Egy egységesített magyar igei vonzatkerettár építése és felhasználása

Vadász Noémi[1,3], Kalivoda Ágnes[1,3], Indig Balázs[2,3]

[1]Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
{vadasz.noemi,kalivoda.agnes,indig.balazs}@itk.ppke.hu

Kivonat A cikkben egy egységesített magyar vonzatkeret-adatbázist ismertetünk, amelyet az elérhető magyar vonzatkeret-gyűjtemények összefűzésével és egységesítésével építettünk. A vonzatkerettár felépítésének, használatának, valamint az erőforrásokban talált szisztematikus hibák javításának ismertetése után bemutatunk egy vonzatkeret-egyértelműsítő eljárást, amely a vonzatkerettár egy részén alapul. Végül ismertetjük a modul teljesítményének kiértékelését, összehasonlítva az irodalomból ismert más módszerekkel.
Kulcsszavak: vonzatkeret-adatbázis, vonzatkeret-egyértelműsítés, korpusznyelvészet, mondatelemző

1. Bevezetés

A mondat feldolgozása során az ember számára triviális lépés az ige vonzatainak megtalálása, míg egy számítógépes nyelvi elemző ugyanezt a feladatot egy vonzatkeret-adatbázis segítségével képes végrehajtani. A cikk első felében egy ilyen adatbázist ismertetünk, a Manócskát[1], amelyet az elérhető magyar vonzatkeret-adatbázisok összefűzésével és egységesítésével készítettünk. A cikk második felében bemutatjuk, hogyan hasznosítottuk a Manócska infinitívuszi vonzatokat tartalmazó vonzatkereteit az AnaGramma [1] nyelvi elemző rendszer vonzatkeret-egyértelműsítő moduljában.

2. A Manócska

Az általunk épített vonzatkeret-adatbázis, a Manócska a jelenleg ingyenesen hozzáférhető magyar vonzatkerettárak összekapcsolásával és harmonizálásával készült, így a legbővebb szabadon elérhető, ilyen jellegű erőforrás magyar nyelvre. A felhasznált erőforrások harmonizálásakor törekedtünk a korántsem teljes és precíz, különböző felépítésű erőforrások gyengeségeinek javítására is. A következő fejezetben a Manócska építéséhez használt erőforrásokat mutatjuk be. Ezután ismertetjük az adatbázis felépítését és használatát, végül pedig sorra vesszük az erőforrásokból származó hibás bemenetek javítási módszereit is.

[1] https://github.com/ppke-nlpg/manocska

2.1. A felhasznált erőforrások

A Manócska vonzatkerettárába integrált erőforrások mindegyike korpuszból nyert adatokon alapul és automatikus módszerekkel készült, amelyet néhányuknál kézi ellenőrzés is követett. Sass Bálint a leggyakoribb magyar vonzatokat és szókapcsolatokat tartalmazó, nyomtatásban megjelent szótára [2] a Magyar Nemzeti Szövegtár első, központozás nélkül 187 millió tokent tartalmazó verziójából (MNSZ) [3] készült egy automatikus vonzatkeret-kinyerő módszerrel [4]. A papírszótár anyagát kézzel ellenőrizték, ugyanakkor a nyersanyagául szolgáló tételek manuálisan nem ellenőrzött, bővebb gyűjteménye [5] – amely 28 millió elemzett mondatot és félmillió igei szerkezetet tartalmaz – elérhető adatbázis formájában is.

A Mazsola igei argumentum böngésző platform[2] [6] vonzatkeret-adatbázisa [5] sokáig a magyar vonzatkeretek egyetlen elérhető statisztikai adatbázisa volt. A Mazsolában megjelenő tételekből generált vonzatkeretek és példamondatok képezik a papírszótár bővebb verzióját. Maga az adatbázis csak később vált elérhetővé, és a szótár nehezebben volt hozzáférhető a platformnál, ezért mindhárom forma (platform, szótár, adatbázis) röviden Mazsola néven is használatos, a szövegkörnyezet által meghatározott módon, egyértelműen behelyettesítve.

A Tádé vonzatkeret-gyakorisági listáját a Webkorpusz [7] „4%-ából" (amely központozással együtt 589 millió token) spektrális klaszterezéssel [8] nyerték ki [9]. A gyakorisági lista az opcionális vonzatok keresésének egyik melléktermé ke. Méretében az eddig bemutatott erőforrásoknál jóval nagyobb, de ez köszönhető annak is, hogy sokkal nagyobb osztályt vizsgáltak, amely magában foglalta az infinitívuszt vonzó igék vonzatkereteit is.

