Ablak által világosan

Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével

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

2017 január 26.

[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

Az előadás felépítése

- az AnaGramma
  - működési alapelvei
  - az *ablak* és a *tározó*
- kolbásztöltés
- vonzatkeret-egyértelműsítés
- korpuszmérések
  - a finit ige és a jobbra kihelyezett igekötőjének távolsága
  - az infinitívusz és a jobbra kihelyezett igekötőjének távolsága
  - a finit ige és a tőle jobbra elhelyezkedő infinitívuszi vonzat távolsága
- a VFrame keresőeljárás
- további kutatást igénylő esetek

Az AnaGramma

AnaGramma

Az AnaGramma elemző (Prószéky et al., 2016; Prószéky and Indig, 2015) főbb tulajdonságai:

- performanciaalapú
- pszicholingvisztikailag motivált
- balról jobbra és szavanként elemez
- új kategóriák és hierarchikus jegyek
- függőségi gráf különböző típusú irányított élekkel

Működési alapelvei:

- kereslet-kínálat keretrendszer
- tározó
- ablak

Kolbásztöltés

Kétfázisú mondatfeldolgozó modell (Frazier and Fodor, 1978):

1. Preliminary Phrase Packager

   a bemenet szócsoportjaiból frázisokat „csomagol”

   itt történik az egyértelműsítés

2. Sentence Structure Supervisor

   a „csomagok” megkapják a szerepüket a mondatban

Vonzatkeret-egyértelműsítés

A *szeret* ige néhány vonzatkerete:

(igekötő) + ige	vonzatok		
szeret[1]	Nom	Acc	
szeret[2]	Nom	Inf	
meg + szeret[1]	Nom	Acc	
meg + szeret[2]	Nom	Inf	
agyon + szeret	Nom	Acc	
viszont + szeret	Nom	Acc	
bele + szeret	Nom	Ill	
ki + szeret	Nom	Ela	
el + szeret	Nom	Acc	Abl

Korpuszmérések

A használt korpuszok

InfoRádió korpusz:

- 54.996 hír, 135.587 mondat, **1.953.419 token**
- cím + 2-3 mondatos hír
- csak szerkesztett szövegek *!* ideális bemenet az elemző számára

MNSZ2 v.2.0.3 (Oravecz et al., 2014):

- **785 millió token** (írásjelek nélkül)
- szerkesztett és szerkesztetlen szöveg (beszédátiratok is)

Pázmány Korpusz (Endrédy, 2016):

- **1,2 milliárd token** több mint 30.000 weboldalról
- főkorpusz (szerkesztett) és kommentkorpusz (szerkesztetlen)

Igeosztályok

Öt igeosztály az igéknek az igekötővel és az infinitívuszi vonzattal való kombinálhatósága alapján:

	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...	?

X: nincs elfogadott infinitívuszi vonzat vagy igekötő az igével kombinálva

?: az igének lehet, hogy van infinitívuszi vonzata

Kié az igekötő?

A finit ige (FIN), az infinitívuszi vonzat (INF) és valamelyik igekötőjének (PreV) egymáshoz viszonyított lehetséges sorrendjei:

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

A finit ige és a tőle jobbra elhelyezett igekötőjének távolsága

FIN	+1	+2	+3	+4	+5	+6	+7
MNSZ2	7.527.308	163.993	5.126	1.193	267	101	27
Inforádió	23.552	220	-	-	-	-	-
MNSZ2%	97,78%	2,13%	0,0666%	0,015%	0,003%	0,001%	3,5e-4%
Inforádió%	99,999%	0,001%	-	-	-	-	-

Szerkesztett szövegekben 99,9%-ban közvetlenül az ige után, szerkesztetlen szövegekben 99,9%-ban maximum két token távolságra helyezkedik el az igekötő.

A finit ige és a tőle jobbra elhelyezett igekötőjének távolsága

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

b. Csábítson téged a retyezáti nagy barna medve oda ahova akarod.

