



Egy egységesített magyar igei vonzatkerettár építése és felhasználása

Vadász Noémi^{1,3}, Kalivoda Ágnes^{1,3}, Indig Balázs^{2,3}

2018. január 18.

¹Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar

²Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

³MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport

- a Manócska igei vonzatkerettár
 - felépítése
 - egységesítés és hibajavítás
- a VFrame keresőeljárás
 - ismertetése
 - kiértékelése

Manócska, egy egységesített igei vonzatkerettár

A felhasznált erőforrások

- **Sass Bálint et al.:** (Sass et al., 2010)
Magyar igei szerkezetek (szótár)
- **Sass Bálint:** (Sass, 2015)
Mazsola – 28 millió szintaktikailag
elemzett mondat és 500 000 igei szerkezet
(lista)
- **Kornai András et al.:** (Kornai et al.,
2016)
Tádé – Igei vonzatkeret-gyakorisági lista
- **Kalivoda Ágnes:** (Kalivoda, 2016)
Menyus – Igekötős igék gyakorisági listája
- **Kalivoda Ágnes:**
Egérke – Infinitívuszi vonzattal bíró igék
gyakorisági listája

A felhasznált erőforrások

- **Sass Bálint et al.:** (Sass et al., 2010)
Magyar igei szerkezetek (szótár)
- **Sass Bálint:** (Sass, 2015)
Mazsola – 28 millió szintaktikailag elemzett mondat és 500 000 igei szerkezet (lista)
- **Kornai András et al.:** (Kornai et al., 2016)
Tádé – Igei vonzatkeret-gyakorisági lista
- **Kalivoda Ágnes:** (Kalivoda, 2016)
Menyus – Igekötős igék gyakorisági listája
- **Kalivoda Ágnes:**
Egérke – Infinitívuszi vonzattal bíró igék gyakorisági listája



Triviális hibák javítása a Humor morfológia segítségével

helytelen igealakok kiszűrése

tövesítés vagy szófaji egyértelműsítés hibája

nyug → törlés

12-15-ért → törlés

rossz igekötők megjavítása

szegmentálási hiba

ki|abál → *kiabál*

alá|z → *aláz*

ráró → *rá|ró*

kölcsön|sikerül → *sikerül*

meg|akar → *akar*

abba|bele|fut → *bele|fut*

	igekötő	nem igekötő
szó	<i>kifolyik</i>	<i>felvételizik</i>
tagadás-teszt	<i>nem folyik ki</i>	<i>nem felvételizik</i>
„is”-teszt	<i>ki is folyik</i>	-

- egybeírás/különírás mint az „igekötőség” mértéke
- a korpuszban nem következetes
- ha az igével egybeírt változat az összes előfordulás min. 15%-a, akkor igekötő
- pl. (MNSZ 2.0.4) *helyt (áll, ad)*:
 - 1 396 külön, az ige előtt
 - 6 548 egybeírva az igével

<i>abba/hagy</i> [ACC]	a keret gyakorisága	összes keret összgyakorisága
Magyar Igei Szerkezetek	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
infinitívus lista	0	1 507
rang	0.00168342	

Az infinitívuszi vonzatok a Manócskában

Ami a Tádében nem megfelelő:

- lexikális vonzatként kezeli az INF-vonzatokat
(pl. *alábukik megnedvesíteni*, ami lehetne *alábukik gyöngyhalászni* is)
- ha az ige vonzatkeretében szerepel INF, akkor az összes többi vonzatot az INF-nek tulajdonítja
(ellenpélda: *[megtanít vkit] vmizni*, *[megérez vmit] vmizni*)

Infinitívuszi vonzattal bíró igék gyakorisági listája:

- Az MNSZ 2.0.4 alapján készült, 1507 lemmából álló, átnézett lista.
- Tartalmazza:
 - az INF-vonzatos igéket,
 - ezek gyakoriságát,
 - egy-egy példamondatot az MNSZ 2.0.4-ből,
 - egyéb megjegyzéseket a lexikális vonzatra vonatkozóan,
pl. ha a finit igének lexikális vonzata van, és ez a komplexum vonzza az INF-et, ld. *kedve támad vmizni*.