A felhasznált erőforrások közül érdemes a Mazsolát és a Tádét összevetni, hiszen különböző módszerrel és különböző elvek mentén készültek. A Mazsola készítésekor inkább a vonzatkeret-adatbázis pontossága lehetett a fő cél, tehát az, hogy ne tartalmazzon hibás vonzatkereteket. Ezzel szemben a Tádé készítői inkább a nagyobb fedésre törekedhettek. Ennek megfelelően a Mazsola alkalmasabbnak tűnik olyan feladatokra, amelyekhez lexikai erőforrásként szükség van egy vonzatkerettárra, ugyanakkor a mérete jóval kisebb. A 4. táblázatban látható, hogy a Tádé az igék száma szerint kicsit több mint háromszorosa a félmillió igei szerkezetet tartalmazó, legbővebb Mazsola adatbázisnak, a keretek számában viszont alig van különbség a Tádé javára. Egy másik nagyon fontos eltérés a két erőforrás között az, hogy míg a Mazsola egyáltalán nem tárolja az infinitívuszi vonzatokat, a Tádé igen – igaz, nem kezeli őket megfelelően (lásd a 2.4. fejezetet).

Két további, korpuszból kinyert listát is beépítettünk a Manócska adatbázisába. Az első az MNSZ 2.0.3 [10] 785 millió tokent (írásjelek nélkül) tartalmazó anyagából automatikusan kinyert gyakorisági lista[3] [11], amely 27 091 igekötő-ige párt tartalmaz. Az igekötő-ige lista készítésénél a legnagyobb pontosság volt a cél, így a lista manuálisan ellenőrzött és javított. Ugyanakkor az

[2] `http://corpus.nytud.hu/mazsola/`

[3] `https://github.com/kagnes/hungarian_verbal_complex`

igekötő-kapcsolódás produktivitása miatt a lista fedése soha nem lehet teljes. A második az infinitívuszi vonzattal rendelkező igék gyakorisági listája[4]. Ez az MNSZ 2.0.4 verzióján [10] készült, amely 1,04 milliárd tokent tartalmaz (írásjelek nélkül). Összesen 1 507 ige szerepel benne, az infinitívuszi vonzatuk gyakoriságával, típusával, a korpuszból vett egy-egy példamondattal, valamint a potenciális lexikális vonzatokra vonatkozó információval. Akárcsak az első listánál, itt is a pontosság volt az elsődleges szempont, ezért ez a lista is manuálisan ellenőrzött.

2.2. Az erőforrások egységesítése

Az erőforrások összefűzése és egységesítése során elvégeztünk néhány hibajavító eljárást is, amelyek általunk definiált szabályok automatikus alkalmazását jelentik. Ezek a legegyszerűbb, triviális javításoktól a komplexebb megoldásokig terjednek. A módosítások – egy kivétellel – a szótár bal oldalát, tehát az igét és a hozzá tartozó igekötőt érintik. A javításokat részletes ismertetésük után a 2. táblázat foglalja össze.

Triviális hibajavításnak nevezzük a helytelen igealakok kiszűrését a vonzatkerettárból. Feltételezzük, hogy ezek a tövesítő hibájából eredően kerülhettek be a felhasznált erőforrásokba. Kiszűrésükhöz minden ige esetében a HUMOR számítógépes morfológiát [12] használtuk. Ha a HUMOR az igéhez „[IGE]" elemzést (is) adott, akkor meghagytuk, ha nem, akkor töröltük a teljes keretet az adatbázisból.

Az erőforrások vizsgálatakor azt találtuk, hogy egy-egy ige gyakran olyan igekötővel szerepel egy keretben, amely valójában nem az övé. Az ilyen hibákat fontosnak tartottuk javítani, hiszen ismert, hogy az igekötőnek nagy szerepe van a vonzatkeret-egyértelműsítésben: ha igekötős az ige, akkor a lehetséges vonzatkeretek száma jelentősen csökkenhet [13]. A hibák kiküszöbölésére egy szabályalapú megoldást javaslunk. Az igekötős igék esetében – ahol az igekötő és az ige egy *függőleges vonal* (pipe, |) karakterrel vannak elválasztva – mind az igekötőt, mind az igét megvizsgáltuk a HUMOR segítségével. Ha sem az igekötő, sem az ige nem kapott megfelelő elemzést, akkor a két tokent összevonva megnéztük, hogy a HUMOR [IGE] elemzést adott-e vissza. Ha igen, akkor az adatbázisban is összevontuk a két tokent egyetlen, igekötő nélküli igévé.

Nem mindig egyértelmű, hogy igekötős-e az ige. A döntéseink során arra támaszkodtunk, hogy míg a semleges mondatokban általában egybeírjuk az igekötőt az igével, a nemsemleges mondatokban speciális szórendben szerepelnek. Két intuitív tesztet alkalmaztunk: (1) a finit igekötős ige tagadásakor a szórend többnyire: tagadószó, finit ige, igekötő, valamint (2) az igekötős ige *is*-sel történő módosítása esetén a szórend: igekötő, *is*, ige. Az 1. táblázat illusztrálja a két tesztet.

