereteket külön kell kódolni akkor is, ha köztük sok formai hasonlóság van.

Egy vonzatkeret gyakran egy szótárbeli jelentésnek felel meg. A gyakorisági szótár alapján lehetőség lesz változtatni szótárak (pl. az ÉKSz.) "jelentéskincsén" illetve a gyakoribb jelentéseket előrevéve a jelentések sorrendjén. A szótárakban általában az a gyakolat, hogy először az alapszó jelentései vannak részletesen kidolgozva, csak aztán következnek a kifejezések. Így fordulhat elő, hogy a *kezébe/nyakába vesz* előrébb szerepel az ÉKSz-ben, mint a nagyságrendekkel gyakoribb *részt vesz*.

Nem utolsó sorban a megbízható vonzatkeret-felismerő hatékony szintaktikai elemző elkészítéséhez járulhat hozzá.

6 Fejlesztési lehetőségek

Egyértelmű, hogy a jelen dolgozatban bemutatott állapot csak egy kezdeti lépcsőt jelent. Nagyban fejleszthető a vonzatkeret-azonosítás megbízhatósága jobb nyelvtanokkal, a szemantikai jegyek tekintetbe vételével, a névszótáblázatok feldolgozásával, integrálásával. Fontos feladat tagmondatra-bontó modul beépítése, mely képessé teszi a rendszert összetett mondatok vonzatkereteinek megtalálására is. A biztosan szabad határozónak minősülő mondatrészeket megfelelő eszközzel ki kell szűrni.

A program korrekt kiértékeléséhez nagy mennyiségű manuális munkára vagy elég nagy kézzel annotált korpuszra van szükség.

Bibliográfia

1. Abney, S.: Partial Parsing via Finite-State Cascades. In *Proceedings of the ESSLLI`96 Robust Parsing Workshop*, 1996, pp. 1-8.
2. Gábor K.: Lexikai adatbázis dokumantációja, MTA Nyelvtudományi Intézet belső dokumentuma, 2004.
3. Gábor K., Héja E., Mészáros Á., Sass B.: Nyílt tokenosztályok reprezentációjának technológiája. `http://www.nytud.hu/oszt/korpusz/resources/ikta_ner.doc`
4. Prószéky G., Tihanyi L.: Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany,* Budapest, 1996, pp. 149-158.
5. Pusztai, F. (szerk.): *Magyar Értelmező Kéziszótár.* Akadémiai Kiadó, 2003.

6. Váradi T.: Főnévi csoport annotálása a CLaRK rendszerben. In *Alexin Z., Csendes D. (szerk.): MSZNY2003*, Szeged, 2003, pp. 65-71.
7. Váradi T., Gábor K.: A magyar INTEX fejlesztésről. In *Alexin Z., Csendes D. (szerk.): MSZNY2004*, Szeged, 2004, pp. 3-10.