A Manócska számokban (javítás előtt és után)

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

A Manócska számokban

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

A VFrame keresőeljárás

Az AnaGrammar 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**

A Manócskából építjük:

- **Kalivoda Ágnes:** Igeköötős igék gyakorisági listája (Kalivoda, 2016)
- **Kalivoda Ágnes:** Infinitívuszi vonzattal bíró igék gyakorisági listája

utál ?, el:X, ki:X, meg:X

felejt ?, el:?, ki:X, le:X, ott:X, rajta:X

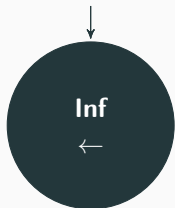
PreV az ígén



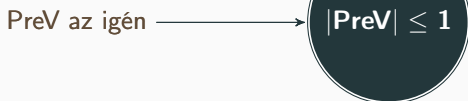
$$|\text{PreV}| \leq 1$$

A VFrame algoritmus

több PreV és lehet Inf

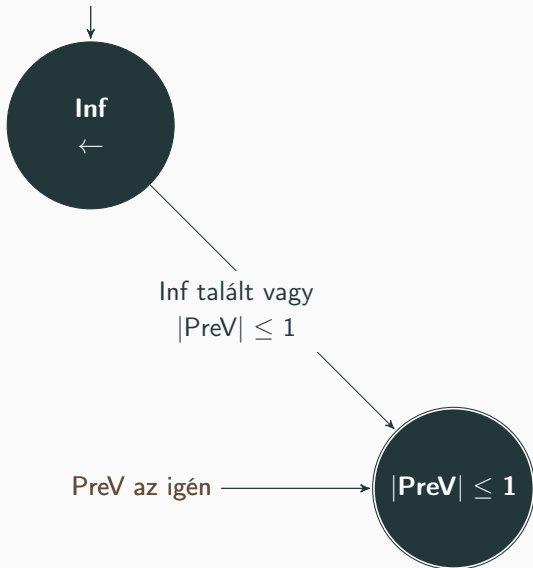


PreV az ígén



A VFrame algoritmus

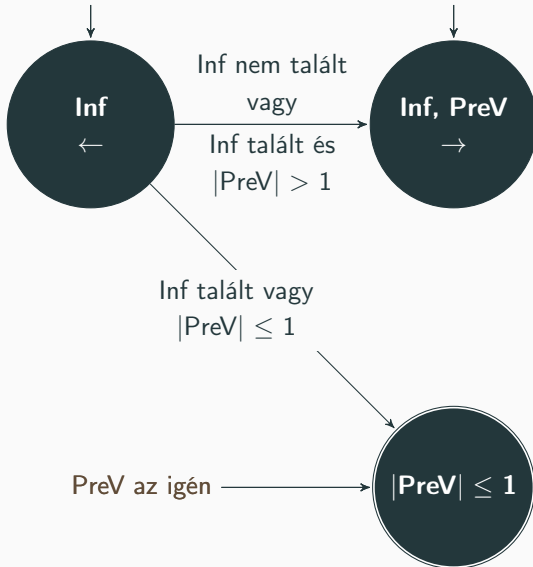
több PreV és lehet Inf



A VFrame algoritmus

több PreV és lehet Inf

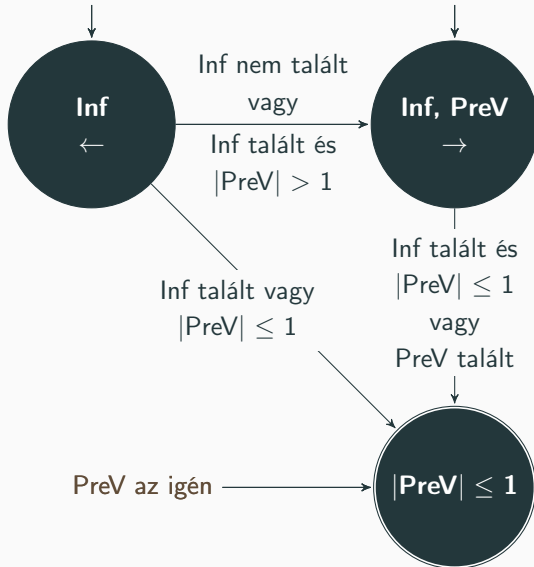
több PreV és nem lehet Inf



A VFrame algoritmus

több PreV és lehet Inf

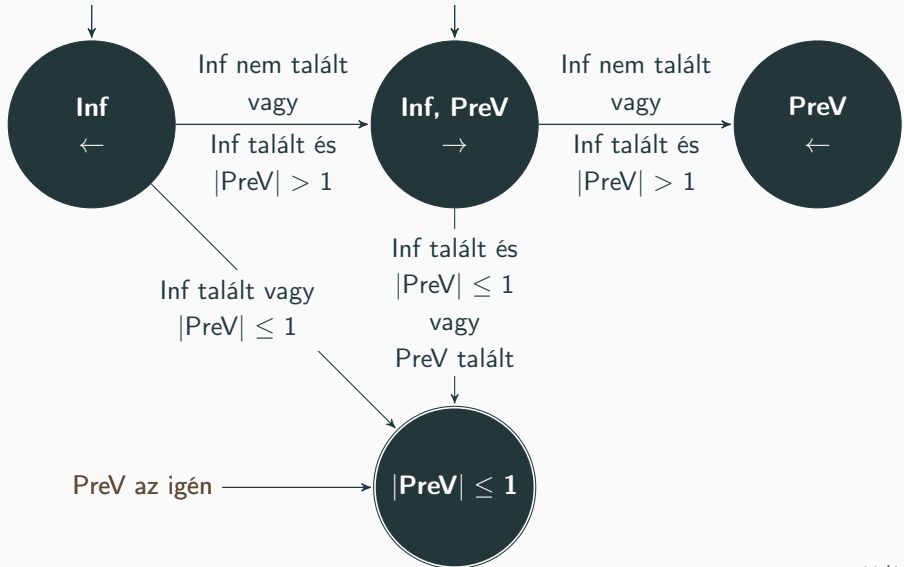
több PreV és nem lehet Inf



A VFrame algoritmus

több PreV és lehet Inf

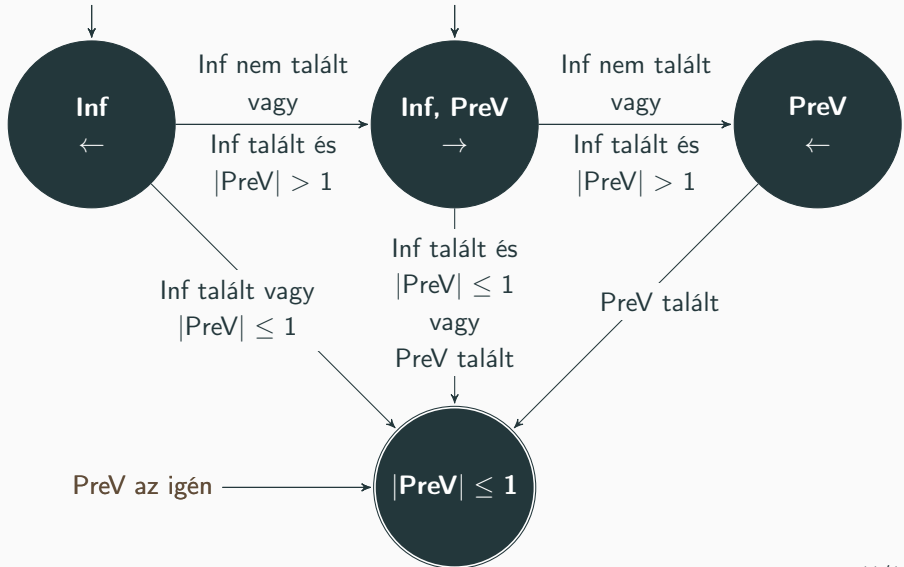
több PreV és nem lehet Inf



A VFrame algoritmus

több PreV és lehet Inf

több PreV és nem lehet Inf



1 000 tesztmondat (tagmondat)

összevetés két másik módszerrel:

- baseline: kösd össze az igét a legközelebbi igekötővel/infinitívusszal (Recski, 2011)
- magyarlanc függőségi elemzője

	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

		Fin-lk	Inf-lk	Fin/Inf-lk	Fin-Inf	Összesen
pontosság	VFrame	97,60	94,71	96,85	97,88	97,23
	baseline	92,39	91,41	92,14	96,98	93,89
	magyarlánc	94,55	90,96	93,56	97,16	94,94
fedés	VFrame	97,60	94,21	96,72	98,34	97,31
	baseline	96,49	93,78	95,78	99,05	96,97
	magyarlánc	86,19	87,69	86,59	96,93	90,36
F-mérték	VFrame	97,60	94,46	96,78	98,11	97,27
	baseline	94,40	92,58	93,92	98,00	95,41
	magyarlánc	90,18	89,30	89,94	97,04	92,59

Köszönjük a figyelmet!

- Kalivoda, Á. (2016).
A magyar igei komplexumok vizsgálata.
Mesterszakos szakdolgozat. PPKE-BTK.
https://github.com/kagnes/hungarian_verbal_complex.
- Kornai, A., Nemeskey, D. M., and Recski, G. (2016).
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., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).
- Prószték, G. and Indig, B. (2015).
Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel.
Alkalmazott Nyelvtudomány, 15(1-2):29–44.
- Prószték, 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.
- Sass, B. (2015).
28 millió szintaktikailag elemzett mondat és 5 00000 igei szerkezet.
In Tanács, A., Varga, V., and Vincze, V., editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 399–403, Szeged. SZTE TTIK Informatikai Tanszékcsoport.
- Sass, B., Váradi, T., Pajzs, J., and Kiss, M. (2010).
Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára.
Tinta Könyvkiadó, Budapest.