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

Néhány gyakori igekötő távolsága a finit igétől:

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

Az igekötők 98,5%-a az igén vagy közvetlenül utána áll, csak 1,01% távolodik el jobban (MNSZ2)

Az inf. és a tőle jobbra lévő igekötőjének távolsága

A méréseket a Pázmány Korpuszon (Endrédy, 2016) végeztük:

webkorpusz *!* sok szerkesztetlen szöveg (2 milliós kommentkorpusz)

INF [...] IK	db.	%
össz.	717	
+1	619	86,3
+2	52	7,3
+3	35	4,9
> +3	11	1,5
max. 2	671	93,6%

(2) *épp foglalni akartam le a buszt*

(3) *már indulni akartam vissza*

(4) a. *vinni kell a kamerát el*

b. *menekülni akartak a városon keresztül vissza*

A finit ige és a tőle jobbra lévő inf. vonzatának távolsága

A méréseket a Pázmány Korpuszon (Endrédy, 2016) végeztük:

webkorpusz *!* sok szerkesztetlen szöveg (2 milliós kommentkorpusz)

FIN [...] INF	db.	%
össz.	727.562	
+1	652.778	89,7
+2	47.669	6,6
> +2	27.115	3,7
max. 2	700.447	96,3%

Ha az infinitívusz túl messze kerül (példa az MNSZ2-ből):

(5) *Ha már valamit nem vonnak le automatikusan a fizetésből, akkor már lehet, hogy be se fogja az a szakszervezeti tagdíjat fizetni*.