További nehézséget okoz, hogy a lexikális vonzatok, az igemódosítók és az igekötők kategóriái között nem húzódik éles határvonal, a három kategória kontinuumot képez. Sok esetben az egybeírás/különírás sem segít, hiszen a problémás szavak helyesírása a korpuszban sem következetes. Az egybeírást/különírást így

[4] `https://github.com/kagnes/infinitival_constructions`

	igekötő	nem igekötő
szó	***ki****folyik*	***fel****vételizik*
tagadás-teszt	*nem folyik* ***ki***	*nem* ***fel****vételizik*
„is”-teszt	***ki*** *is folyik*	-

1. táblázat. Két teszt annak eldöntésére, hogy igekötős-e az ige. Az igekötő-gyanús szavakat **félkövérrel** szedtük.

az „igekötőség” mértékének tekintettük: minél több esetben írják egybe az igével, annál közelebb áll a kérdéses szó a prototipikus igekötőkhöz. Megszámoltuk az egybeírt és a különírt változatok gyakoriságát, és ha az egybeírás gyakorisága elérte az összes előfordulás 15%-át, akkor igekötőként kezeltük. Például az MNSZ 2.0.4-ben a *helyt* szó 1 396-szor szerepel különírva az ige előtt, és 6 548-szor egyeírva, így igekötőként kezeltük.

A 2. táblázat a triviális hibákat, a lehetséges okokat és a megoldásukat ismerteti. Az első három oszlopban az látható, hogy a HUMOR milyen szófaji címkét adott az igekötőre és az igére külön, valamint a két elemet egybevonva. A HUMOR által adott szófaji címkék alapján (1) töröltünk olyan kereteket, amelyekben az igealak semmilyen átalakítással nem volt elfogadható, (2) átalakítottuk a hibásnak talált igealakokat.

igekötő	HUMOR ige	egyben	példa	hiba	javítás
-	-	-	*nyug*	tövesítés	törlés
-	-	[FN]	*12-15-ért*	szófaji egyértelműsítés	törlés
[IK]	[IGE]	[IGE]	*ki/abál*	szegmentálás	*kiabál*
[IK]	-	[IGE]	*alá/z*	szegmentálás	*aláz*
-	[IGE]	-	*kölcsön/sikerül*	szegmentálás	*sikerül*
[IK]	[IGE]	-	*meg/akar*	szegmentálás	*akar*
[IK]	[IGE]	[IK][IGE]	*ráró*	szegmentálás	*rá/ró*
[IK]	[IGE]	-	*össze/tud*	szegmentálás	*tud*
[IK]	[IK][IGE]	-	*abba/bele/fut*	szegmentálás	*bele/fut*

2. táblázat. Triviális hibák és esetleges okaik. A hibák kombinálódhatnak is.

A fenti hibajavító eljárások a vonzatkerettár bejegyzéseinek bal oldalát javítják, tehát azt az oldalt, amely az igét és az esetleges igekötőjét tartalmazza. Ezen felül a bejegyzések jobb oldalán is elvégeztünk egy triviális javítást. A több azonos vonzatból csak egyet őriztünk meg, azt feltételezve, hogy ezek tévesen kerültek az erőforrásokba (valószínűleg mellérendelő szerkezetek vagy felsorolások következtében, a vonzatkeretek korpuszból történő kinyerése során). A

javítóeljárásokban megfogalmazott szabályok az adatbázissal együtt, forráskód formájában elérhetőek.

2.3. A Manócska felépítése

A Manócska egy `tsv` fájl formájában használható, amely az igék vonzatkereteit tartalmazza az egyes erőforrásokban talált gyakoriságukkal együtt. A 3. táblázat sorai a vonzatkerettár egy vonzatkeretéhez tartozó információt mutatják.

***abba/hagy* [ACC]**	a keret gyakorisága	összes keret összgyakorisága
Magyar Igei Szerkezetek (szótár)	3 622	6 117 057
Félmillió igei szerkezet	3 771	18 303 463
Tádé	397	8 867 536
igekötő-ige lista	11 528	13 715 465
infinitívusz lista	0	1 507
rang	0.00168342	

3. táblázat. A Manócska egy vonzatkerete. A vonzatkerettár oszlopait itt sorokba rendeztük.

Az első oszlop maga az ige (az igekötőjével együtt). A vonzatkeret második oszlopa (a vonzatokat tartalmazó oszlop) felépítése a következő: (1) ha az igének semmilyen vonzata nincs, azt egy „@” karakter jelöli, (2) a főnévi vonzatokat az esetragjuk vagy a névutójuk képviseli. A lexikális vonzatok az esetragon vagy a névutón kívül magát a lexikális tövet is tartalmazhatják. A felhasznált erőforrások mindegyike korpuszadatok alapján készült, így feltételezhetjük, hogy a viszonylag gyakori vonzatkeretek szerepelnek bennük[5].

A következő oszlopokban az egyes kereteknek a felhasznált erőforrásokban talált gyakorisága szerepel. A legutolsó oszlopban a rang, egy összesített gyakorisági szám látható, amely a kerethez az egyes erőforrásokra kiszámított normalizált gyakoriságok összege[6]. A rang alapján rendezett vonzatkeret-szótárban a legtöbb adatbázisban megtalálható és gyakori keretek kerülnek előre – a normalizálás miatt az adatbázis méretétől függetlenül – az egyes igékhez tartozó vonzatkeretek közül.

A Manócska adatbázisa az eredeti erőforrásokból az összefűző, egységesítő és hibajavító szkript futtatásával bármikor reprodukálható[7]. A reprodukálhatóság megőrzése fontos célunk, hiszen így biztosíthatjuk a további hibák javítása-

[5] A kutatás jelenlegi fázisában nem foglalkozunk a vonzatok kötelezőségével vagy opcionalitásával, illetve a szabad határozókról sem hozunk ítéletet.

[6] A normalizált gyakoriság a vonzatkeret gyakorisága elosztva az összes vonzatkeret gyakoriságának összegével.

[7] Jelenleg a reprodukáláshoz szükség van néhány kiinduló erőforrás beszerzésére – amelyeket megfelelő formátumban a megfelelő mappákba kell elhelyezni a működés-

nak lehetőségét. A Manócska reprodukálhatóságával kiemelkedik a felhasznált erőforrások közül.

2.4. Az infinitívuszi vonzatok a Manócskában

A vonzatkerettár építésekor fontosnak tartottuk az infinitívuszi vonzatok megfelelő kezelését, ezért a Manócska ezeket is tárolja. Amint a 2.1. fejezetben már utaltunk rá, a felhasznált erőforrások közül egyedül a Tádé tartalmazza az infinitívuszi vonzatokat, azonban ez az erőforrás sem kezeli őket megfelelően. Ha egy ige vonzatkeretében infinitívuszi vonzat van, akkor az infinitívuszi vonzatot a tövével együtt (tehát lexikális vonzatként) sorolja fel a mondatban található többi főnévi vonzattal együtt. Ez azt jelenti, hogy a keretben összekeveredhetnek a finit és infinit ige saját vonzatai. Ráadásul a Tádé leírásában[8] azt olvashatjuk, hogy ha a finit igének van infinitívuszi vonzata, akkor a vonzatkeret többi eleme az infinitívuszhoz tartozik. Ez azonban nem mindig igaz (pl. a *megtanít*, *megérez*, *megvár* tárgyesetű vonzatainál), valamint az összes infinitívuszi vonzat lexikális vonzatként történő kezelése sem megfelelő.

Az infinitívuszi vonzatok kezelésére a következő javaslatot tesszük: egyelőre nem próbáljuk meg szétválogatni az infinitívuszi vonzatot is tartalmazó keretek többi vonzatát aszerint, hogy a finit vagy az infinit igéhez tartoznak-e, hanem az infinitívuszt tartalmazó keretekből csak az infinitívuszi vonzatokat tároljuk (a többi vonzatot töröljük). Azt állítjuk, hogy az infinitívuszok töve az ő vonzat tulajdonságuk szempontjából irreleváns, ezért az infinitívuszi vonzatokhoz tartozó tövet sem tároljuk (az egyes tövek gyakoriságát természetesen megőrizzük, és a tövek gyakoriságát összegezzük az infinitívuszt vonzó ige keretének gyakoriságában). Feltételezzük, hogy minden igető, amely infinitívuszi vonzatként előfordul, szerepel finit igeként is a vonzatkeret-adatbázisban az ő többi saját vonzatával együtt. Például a Tádé keretei között talált *alá/bukik INF_ meg/nedvesíteni* keretet átalakítjuk *alá/bukik INF* keretté, ami azt jelenti, hogy az *alábukik* igének egy bármely tövű infinitívuszi vonzata lehet (pl. *úszni*, *búvárkodni*, *gyöngyöt halászni*, *megnedvesíteni* stb.).

A Tádé infinitívuszi vonzatai mellett használtuk a már bemutatott gyakorisági listát is, amely az infinitívuszt vonzó igéket (és igekötőjüket) tartalmazza. A Manócskában az infinitívuszt tartalmazó keretek tehát két forrásból származnak, így a forrásoknak megfelelő gyakoriságuk is szerepel az adatbázisban.

Ezt az infinitívuszt vonzó ige-igekötő listát használtuk a 3. fejezetében ismertetett vonzatkeret-egyértelműsítő eljáráshoz.

2.5. A Manócska számokban

Az erőforrások eredeti méretét összevetettük a javítóeljárások alkalmazása után kapott méretükkel. Az eredményeket a 4. táblázat mutatja. Feltételezzük, hogy

hez –, mivel azok licencbeli korlátozások miatt nem szerepelhetnek a Manócska repozitóriumban.

[8] `https://hlt.bme.hu/en/resources/tade`

a 2.2. fejezetben ismertetett, a vonzatkerettár bal oldalán (az igéken) végzett szabályalapú javításaink nem eredményeztek hibás kereteket. A táblázaton látható, hogy az egyes erőforrások méretéből különböző százalékban vágott le a javítóeljárások alkalmazása. Ennek az az oka, hogy az erőforrásokat különböző korpuszokon és különböző technikákkal állították elő.

erőforrás	igék száma			keretek száma		
	eredeti	javított	(%)	eredeti	javított	(%)
Igei szerkezetek	2 226	2 185	1,84	6 266	6 203	1,00
Félmillió...	9 999	9 629	3,70	535 607	527 059	1,59
TÁDÉ	30 263	29 269	3,28	864 281	549 753	36,39
ige-igekötő	27 091	27 091	0,00			
ige-infinitívusz	1 507	1 507	0,00			
ÖSSZESEN	31 455	30 119	4,25	1 242 354	972 524	21,72

4. táblázat. A javítóeljárások alkalmazása előtti és utáni keretszámok az egyes erőforrásokhoz, valamint a különbség mértéke %-ban feltüntetve.

Az 5. táblázat azt mutatja, hogy az egyes erőforrásokban talált keretek hány százalékban voltak megfeleltethetőek egymásnak. Feltételezzük, hogy minél több erőforrásban szerepel egy-egy vonzatkeret, annál valószínűbb, hogy helyes keretről van szó.

típus	keretszám	erőforrások száma	arány (%)
összes keret	972 524	5	100,00
összes infinitívusz	2 772	2	0,29
csak infinitívusz	729	2	26,30
csak infinitívusz	2 043	1	73,7
összes nem infinitívusz	969 752	4	99,71
csak nem infinitívusz	1 734	4	0,18
csak nem infinitívusz	47 900	3	4,93
csak nem infinitívusz	377 123	2	38,89
csak nem infinitívusz	54 2995	1	56

5. táblázat. A keretek számának eloszlása aszerint, hogy hány erőforrásban szerepelnek. Az arányszámok esetében az infinitívuszi kereteket az összes infinitívuszos keret számához, a nem infinitívuszi kereteket az összes, infinitívuszt nem tartalmazó keret számához viszonyítottuk.

3. Az infinitívuszi vonzatok szerepe a vonzatkeret-egyértelműsítésben

A fejezet a [13] által ismertetett vonzatkeret-egyértelműsítő eljárás, a VFrame alkalmazását mutatja be az AnaGramma nyelvi elemző rendszer kereteiben. Az AnaGramma mint pszicholingvisztikailag motivált elemző az emberi mondatfeldolgozás kétfázisú működését modellálja [14], amelynek első fázisában történik meg a vonzatkeret-egyértelműsítés. Ezután az igei komplexum készen áll arra, hogy a mondatelemzés második fázisában kielégíthesse az ige vonzatigényét. A kétfázisú mondatelemzés első fázisát a balról jobbra és szavanként történő elemzésben egy +2 token méretű előretekintő elemzési ablak valósítja meg.

Az igék, igekötők és infinitívuszok korpuszban mérhető viselkedése alapján előállított VFrame algoritmus képes arra, hogy az ige szűk kontextusa ismeretében a felmerülő összes lehetséges vonzatkeret közül kizárja a nem aktuális kereteket [13]. Ebben a vonzatkeret-egyértelműsítő algoritmusban az igekötők és az infinitívuszi vonzatok segítik a nem aktuális vonzatkeretek kizárását. Az ismertetett eljárást az igekötők és az infinitívuszok viselkedésének alapos megfigyelésével és korpuszmérésekkel támasztottuk alá, ugyanakkor az eljárás kiértékelése a cikkben nem történt meg.

A VFrame eljárás működésének alapfeltétele egy szótár, amelyben az igékhez rendelt igekötők tárolódnak azzal az információval együtt, hogy az ige-igekötő pár vonzatkeretei között szerepel-e olyan, amely infinitívuszi vonzatot tartalmaz. Ez a szótár a Manócska felhasználásval készült, pontosabban annak az ige-igekötő és az ige-infinitívusz listájából. Az 1. ábra két bejegyzést mutat ebből a szótárból[9].

```
utál {?, el:X, ki:X, meg:X}
felejt {?, el:?, ki:X, le:X, ott:X, rajta:X}
```

1. ábra: Két bejegyzés a VFrame által használt szótárból. Az igekötőhöz rendelt *?* azt jelenti, hogy azzal az igekötővel együtt az igének lehet infinitívuszi vonzata, az *X* azt jelenti, hogy nem. Az igéhez rendelt *?* azt jelenti, hogy az igének igekötő nélkül lehet infinitívuszi vonzata.

Ennek a szótárnak a segítségével ki tudjuk értékelni az eljárás teljesítményét. A VFrame algoritmusának ismertetése után kiértékeljük az eljárást[10].

[9] Az 1. ábrán szereplő *megutál* igének elvileg lehet infinitívuszi vonzata, de az MNSZ 2.0.4 korpuszban mindössze egy példa volt erre: *már megutáltam folyton hasznos lenni*. Az ebből előállított gyakorisági lista ötös gyakoriságnál kezdődik, így néhány ritka eset nem került be az erőforrásba.

[10] Az implementációnk elérhetősége: `https://github.com/ppke-nlpg/vframe`

3.1. A VFrame algoritmusa

A mondatelemzés során az igék elemzésénél[11] elindul a VFrame eljárás, amely leírható a 2. ábrán látható automata formájában. Az eljárás három keresőből áll, amelyek sorban a következők: (1) infinitívuszkereső balra, (2) infinitívusz- vagy igekötőkereső az ablakban, majd (3) igekötőkereső balra. A három keresés célja az, hogy a lehető legjobban lerövidüljön az ige elemzésekor felmerülő lehetséges vonzatkeretek listája.

Az automatának két belépési pontja van. Az olyan igék esetében, amelyekhez a szótárban – bármilyen igekötővel vagy igekötő nélkül – van infinitívuszi vonzat, az első keresőeljárás indul el. Ekkor az elemzett igét megelőző mondatszakaszban esetlegesen talált infinitívusz – feltételezve hogy az ige vonzata lesz – az igéhez esetlegesen tartozó igekötők listáját megszorítja azokra az igekötőkre, amelyekkel párosulva az igének lehet infinitívuszi vonzata. A második keresőre már ezzel a megszorított igekötőlistával lép tovább. Ha nem volt infinitívusz az elemzett igét megelőző mondatrészben, akkor a második keresés következik.

A második keresésnél az igét követő néhány elem vizsgálata történik aszerint, hogy van-e közöttük igekötő vagy infinitívusz. Találat esetén a talált elemmel megszorításra kerül az igéhez tartozó ige-infinitívusz lista.

A harmadik keresés ismét az elemzett ige előtti mondatrészt vizsgálja, amelyben igekötőt keres. Ha korábban volt infinitívusz-találat, amely megszorította az igekötőlistát, akkor csak olyan találatot fogad el, amely szerepel ebben a megszorított listában.

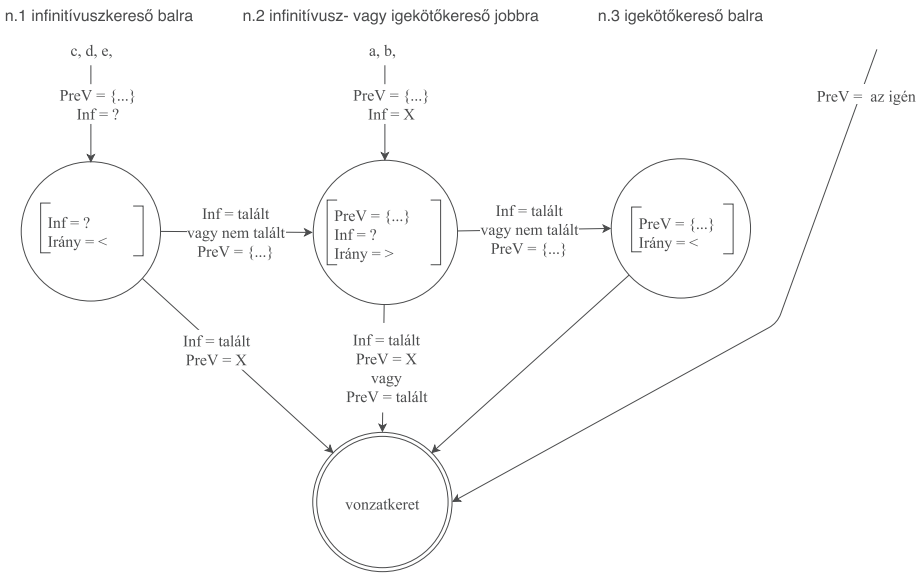
Mindhárom keresés után kiléphet az automatából abban az esetben, ha sikerült annyira megszorítani az igekötőlistát, amennyire csak lehetséges. Ehhez elsősorban a megfelelő igekötő megtalálása szükséges, amelyet az infinitívusz megjelenése segít. A megfelelő igekötő igéhez kapcsolása után (vagy annak tisztázása után, hogy nincs igekötő) a vonzatkeretek lehívása következik. Ezek között már csak olyanok szerepelnek, amelyeket az ige szűk környezete alapján nem tud az algoritmus kizárni.

Ha az elemzett igén eleve volt igekötő, akkor a VFrame eljárás kimarad, egyből a vonzatkeretek lehívása következik.

3.2. Kiértékelés

A VFrame teljesítményét 1 000 tesztmondaton mértük ki. A tesztmondatokat (egész pontosan a tagmondatokat, amelyekben finit ige szerepel) az MNSZ 2.0.4 szolgáltatta. A VFrame teljesítményét három jelenség kezelése teszi ki: 1) a finit ige és az igekötőjének összekapcsolása, 2) az infinitívusz és az igekötőjének összekapcsolása, valamint 3) a finit ige és az infinitívuszi vonzatának összekapcsolása. A kiértékelés során a finit igékre koncentráltunk, a többi igének (így a melléknévi és határozói igeneveknek) és az igekötőjének vagy infinitívuszi vonzatának az összekapcsolását kihagytuk a vizsgálatból.

[11] A VFrame jelenleg a finit igéket és az infinitívuszokat kezeli, a melléknévi és határozói igenevek kezelése még nem megoldott.



2. ábra: A *VFrame* keresőeljárás állapotainak véges állapotú automata reprezentációja.

A tesztmondatokat ennek megfelelően úgy válogattuk, hogy egy finit igét tartalmazzanak, ezen kívül a tagmondatban legyen legalább vagy egy igekötő, vagy egy infinitívusz. A finit ige és az infinitívusz is lehet igekötős, hiszen a VFrame ezt külön esetként kezeli. Az így leszűrt mondatok közül véletlenszerűen választott 1 000 darabot vettünk fel a teszthalmazba. A tesztmondatok, valamint az összetételükkel kapcsolatos részletes információ megtalálható a VFrame git repozitóriumban[12].

