Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével

Vadász Noémi[1,3], Kalivoda Ágnes[1], 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 mondat elemei közötti viszonyok feltárásakor a vonzatkeretek mielőbbi egyértelműsítésére törekszünk. A vonzatkeret-egyértelműsítésben az igekötők és az infinitívuszi vonzatok kiemelt szerepet kapnak. Ha ezek az elemek az igétől jobbra helyezkednek el, az a balról jobbra történő elemzés során problémát jelenthet. Cikkünkben az ige–igekötő, ige–infinitívuszi vonzat távolságok korpuszokon történő kimérése után egy megoldást kínálunk erre a problémára. Az általunk készített *VFrame* keresőeljárás az igékhez esetlegesen tartozó igekötőket és infinitívuszi vonzatokat egységesen, egy eljárásban kezeli, így a vonzatkeret-egyértelműsítés azonnal elvégezhető.

1. Bevezetés

Az AnaGramma elemzőrendszer [1] egy pszicholingvisztikai indíttatású nyelvelemző modell [2], amely az emberi mondatmegértés mintájára balról jobbra és szavanként elemez. A rendszer működésének három alapvető eleme a *kereslet-kínálat* elvű keretrendszer [3], a *tározó* és az *ablak* (a részleteket lásd 1.1). Az elemző kimenete egy függőségi éleket tartalmazó gráf.

A kereslet-kínálat elvű keretrendszer azt jelenti, hogy az elemző látóterébe kerülő tokenek más – korábbi vagy éppen később érkező – elemek számára kínálatként szerepelhetnek, ugyanakkor saját keresleteik is lehetnek (pl. az ige keresi vonzatait). A keresleteket *keresőeljárások* valósítják meg, amelyek különböző megszorításokat tartalmaznak pl. a keresés irányára. Emellett olyan egyéb információkat is hordozhatnak, amelyek a mondat elemzése során később lehetnek szükségesek. A keresőeljárások egyik legfontosabb funkciója a *találati függvény*, amely utasítja az elemzőt, hogy mit tegyen, ha a keresés a keresett elem megtalálásával vagy sikertelenül fejeződött be. Az eljárás az eredménye alapján (talált/nem talált) az elemzés következő lépésében újabb keresőeljárást indíthat.

Az elemzés során az adott pillanatig még be nem kötött éleket, tehát az eddigi kínálatokat (az éppen elemzett tokent megelőző tokeneket, azok morfológiai elemzését és az esetleges kész szerkezeteket), a keresleteket, valamint a kész részszerkezeteket egy rövidtávú munkamemória [4], az ún. tározó tartalmazza. Az

elemzés bármely pillanatában lekérdezhető a tározó tartalma a keresletek számára. A kereslet és kínálat találkozásakor az elemek jegyeinek unifikálódásával jön létre közöttük kapcsolat.

Az elemzés során a függőségi élek bekötéséhez a tározón kívül rendelkezésre áll egy, az aktuálisan elemzett szótól jobbra tekintő, két token méretű elemzési ablak is. Ebben az ablakban az éppen elemzett token bizonyos keresleteit kielégítő kínálatok is előfordulhatnak, amelyek befolyásolhatják az elemzést.

1.1. Az ablak és a tározó

Az AnaGramma az emberi szövegfeldolgozást modellálja. Ehhez a *Sausage Machine* kétfázisú mondatfeldolgozását [5] veszi alapul. Az első fázis a *Preliminary Phrase Packager* (PPP), amely a szöveges bemenet szócsoportjaiból frázisokat „csomagol". A második fázisban (*Sentence Structure Supervisor*) a „csomagok" további nemterminális csomópontok hozzáadásával megkapják a szerepüket a mondatban[1].

A PPP fázis eredetileg egy kb. hat szó méretű ablakban dolgozik (angol szöveggel). Agglutináló jellege miatt a magyar nyelvre célszerűbb egy kisebb ablak alkalmazása, így a PPP fázist az AnaGramma keretében egy három token méretű elemzési ablak modellálja[2].