VFrame

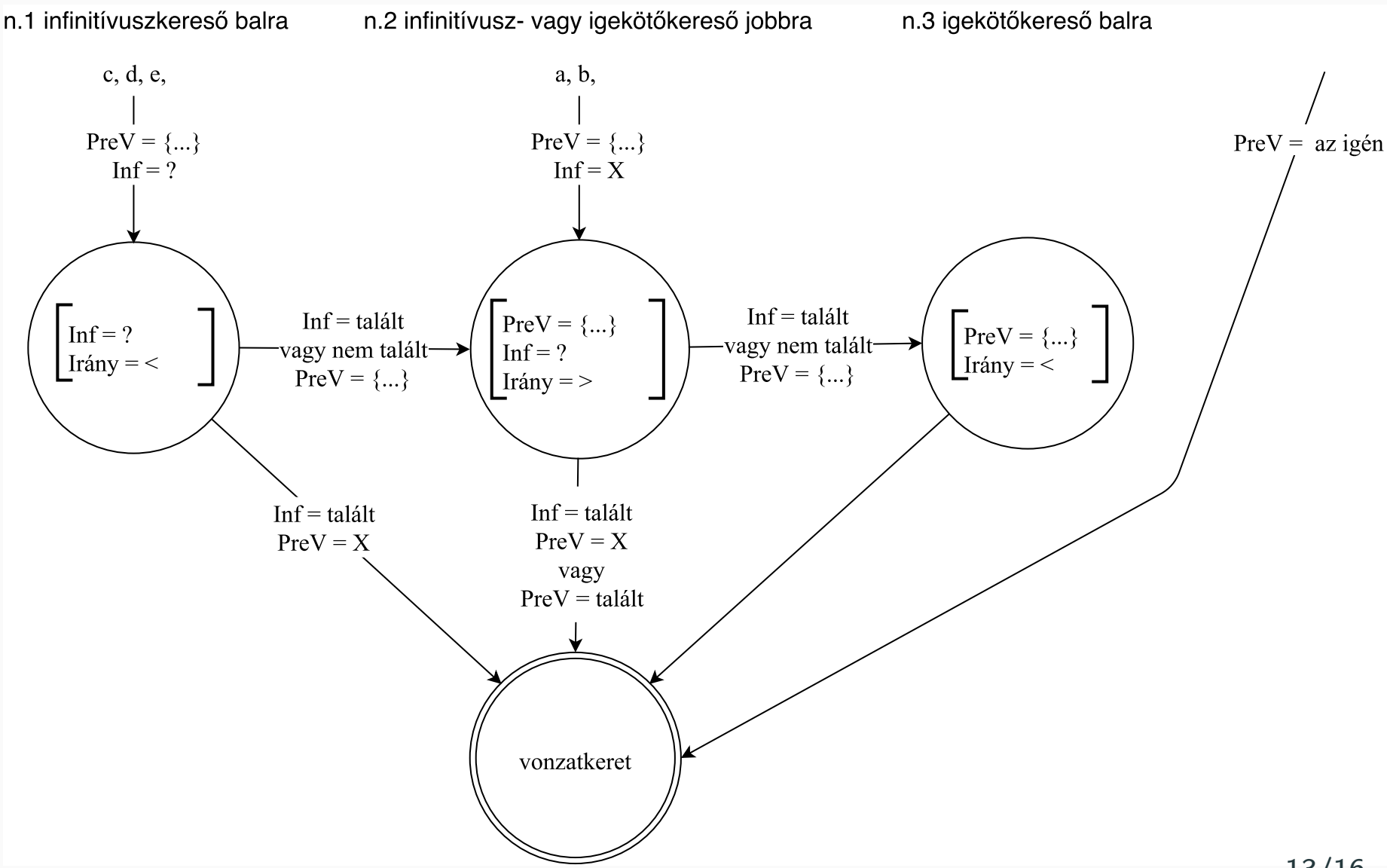
A VFrame kerespeljárás architektúrája

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

$$\text{szeret}\begin{bmatrix}\text{Igekötő} = \{\text{meg, agyon, viszont, bele, ki, el}\} \\ \text{Infinitívusz} = ?\end{bmatrix}$$

$$\text{szeret}\begin{bmatrix}\text{Igekötő} = \{\text{meg}\} \\ \text{Infinitívusz} = \textit{talált}\end{bmatrix}$$

VFrame


n.1 infinitívuszkereső balra
n.2 infinitívusz- vagy igekötőkereső jobbra
n.3 igekötőkereső balra
c, d, e,
PreV = {...}
Inf = ?
a, b,
PreV = {...}
Inf = X
PreV = az igén
Inf = ?
Irány = <
Inf = talált
vagy nem talált
PreV = {...}
PreV = {...}
Inf = ?
Irány = >
Inf = talált
vagy nem talált
PreV = {...}
PreV = {...}
Irány = <
Inf = talált
PreV = X
Inf = talált
PreV = X
vagy
PreV = talált
vonzatkeret

Problémás esetek

Problémás esetek

1. több infinitívusz

   (6) a. *el kell kezdeni keringőzni tanulni*

   b. *el fogod tudni dönteni*

   c. *pisilni el tudtál menni*

2. hominímia

   (7) a. *akkor csak lámpát kell vennem meg rácsot*

   b. *az mennyibe fog kerülni és ki fogja rá adni a pénzt*

3. „megírni meg kell”

   (8) a. *elképzelni bármit el lehet*

   b. *becsajozni be tudnék*

4. más igei elemek
5. produktív igekötők

   (9) *ki|facebookozta a szemét*

6. mellérendelés és elliptikus szerkezetek

Összefoglalás

ismerjük az igék igekötőit és lehetséges vonzatkereteit

az igekötő általában elég közel van az igéjéhez

az infinitívusz általánban elég közel van a főigéhez

! a VFarme megtalálja őket a vonzatkeret-egyértelműsítéshez

Köszönöm a figyelmet!

Hivatkozások

Endrédy, I. (2016).
Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz.
PhD disszertáció. PPKE-ITK.

Frazier, L. and Fodor, J. D. (1978).
The Sausage Machine: A New Two-Stage Parsing Model.
Cognition, 6(4):291–325.

Oravecz, C., Váradi, T., and Sass, B. (2014).
The Hungarian Gigaword Corpus.
In Calzolari, N. and et al., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation, May 26-31, 2014*, pages 1719–1723, Reykjavik, Iceland. ELRA.

Prószéky, G. and Indig, B. (2015).
Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel.