Egy igei vonzatkereteket szűrő eljárás implementálása és kiértékelése

Vadász Noémi, Kalivoda Ágnes

2018. január 5.

Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport

Az előadás felépítése

- Az AnaGramma
  - Alapelvek
  - Kolbásztöltés
- Igei vonzatkeretek a számítógépben
  - Manócska
  - Vonzatkeret-egyértelműsítés
- VFrame
  - Felépítése
  - A szótár
  - Kiértékelés

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. **P**reliminary **P**hrase **P**ackager
   - a bemenet szócsoportjaiból frázisokat „csomagol”
   - itt történik az egyértelműsítés
2. **S**entence **S**tructure **S**upervisor
   - a „csomagok” megkapják a szerepüket a mondatban

Igei vonzatkeretek a számítógépben

Manócska: egy egységesített igei vonzatkerettár

Indig Balázs: Manócska – integrált igei vonzatkeret adatbázis (2017)

Bárki számára elérhető: `https://github.com/ppke-nlpg/manocska`

Felhasznált erőforrások:

- Sass Bálint et al.: Magyar igei szerkezetek (szótár)
- Sass Bálint: 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet (lista)
- Kornai András et al.: Tádé – Igei vonzatkeret-gyakorisági lista
- Kalivoda Ágnes: Igekötős igék gyakorisági listája
- Kalivoda Ágnes: Infinitívuszi vonzattal bíró igék gyakorisági listája

Igekötős igék gyakoriségi listája

Elérhető: `https://github.com/kagnes/hungarian_verbal_complex`

Az MNSZ 2.0.3 alapján készült, 27091 lemmából álló, átnézett lista.

Néhány példa...

a lista elejéről		**a lista végéről**	
gyakoriság	**IK+IGE**	**gyakoriság**	**IK+IGE**
292 019	el+mond	5	agyon+fárad
145 257	el+fogad	5	agyon+csócsál
141 467	ki+derül	5	agyon+büntet
124 530	meg+jelenik	5	agyon+adóztat
104 184	hozzá+tesz	5	abba+hagyódik

Infinitívuszi vonzattal bíró igék gyakorisági listája

Elérhető:
`https://github.com/kagnes/infinitival_constructions`

Az MNSZ 2.0.4 alapján készült, 1507 lemmából álló, átnézett lista.

Az egyes sorok szerkezete, példákkal:

gyak.	finit ige	"INF-típus"	példa	megjegyzés
20 054	készül	arra	kitörni készült	
11 500	el+megy	azért	elmegyek megkeresni	
1 400	támad	x	kedvem támadt elvonulni	@lex: kedve

Leggyakoribb "INF-típusok":

- *azért*: 59,6% (mozgásigék)
- *azt*: 11,5%
- *x*: 8,9% (a finit igének lexikális vonzata is van)

Vonzatkeret-egyértelműsítés

A *szeret* ige néhány vonzatkerete:

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

Korpuszmérések (Kalivoda, 2016) I.

A finit ige és a posztverbális igekötő távolsága:

(Felhasznált korpuszok: MNSZ 2.0.3, Inforádió)

	+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%	-	-	-	-	-

(1) a. *Kollár doktor körzetében azért nem **merül** ez a kérdés ilyen sarkallatosan **föl**, ...*

b. *Azért **mentem** egy kicsit a popzene felé **el**, ...*

Korpuszmérések (Vadász et al., 2017) II.

Az infinit ige és a posztverbális igekötő távolsága:

(Felhasznált korpusz: Pázmány Korpusz (Endrédy, 2016)):

INF [...] IK	db.	%
össz.	717	
+1	619	86,3
+2	52	7,3
>+2	46	6,4
max. 2	671	**93,6%**

(2) *épp* ***foglalni*** *akartam* ***le*** *a buszt*

(3) *már* ***indulni*** *akartam* ***vissza***

(4) ***vinni*** *kell a kamerát* ***el***

Korpuszmérések (Vadász et al., 2017) III.

A finit ige és a tőle jobbra lévő INF vonzat távolsága

(Felhasznált korpusz: Pázmány Korpusz (Endrédy, 2016)):

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: egy vonzatkeret-egyértelműsítő eljárás

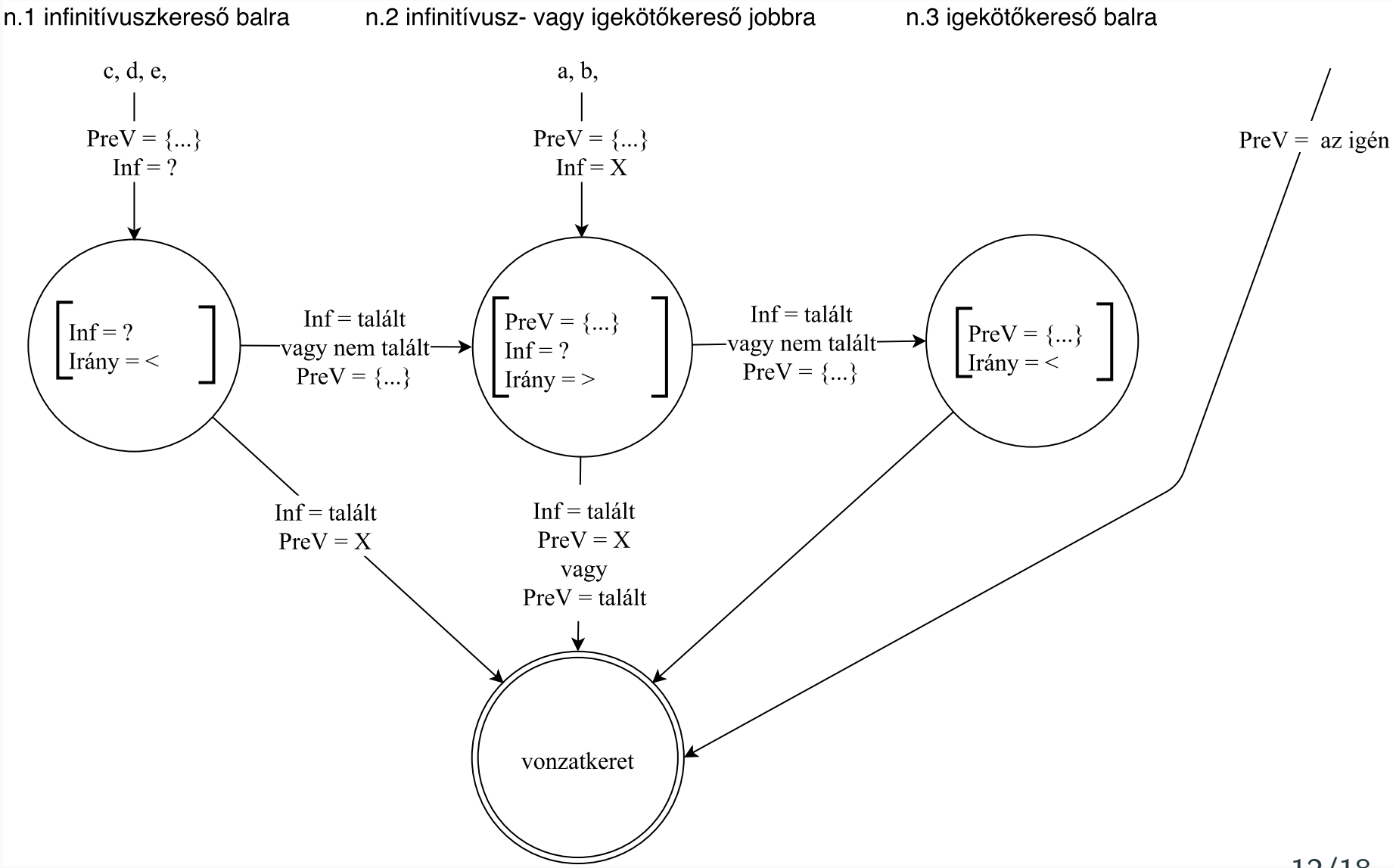
A kereső architektúrája

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

$$\text{kíván}\begin{bmatrix}\text{Igekötő} = \{\text{meg, -}\} \\ \text{Infinitívusz} = ?\end{bmatrix}$$

$$\text{kíván}\begin{bmatrix}\text{Igekötő} = \{\text{-}\} \\ \text{Infinitívusz} = \textit{talált}\end{bmatrix}$$

Működése


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

A VFrame szótára

Két szólistából készült:

- `https://github.com/kagnes/hungarian_verbal_complex`
- `https://github.com/kagnes/infinitival_constructions`

igénként (minden igekötőhöz külön):
lehet-e infinitívuszi vonzata

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

Kiértékelés I.

Tesztmondatok:

- 1 000 tagmondat
- MNSZ.2.0.4

Módszerek:

- **kézi annotáció**
- **VFrame**
- **baseline** (Recski, 2011)
- `magyarlanc` (Zsibrita et al., 2013)

```
%mit lélegeznek ki a falak ,
ki      lélegeznek      -       -
-       lélegeznek      -       -
ki      lélegeznek      -       -
ki      lélegeznek      -       -

%de az új fegyvereket ki kell próbálni .
-       kell    ki      próbálni
-       kell    ki      próbálni
-       kell    -       próbálni
-       kell    ki      próbálni
```

Kiértékelés II.

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

mérőszámok:

- **pontosság** a találatokból hány volt eredetileg jó
- **fedés** az eredetileg jók közül hányat találtunk meg
- **F-mérték** a pontosság és a fedés harmonikus közepe

Eredmények

		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

Köszönjük 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.

Kalivoda, Á. (2016). A magyar igei komplexumok vizsgálata. Mesterszakos szakdolgozat. PPKE-BTK. `https://github.com/kagnes/hungarian_verbal_complex`.

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

Prószéky, G., Indig, B., and Vadász, N. (2016). Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Bence, K., editor, *"Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére*, pages 223–232. MTA NYTI, Budapest.

Recski, G. (2011). A sekély mondattani elemzés további lépései. In Tanács, A. and Vincze, V., editors, *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 113–118.

Vadász, N., Kalivoda, Á., and Indig, B. (2017). Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In Vincze, V., editor, *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*, pages 3–12, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Zsibrita, J., Vincze, V., and Farkas, R. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP 2013. Hissar, Bulgária, 2013.09.08-2013.09.13.*, pages 763–771.