A Sausage Machine alkalmazása az AnaGramma modellben azt jelenti, hogy a balról jobbra szavanként történő feldolgozás egy lépésében az elemző az – egyelőre három token méretűnek meghatározott – ablak bal szélén lévő elemmel foglalkozik, elindítja az összes, az adott token morfológiai elemzése által indítandó folyamatot, megvizsgálja, hogy a token mint kínálat kielégíthet-e egy keresletet a tározóban. Az ablakban található további tokenek – az elsődleges morfológiai információjukkal együtt – módosíthatják az aktuális elemzési lépéseket (pl. jelentős szerepük van a szófaji egyértelműsítésben is).

1.2. Vonzatkeret-egyértelműsítés az ablakban

Amikor az elemző egy igei elemet talál – legyen az finit vagy infinit ige, melléknévi vagy határozói igenév, amelynek vonzatkerete van és lehet igekötője –, többféle lehetséges vonzatkeret is felmerülhet. Az elemzés során a vonzatkeret-egyértelműsítésnek minél előbb meg kell történnie ahhoz, hogy a megfelelő keresőeljárások elindulhassanak. Az olyan, saját vonzatkerettel rendelkező igei elemek esetében, amelyeknek lehet igekötőjük, a balról jobbra történő elemzés

[1] Az elemzési ablak fontosságát jól mutatja, hogy nehezebben dolgozzuk fel a hírcsatornákon a képernyő alján végigfutó hírszalagot, ahol nem tudunk előretekinteni (lásd: http://users.itk.ppke.hu/∼yanzigy/olvaso/).

[2] Az ablaknak flexibilisnek is kell lennie, amely azt jelenti, hogy az elemző az elemzési ablakban „átugorja" az aktuális állapota számára nem releváns elemeket. Ezek jellemzően olyan rövid, nem tartalmas funkciószavak, amelyek nem játszanak szerepet a PPP fázisban. Az ablak megfelelő méretének és a flexibilitásának alátámasztásához szemmozgáskövetővel végzett kísérletek eredményei szükségesek, ám magyarra – eddig – ilyen megközelítésű kísérletről nem tudunk.

során felmerül az igei elemtől jobbra elhelyezkedő igekötő és infinitívuszi vonzat problémája. Ekkor az igekötő és az infinitívuszi vonzat mint a vonzatkeret-egyértelműsítésben fontos szerepet játszó elem „később" kerül be az elemzés folyamatába. Az ablak és a tározó, valamint a keresőeljárások segítségével az igék és az esetlegesen hozzájuk tartozó igekötők és infinitívuszi vonzatok egységesen kezelhetők.

1.3. A korpuszok

Méréseinket három különböző korpuszon végeztük. Az általunk készített InfoRádió korpusz rövid politikai és gazdasági híreket tartalmaz az InfoRádió hírportálról. A rövid hírek egy címből és egy két-három mondatos hírből állnak. A szerkesztett szöveg egyfajta „ideális" bemenetként szolgál az elemzőmodellünk számára. A korpusz 54996 hírből, 135587 mondatból és 1953419 tokenből áll.

A Magyar Nemzeti Szövegtár (MNSZ2) [6] v.2.0.3 verzióján is végeztünk méréseket, amely 785 millió tokent tartalmaz (írásjelek nélkül). A korpusz sokféle forrásból épül fel, közöttük fórumhozzászólásokból és beszélgetések leirataiból. Az MNSZ2 kontrasztot képez az InfoRádió korpusszal, hiszen ez utóbbi egy műfajból származó, szerkesztett szöveget tartalmaz.

A Pázmány Korpusz [7] 1,2 milliárd tokenből áll, amit több, mint 30000 weboldalról gyűjtöttek. Megkülönbözteti a főkorpusz (szerkesztett) és a kommentkorpusz (szerkesztetlen) szövegrészeket, amelyeket mi egyben kezeltünk.