A mondatokhoz kézzel megjelöltük a bennük található ige-igekötő, infinitívusz-igekötő, ige-infinitívusz kapcsolatokat, amely referenciaadatként szolgált a kiértékeléshez. A kézi annotációt és a VFrame kimenetét automatikusan összevetettük, és a megegyező vagy különböző eredményeket a megfelelő kategóriákba soroltuk, amelyeket a 6. táblázat tartalmaz.

A kategóriák számosságát mindhárom feladatra külön megnéztük, így megvizsgálhatjuk a VFrame teljesítményét az ige-igekötő, az infinitívusz-igekötő, a finit ige/infinitívusz-igekötő, valamint az ige-infinitívusz összekapcsolására is, de a VFrame teljesítményére összességében is. A VFrame teljesítményét összevetettük két egyéb eljárásával is. Az eredményeket a 7. táblázat mutatja.

A VFrame teljesítményét összevetettük egy baseline eljárással, amely a finit igét és az infinitívuszt az igekötővel, valamint a finit igét az infinitívuszi vonzatával azok közelsége alapján kapcsolja össze. Az eljárást Recski Gábor módszere [15] alapján dolgoztuk ki úgy, hogy az minden igekötőhöz a hozzá legközelebb

[12] `https://github.com/ppke-nlpg/vframe`

kategória	finit ige/infinitívusz–igekötő	finit ige–infinitívusz
TP	van igekötő és megtalálta	van infinitívusz és megtalálta
TN	nincs igekötő és nem találta meg	nincs infinitívusz és nem találta meg
FP	rossz igekötőt talált	rossz infinitívuszt talált
FN	nem találta meg az igekötőt	nem találta meg az infinitívuszt

6. táblázat. Az egyes kategóriák, amelyek az igekötő-ige és az ige-infinitívusz összekapcsolásánál felmerülnek. **TP**: valós pozitív, **TN**: valós negatív, **FP**: álpozitív és **FN**: álnegatív

álló igét (finit igét vagy infinitívuszt) rendeli. Ez a baseline eljárás tehát nem támaszkodik arra az információra, hogy az igének lehet-e infinitívuszi vonzata, illetve hogy milyen igekötője lehet egyáltalán, csupán annyi megszorítással él, hogy bizonyos finit igéknek nem keres igekötőt (ezek a segédige-szerű igék [16] alapján az *akar, bír, fog, kell, kezd, kíván, lehet, mer, óhajt, próbál, szabad, szándékozik, szeret, szokik, talál, tetszik* és a *tud* a létigével kiegészítve). A baseline eljárás az igéhez az infinitívuszt is annak közelsége alapján kapcsolja. Mind az igekötő, mind az infinitívusz összekapcsolásának feltétele, hogy egy tagmondatban szerepeljenek a finit igével, ez a feltétel a tesztmondatainkban mindig teljesül.

A baseline módszer mellett a `magyarlanc` függőségi elemzőjének [17] eredményével is összevetettük a VFrame teljesítményét. A függőségi elemzésben megnéztük, hogy hányszor egyezett meg a kézi annotációval az ige-igekötő és az ige-infinitívusz összekapcsolása. Ez az összekapcsolás gyakran amiatt volt hibás, hogy az elemző eleve rosszul állapította meg a finit igét (összesen 40 alkalommal).

		Fin–Ik	Inf–Ik	Fin/Inf–Ik	Fin–Inf	Összesen
pontosság	VFrame	**97,57**	**94,71**	**96,82**	**97,88**	**97,21**
	baseline	92,39	90,40	91,87	96,98	93,72
	`magyarlánc`	88,22	89,36	88,53	89,93	89,08
fedés	VFrame	96,30	**94,21**	**95,76**	98,34	96,70
	baseline	**96,49**	92,75	95,50	**99,05**	**96,80**
	`magyarlánc`	79,20	86,15	80,96	89,74	84,23
F-mérték	VFrame	**96,93**	**94,46**	**96,29**	**98,11**	**96,95**
	baseline	94,40	91,56	93,65	98,00	95,24
	`magyarlánc`	83,47	87,73	84,58	89,83	86,59

7. táblázat. A különböző alfeladatok és a VFrame teljesítményének kiértékelése összevetve egy baseline eljárással és a `magyarlanc` függőségi elemző eredményével. **Vastag betűvel** szedtük a legmagasabb értékeket.

Az eredmények azt mutatják, hogy a VFrame és a baseline módszer teljesítménye között csupán kis különbség van. A baseline módszer néhány alfeladatban a fedés szempontjából valamivel jobban teljesített, míg a VFrame minden alfeladatban és összesítve is a pontosságban volt jobb. A baseline módszer a tesztmondatok korpuszból származó szófaji címkéjére támaszkodik, így előfordult, hogy nem helyesen állapította meg a finit igét (pl. a *vagy* kötőszót vette finit igének). Ebből a hibából összesen 4 darab fordult elő. A baseline módszerhez képest a VFrame a felhasznált szótárnak köszönhetően tudott jobban teljesíteni, amely segítségével kizárhatóak a helytelen igekötő-ige vagy infinitívusz-ige kapcsolatok.

A baseline módszer és a VFrame esetében a két hibatípusba (FP, FN) tartozó hibákat megvizsgálva kiderül, hogy a legtöbbjük az eleve hibás bemenetből fakad. Mindkét eljárás esetében a korpuszból vett tesztmondatokban a szófajegyértelműsítő hibát vétett (például az *elég* főnevet igekötős finit igének jelölte meg).

Egy másik, hibát okozó jelenség az, amikor a példamondatban az ige töve nem a megfelelő módon van feltüntetve – elsősorban az ikes igék esetében. Például a *mit lélegeznek ki a falak* tesztmondatban a *lélegeznek* ige töve a korpuszban a következő formában jelenik meg: *lélegezik/lélegzik*. Az ige-igekötő-infinitívusz listában azonban ez a tő nem szerepel (hiába van felsorolva a *ki* igekötő a *lélegez* tőhöz a szótárban). A VFrame esetében más, relevánsabb hibatípust nem találtunk, tehát a hibás eredmény lényegében a hibás bemenetből adódik.

A `magyarlanc` eredménye mind a baseline módszerrel, mind pedig a VFrame módszerrel szemben alulmaradt. A hibák számát jelentősen növelte, hogy a másik két módszerhez képest jóval többször rontotta el a finit ige megtalálását.

Mindent összevetve a VFrame előnye elsősorban abban áll – a legmagasabb pontosság és F-mérték mellett –, hogy a balról jobbra és szavanként történő feldolgozás miatt beépíthető az AnaGramma elemzőbe.

4. Összegzés

A cikk első felében a Manócskát ismertettük, egy egységesített vonzatkerettárat, amelyet az eddig elérhető erőforrások felhasználásával és javításával készítettünk. A vonzatkerettár a legbővebb szabadon elérhető, ilyen jellegű erőforrás magyar nyelvre. A cikk második felében az infinitívuszi vonzatok vonzatkeret-egyértelműsítő szerepe mellett érveltünk, majd bemutattunk egy eljárást, amely az infinitívuszi vonzatok és az igekötők segítségével oldja meg a vonzatkeret-egyértelműsítés feladatát. A VFrame eljárás kiértékelése azt eredményezte, hogy mind a finit ige-igekötő, infinitívusz-igekötő és infinitívusz-finit ige összekapcsolását igen magas pontossággal és fedéssel hajtja végre. Az eljárás az AnaGramma nyelvi elemzőrendszer keretébe illeszkedik.

Hivatkozások

1. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott Nyelvtudomány **15**(1-2) (2015) 29–44

2. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára. Tinta Könyvkiadó, Budapest (2010)
3. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002) European Language Resources Association, Paris, ELRA (2002) 385–389
4. Sass, B.: Igei szerkezetek gyakorisági szótára – Egy automatikus lexikai kinyerő eljárás és alkalmazása (2011) Doktori disszertáció. Pázmány Péter Katolikus Egyetem ITK.
5. Sass, B.: 28 millió szintaktikailag elemzett mondat és 5 00000 igei szerkezet. In Tanács, A., Varga, V., Vincze, V., eds.: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015), Szeged, SZTE TTIK Informatikai Tanszékcsoport (2015) 399–403
6. Sass, B.: „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi, T., ed.: Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, Budapest, MTA Nyelvtudományi Intézet (2009) 117–129 `http://corpus.nytud.hu/mazsola`.
7. Halácsy, P., Kornai, A., László, N., András, R., Szakadát, I., Viktor, T.: Creating open language resources for Hungarian. In N, C., ed.: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). (2004) 203–210
8. Brew, C., Schulte im Walde, S.: Spectral clustering for german verbs. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 117–124
9. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting Optional Arguments of Verbs. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
10. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Calzolari, N., et al., eds.: Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26-31, 2014, Reykjavik, Iceland, ELRA (2014) 1719–1723
11. Kalivoda, Á.: A magyar igei komplexumok vizsgálata (2016) Mesterszakos szakdolgozat. PPKE-BTK. `https://github.com/kagnes/hungarian_verbal_complex`.
12. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
13. Vadász, N., Kalivoda, Á., Indig, B.: Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In Vincze, V., ed.: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport (2017) 3–12
14. Frazier, L., Fodor, J.D.: The Sausage Machine: A New Two-Stage Parsing Model. Cognition **6**(4) (1978) 291–325
15. Recski, G.: A sekély mondattani elemzés további lépései. In Tanács, A., Vincze, V., eds.: VIII. Magyar Számítógépes Nyelvészeti Konferencia. (2011) 113–118
16. Kálmán C., Gy., Kálmán, L., Nádasdy, Á., Prószéky, G.: A magyar segédigék rendszere. Általános nyelvészeti tanulmányok – Tanulmányok a magyar mondattan köréből **17**(1) (1989) 49–103
17. Zsibrita, J., Vincze, V., Farkas, R.: MAGYARLANC: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP 2013. Hissar, Bulgária, 2013.09.08-2013.09.13. (2013) 763–771