2. Korpuszmérések

Két korpuszelemzést végeztünk. Az egyikkel a finit ige és a tőle jobbra álló igekötő lehetséges távolságát mértük ki, a másikkal pedig az infinitívusznak és jobbra kihelyezett igekötőjének a távolságát. Módszerünk lényege egy pozíció szerinti összehasonlítás, amelyben 0 pozíciónak az igét tekintjük, és ehhez képest határozzuk meg az igekötő helyét (+1 pozíció tehát pl. *látta meg*).

Mindhárom korpusz tartalmaz hibás annotációkat, emiatt sok rossz találatot is kaptunk. Ezeknek az automatikus szűrésére egy több mint 27 ezer igekötős igelemmát tartalmazó (manuálisan ellenőrzött) listát [8] használtunk fel. Ezzel a listával átszűrtünk minden korpuszbéli adatot, és csak azt az igekötő–ige párt fogadtuk el, amely a lista alapján létező kombináció. Ezzel állapítottuk meg azt is, hogy a gyűjtött anyagban az adott igekötő a finit vagy infinit igéhez tartozike (esetleg tartozhat-e elvben mindkettőhöz). Így vesztettünk néhány egyébként releváns találatot olyan neologizmusok esetében, amelyek nem szerepeltek a listában, de a módszerrel a hibás találatok jelentős részét hatékonyan, automatikusan ki lehetett szűrni.

2.1. Finit igék, infinitívuszok és igekötőik

Az igekötő mellett az infinitívuszi vonzatnak is van vonzategyértelműsítő szerepe, és mint ilyen, megerősítheti vagy kizárhatja bizonyos igekötők főigéhez való

kapcsolhatóságát. Öt igeosztályba sorolhatjuk az igéket aszerint, hogy igekötő nélküli, illetve igekötős vonzatkeretükben szerepelhet-e infinitívuszi vonzat. Az 1. táblázat foglalja össze az öt igeosztály tulajdonságait.

	igeosztály	példa		
	PreV vagy INF vonzat	tő	PreV	INF
(a)	nincs PreV, nincs INF	*villog*	X	X
(b)	nincs INF	*esik*	el, le...	X
(c)	nincs PreV	*kell*	X	?
(d)	PreV és INF kölcsönösen kizárják egymást	*tud*	le, meg...	?
(e)	INF bizonyos PreV-vel	*megy*	ki, el...	?

1. táblázat. Öt igeosztály az igéknek az igekötővel és az infinitívuszi vonzattal való kombinálhatósága alapján (X: nincs elfogadott infinitívuszi vonzat vagy igekötő az igével kombinálva, ?: az igének lehet, hogy van infinitívuszi vonzata)

A balról jobbra történő elemzés során az ige elemzésének pillanatában az ideális az, ha rendelkezésünkre áll minden vonzatkeret-egyértelműsítő információ. Megvizsgáltunk minden lehetséges kombinációt, amely az igekötő (PreV), finit ige (FIN) és az infinitívusz (INF) egymáshoz viszonyított sorrendjéből jön létre. Az eredményeket a 2. táblázat foglalja össze.

PreV – **FIN** – INF	**meg** sem **próbálták** csökkenteni
PreV – FIN – **INF**	**le** is akartam **fényképezni**
FIN – **PreV** – INF	**szűnjön meg** létezni
FIN – **PreV** – **INF**	sikerült két példányt **el** is **ejtenie**
INF – **FIN** – **PreV**	csodálni **járok vissza**
INF – FIN – **PreV**	**rohannia** kell **vissza**
FIN – INF – **PreV**	-
FIN – **INF** – **PreV**	kellett egészben új állami rendet [...] **építeni fel**
INF – **PreV** – **FIN**	kártyázni **le** ne **ülj**
INF – **PreV** – FIN	**feledni el** nem tudlak
PreV – INF – **FIN**	-
PreV – **INF** – FIN	**el** nem **utasítani** kegyeskedjék

2. táblázat. A finit ige (FIN), az infinitívuszi vonzat (INF) és valamelyik igekötőjének (PreV) egymáshoz viszonyított lehetséges sorrendjei példákkal (vastag betűvel az összetartozó párokat jelöltük)

A leggyakoribb szerkezet az olyan PreV–FIN–INF, amelynél az igekötő az infinitívuszhoz tartozik. Szintén gyakori a FIN–PreV–INF sorrend, ekkor az igekötő legtöbbször a finit igéhez tartozik. Ez a szórend jellemzi a non-neutrális (ezen belül főként a felszólító, tagadó) mondatokat. Az INF–FIN–PreV szerkezetnél az a tendencia figyelhető meg, hogy maga az infinitívusz áll fókuszpozícióban, és ez mozdítja el az igekötőt a preverbális pozícióból. A további három lehetséges

kombinációra is találtunk példát a korpusz mondatai között (lásd a 2. táblázat utolsó három sora), ezek a szerkezetek azonban ritkák.

2.2. A finit ige és a tőle jobbra elhelyezkedő igekötő

Bár a finit ige utáni mondatszakaszban a fő összetevők sorrendje szabad [9], az MNSZ2-n végzett mérések [8] azt mutatták ki, hogy a posztverbális igekötők az esetek 99%-ában +1 vagy +2 pozíciót foglalnak el. Az InfoRádió korpuszban ez az érték 100%, vagyis nem található benne olyan példa, ahol egynél több szó áll a finit ige és annak posztverbális igekötője között. A különbség azzal magyarázható, hogy az InfoRádió korpusz hivatalos stílusú, szerkesztett szövegeket tartalmaz. A mérések eredményéhez lásd a 3. táblázatot.

FIN	+1	+2	+3	+4	+5	+6	+7
MNSZ2	7527308	163993	5126	1193	267	101	27
Inforádió	23552	220	-	-	-	-	-
MNSZ2%	97,78%	2,13%	0,0666%	0,015%	0,003%	0,001%	3,5e-4%
Inforádió%	99,999%	0,001%	-	-	-	-	-

3. táblázat. A finit ige és a tőle jobbra elhelyezett igekötőjének távolsága – szerkesztett szövegekben 99,9%-ban közvetlenül az ige után, szerkesztetlen szövegekben 99%-ban maximum két token távolságra helyezkedik el az igekötő

Az eredmények tehát azt mutatják, hogy az igekötő nagyon ritkán kerül 2 tokennél távolabbra jobbra az igéjétől. Két extrém példa az MNSZ2 korpuszból:

(1) Azért **mentem** egy kicsit a pop zene fele **el**, mert szeretem a nívós, könnyed jó popzenét.

(2) 27 gyereket **vitt** egy feltehetően részeg buszsofőr Szentesen még csütörtökön egy sportrendezvény után **vissza** az iskolába.

A posztverbális igekötők pozíció szerinti eloszlásában az a tendencia fedezhető fel, hogy legtávolabb a testes igekötők kerülhetnek, amelyek határozószóként is funkcionálnak (pl. *haza*, *vissza*). Ezek jellemzően nem befolyásolják az ige vonzatkeretét. A legkevésbé eltávolodó igekötők grammatikalizált, rövid, prototipikus igekötők, amelyek eloszlásukat tekintve nagyon hasonló százalékokat produkálnak (lásd a 4. táblázatot)[3].

A magyar helyesírás szerint egy igekötő nem előzheti meg közvetlenül az igét, amelyhez tartozik – ekkor egy szót alkot az igével –, ezért ezt a pozíció nem része a kiértékelésnek. Az igét közvetlenül megelőző pozícióban szereplő igekötő vagy egy másik igei elem (pl. az ige infinitívuszi vonzatának) igekötője, vagy elírás eredménye (ekkor valóban az igéé).

[3] A preverbális igekötők pozíció szerinti eloszlását lásd [8]

	-2	0	+1	+2	+3
meg, ki, be, le, fel, föl, el, át, rá	0,49%	58,5%	40%	1%	0,01%

4. táblázat. Néhány gyakori igekötő távolsága a finit igétől – az igekötők 98,5%-a az igén vagy közvetlenül utána áll, csak 1,01% távolodik el jobban (MNSZ2)

2.3. Az infinitívusz és a tőle jobbra elhelyezkedő igekötője

Az infinitívusszal kapcsolatos méréseinket a Pázmány Korpuszon [7] végeztük. Mivel ez web-alapú korpusz, még az MNSZ2-nél is nagyobb arányban tartalmaz szerkesztetlen szövegeket (a kommentkorpusz mérete 2 millió token). Emiatt várható, hogy az infinit ige és a hozzá tartozó igekötő gyakran több szónyi távolságban álljanak egymástól. Az eredményeink mégis azt mutatják, hogy igekötő az esetek 86%-ában közvetlenül az infinit igealak után áll (lásd az 5. táblázatot).

INF [...] IK	db.	%
össz.	717	
+1	619	86,3
+2	52	7,3
+3	35	4,9
>+3	11	1,5

5. táblázat. Az infinitívusz és a tőle jobbra elhelyezkedő igekötőjének távolsága – 93,6%-ban maximum két token van köztük

FIN [...] INF	db.	%
össz.	727562	
+1	652778	89,7
+2	47669	6,6
>+2	27115	3,7

6. táblázat. A finit ige és a tőle jobbra elhelyezkedő infinitívuszi vonzatának távolsága – 96,3%-ban maximum két token van köztük

A kiugróan gyakori +1 pozícióban még sok prototipikus igekötőt találunk, pl. *iparkodott ellentétet* ***mutatni ki***, *javasolt a lapokat lazán* ***helyezni el***. A +2 pozícióról elmondható, hogy az infinitívusz és annak igekötője között csak finit ige állhat, és – bár van példa prototipikus igekötőre, pl. *épp* ***foglalni*** *akartam* ***le*** *a buszt* – nagyobb arányban jelennek meg a testes igekötők (pl. *már* ***indulni*** *akartam* ***vissza***). A nagyon ritka +3 pozícióban testes igekötők állnak, pl. *de már* ***jönni*** *kellett sajnos* ***haza***. A +4 és +5 pozícióra mindössze 15 példát találtunk, ez statisztikailag irreleváns mennyiség. Az itt álló igekötők nem befolyásolják az ige vonzatkeretét (csak az ige által kifejezett mozgás irányát módosítják), pl. ***vinni*** *kell a kamerát* ***el***, ***menekülni*** *akartak a városon keresztül* ***vissza***.

2.4. A finit ige és a tőle jobbra elhelyezkedő infinitívuszi vonzata

Kimértük azt is, hogy az infinitívuszi vonzat mint vonzatkeret-egyértelműsítő elem milyen távol állhat a főigétől. A 6. táblázatban látható, hogy az esetek 89%-ában az infinitívusz közvetlenül a finit ige után áll, 6,5%-ban egy szót enged

maga elé. Az eredmény alapján a legtöbb esetben a főige elemzési ablakába esik az infinitívuszi vonzat, ezzel elősegítve a vonzatkeret-egyértelműsítést.

A fenti eredményeket összefoglalva elmondható, hogy ha az igekötő a finit igétől jobbra helyezkedik el, akkor a legtöbb esetben beleesik az ige 1.1. fejezetben említett elemzési ablakába, tehát az ige elemzésének pillanatában elérhető az elemző számára a vonzatkeret-egyértelműsítéshez. Ugyanerre az eredményre jutottunk az infinitívusztól jobbra elhelyezdedő igekötő és a finit igétől jobbra elhelyezkedő infinitívuszi vonzat esetében is.

3. VFrame

Eredményeink alátámasztják, hogy egy viszonylag kisméretű előretekintő elemzési ablak elegendő ahhoz, hogy a vonzatkeret-egyértelműsítés az igei elem elemzésekor megtörténhessen, hiszen – amint a 2. fejezet eredményei mutatták – az igekötő és az infinitívuszi vonzat az esetek legnagyobb részében elérhető az igei elem számára (a tározóban vagy az ablakban). A következőkben a méréseink alapján létrehozott *VFrame* keresőeljárást ismertetjük, amely az igekötők igei elemekhez kapcsolásával segít előhívni a mondatban előforduló finit és infinit igék megfelelő vonzatkeretét.

Az igei elemet megelőző összes, és az azt követő néhány token ismerete egyértelműsíti a vonzatkeretet az igekötő és az infinitívusz vonzat tekintetében, mely szükséges, de nem elégséges. A 2.1. fejezetben ismertetett öt igeosztály egységes kezeléséről a VFrame gondoskodik, amely minden igei elem összes igekötő–infinitívuszi vonzat kombinációját kezeli (azokat az eseteket is, amikor nincs igekötő és/vagy infinitívuszi vonzat). A VFrame szerkezetét az 1. ábra mutatja.

$$\textsc{VFrame}\begin{bmatrix}\textsc{Irány} = > \mid < \\ \textsc{Igekötő} = \textit{lehetséges igekötők halmaza} \mid X \mid \textit{talált token} \\ \textsc{Infinitívusz} = ? \mid X \mid \textit{talált} \\ \textsc{Találati függvény} = \textit{találatkor vagy a sikertelen keresés végén fut le} \\ \textsc{Egyéb}\begin{bmatrix}\textsc{Tő} = \textit{az ige töve} \\ \textsc{megszorítási függvény} = \textit{a találatok megszorítási szabályai}\end{bmatrix}\end{bmatrix}$$

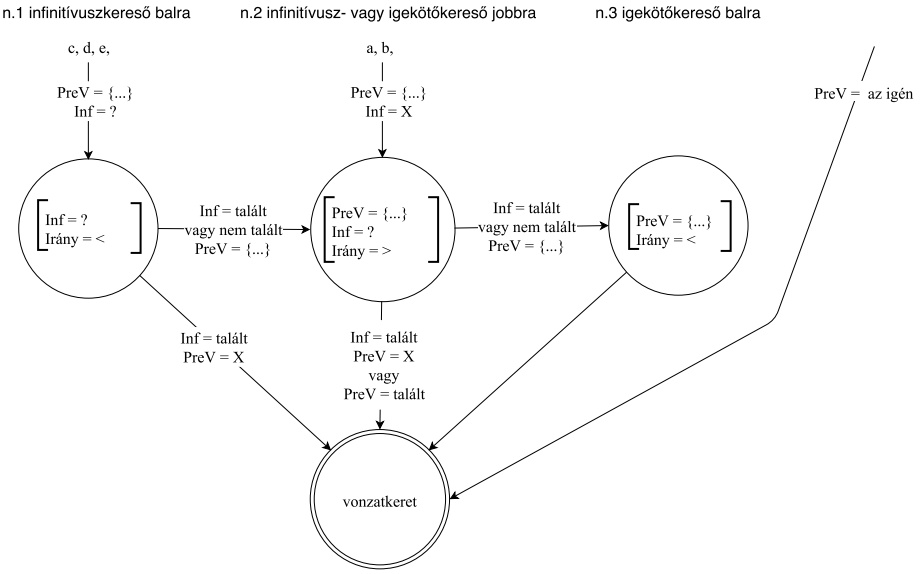
1. ábra. A *VFrame* keresőeljárás architektúrája

A VFrame **Irány** jegye a keresés aktuális irányát mutatja – balra a tározóban vagy jobbra az ablakban. Az **Igekötő** jegy az igei elemmel kompatibilis *összes lehetséges igekötő halmazát* tartalmazza (tekintet nélkül arra, hogy az adott igekötő kizárja-e az infinitívuszi vonzat meglétét), vagy X-et (ha az adott igei elemnek semmilyen igekötője nem lehet), vagy az igei elem az aktuális mondatban már *megtalált igekötőjét*[4]. Az **Infinitívusz** jegy jelzi, hogy az igei elemnek lehet-e infinitívuszi vonzata ($?$) vagy sem (X). Ha az elemző talál infinitívuszi vonzatot,

[4] Abban az esetben, ha az igekötő az igén van, az elemző kihagyja a VFrame keresőeljárást, és elindítja a vonzatok keresését.

akkor azt az Infinitívusz jegy *Talált* állása jelzi. A **Találati függvény** tartalmazza azt a függvényt, amelyet az elemző egy – a VFrame szempontjából fontos – elem megtalálásakor, vagy annak hiányában meghív. A Találati függvény kezeli továbbá a különböző keresőeljárások állapotai közötti átmeneteket is (lást a 2. ábrát). A **Megszorítási függvény** kezeli a találat megszorításait (pl. ha az INF és a PreV kölcsönösen kizárják egymást).

A VFrame aktuális tartalma alapján képes arra, hogy a megfelelő keresőeljárások elindításával egyértelműsítse az aktuális vonzatkeretet. A mondatban több olyan elem is lehet, amelynek lehet igekötője, sőt lehetséges igekötőik halmazában is lehet átfedés, a VFrame keresőeljárással az igék, igekötők és infinitívuszi vonzatok helyesen kapcsolhatók össze. A keresőeljárások sorozata egy véges állapotú automata segítségével írható le három valódi állapottal. A 2. ábra mutatja a folyamat lépéseit[5].



2. ábra. A *VFrame* keresőeljárás állapotainak véges állapotú automata reprezentációja, amely lefedi a 2.1. fejezetben ismertetett öt igeosztályt

4. Problémás esetek

Az igekötőkkel és infinitívuszokkal kapcsolatban egyéb problémák is felmerülhetnek, amelyek megnehezíthetik az igekötő–igei elem vagy infinitívuszi vonzat-igei elem összekapcsolását. Ebben a fejezetben a problémás esetek lajstromba vételével foglalkozunk. A VFrame keresőeljárás önmagában csak az elsőre – a több infinitívuszt tartalmazó mondatok problémájára – nyújt megoldást.

[5] A VFrame pontos működéséről és implementációjáról lásd [10].

4.1. Több infinitívusz

Az olyan példákban, ahol egynél több infinitívusz jelenik meg az igei komplexumban, az infinitívuszok jellemzően egymás mellett állnak, pl. *el kell kezdeni keringőzni tanulni*, *el fogod tudni dönteni*. Azonban arra is vannak példák, hogy az egyik infinitívusz az igei komplexum élére kerül, pl. *pisilni el tudtál menni*.

A VFrame keresőeljárás a több infinitívuszt tartalmazó mondatokkal is megbirkózik. Minden igei elem, így a főige és a mondatban szereplő infinitívuszi vonzatok egyaránt elindítják a saját VFrame keresőeljárásukat a megfelelő beállításokkal. Az olyan mondatok esetén, amelyeket a vonzatok nem természetes sorrendje miatt az ember is nehezen elemez, az elemző visszalépéssel és újraelemzéssel alakítja ki a megfelelő igei elem–igekötő és igei elem–infinitívuszi vonzat viszonyokat.

4.2. A homonímia

Két gyakori igekötő, a *meg* és a *ki* esetében gondot jelent az, hogy mindkét szó homonim, és sokszor hibás annotációval szerepel a korpuszban. A *meg* gyakran IK (azaz igekötő) címkét kap akkor is, ha mellérendelő kötőszó, a *ki* igekötő pedig gyakran keveredik az azonos alakú vonatkozó- illetve kérdő névmással. A hibásan annotált szavak automatikus azonosítását nehezítik az olyan esetek, amikor ezek valóban létező kombinációt alkothatnának az igével, ráadásul olyan pozícióban is állnak, amely az igekötők számára is elérhető (lásd az 3. példát).

(3) a. *akkor csak lámpát kell vennem* ***meg*** *rácsot*
a *meg* igekötőként a létező *meg+vesz* igét eredményezheti

b. *az mennyibe fog kerülni és* ***ki*** *fogja rá adni a pénzt*
a *ki* igekötőként a *ki+ad* létező igét eredményezheti

4.3. „Megírni meg kell”

A korpuszból kinyert mondatokban több mint 200 példát találtunk egy különleges szerkezetre, amelyben látszólag nem tartozik ige az igekötőhöz. A szerkezet egy infinitívuszból, egy finit igéből (jellemzően segédigéből) és egy olyan igekötőből áll, amely az infinitívuszon is megjelenő igekötő hangsúlyos alakja. Például: ***elképzelni*** *bármit* ***el*** *lehet,* ***becsajozni be*** *tudnék.*

4.4. Más igei elemek

A többi, vonzatkerettel és igekötővel rendelkező igei elem (melléknévi és határozói igenév) is rendelkezik VFrame keresőeljárással, ám ezek esetében számos más probléma is felmerül – például a befejezett melléknévi igenév–melléknév–múlt idejű ige szófaji többértelműség kezelése –, amelyek további kutatások tárgyát képezik. Ezen igei elemek esetében a VFrame kiegészül egy olyan megszorítással, amely szerint az igekötőt vagy az infinitívuszi elemet az igenevet tartalmazó NP határain belül és csak balra (a tározóban) keresi.

5. Összefoglalás

Korpuszméréseink alapján bizonyítottuk, hogy az AnaGramma elemzőrendszer keretein belül a finit ige–igekötő kapcsolat létrehozása mellett [10] az infinitívusz–igekötő és a finit ige–infinitívuszi vonzat kapcsolatok létrehozásához is elegendő a feltételezett két token méretű elemzési ablak használata. A tározó és az ablak segítségével a VFrame keresőeljárás a mondatban szereplő igei elemeket (finit és infinit igéket) valamint az igekötőket a megfelelő módon kapcsolja össze.

Az aktuális finit ige–igekötő–infinitívuszi vonzat kapcsolatok létrejötte után elindulnak a megfelelő vonzatkeresők, amelyek mind a tározóban, mind a mondat hátralévő részében keresik a vonzatkeret elemeit. Amennyiben a VFrame nem egyértelműsíti teljesen a vonzatkeretet (mert egy ige ugyanazzal a finit ige–igekötő–infinitívuszi vonzat viszonnyal többféle vonzatkerettel is rendelkezhet), akkor az összes ennek megfelelő vonzatkeret vonzatkeresője elindul. Ekkor a mondatban aktuálisan szereplő többi vonzat egyértelműsíti a vonzatkeretet.

Hivatkozások

1. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Bence, K., ed.: "Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére. MTA NYTI, Budapest (2016) 223–232
2. Indig, B., Vadász, N., Kalivoda, Á.: Decreasing Entropy: How Wide to Open the Window? In Martín-Vide, C., Mizuki, T., Vega-Rodríguez, M.A., eds.: Theory and Practice of Natural Computing: 5th International Conference, TPNC 2016, Sendai, Japan, December 12-13, 2016, Proceedings, Cham, Springer (2016) 137–148
3. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott nyelvtudomány **15**(1-2) (2015) 29–44
4. Turi, Z., Németh, D., Hoffmann, I.: Nyelv és emlékezet. In Pléh, C., Lukács, A., eds.: Pszicholingvisztika 2. Akadémiai Kiadó, Budapest (2014) 743–776
5. Frazier, L., Fodor, J.D.: The Sausage Machine: A New Two-Stage Parsing Model. Cognition **6**(4) (1978) 291–325
6. Oravecz, C., 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 1719–1723
7. Endrédy, I.: Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz (2016) PhD disszertáció. PPKE-ITK.
8. Kalivoda, Á.: A magyar igei komplexumok vizsgálata (2016) MA szakdolgozat. PPKE-BTK. `https://github.com/kagnes/hungarian_verbal_complex`.
9. É. Kiss, K.: Az ige utáni szabad szórend magyarázata. Nyelvtudományi Közlemények **104** (2007) 124–152
10. Indig, B., Vadász, N.: Windows in Human Parsing – How Far can a Preverb Go? In Tadić, M., Bekavac, B., eds.: Tenth International Conference on Natural Language Processing (HrTAL2016) 2016, Dubrovnik, Croatia, September 29-30, 2016, Proceedings, Cham, Springer (2016) (Elfogadva, nyomtatás alatt)