Véges erőforrás végtelen sok igekötős igére

Kalivoda Ágnes

MTA Nyelvtudományi Intézet
MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport
PPKE BTK Nyelvtudományi Doktori Iskola

Bevezetés

Véges erőforrás: PrevLex

- 54 955 igekötős ige gyakorisági adatokkal
- az MNSZ 2.0.4 alapján (Oravecz–Váradi–Sass 2014)
- manuálisan átnézett
- elérhető: `https://github.com/kagnes/prevlex`

Végtelen sok igekötős ige:

- sok igekötővel tetszőleges számú új szó képezhető
  $\rightarrow$ a PrevLex nem lehet teljes
- hogyan mérhető az igekötők morfológiai produktivitása?
- mire használhatók a kvantitatív eredmények?

A PrevLex

Az adatfeldolgozás menete

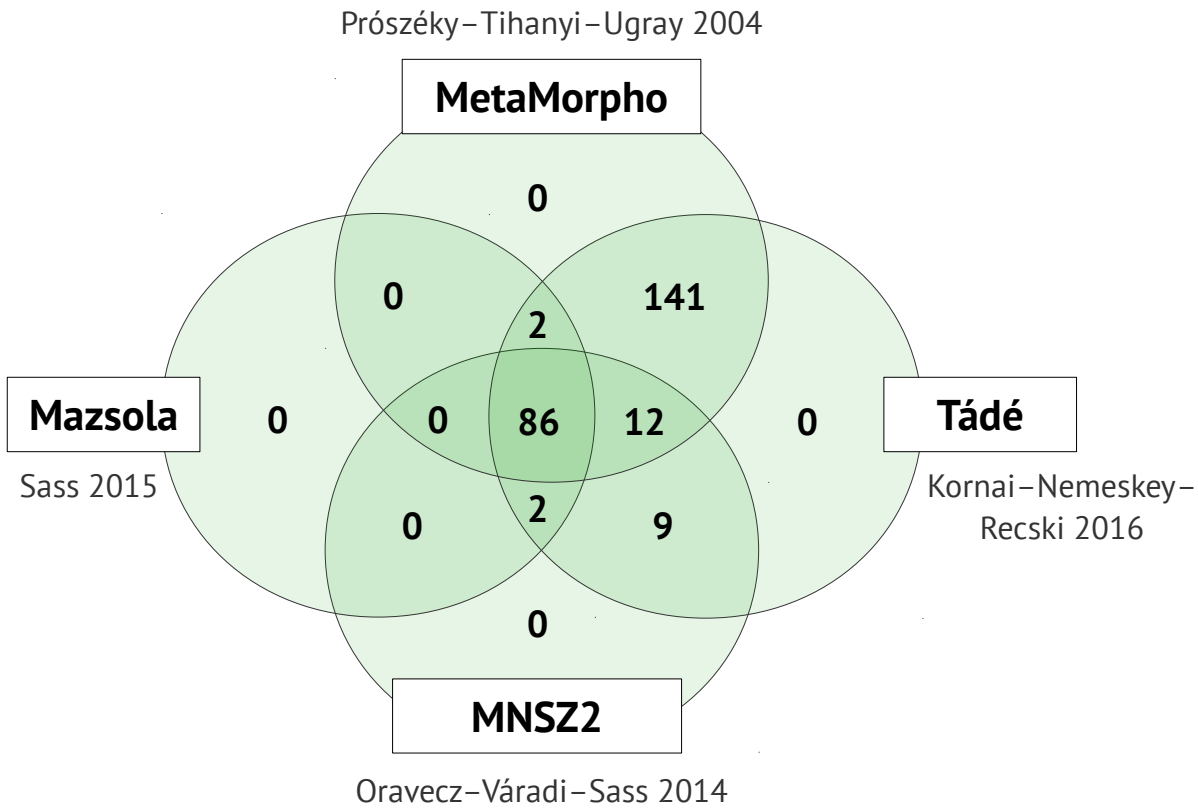
Kiinduló korpusz: MNSZ 2.0.4 forrásfájl

A módosított korpusz az alábbiak szerint alakult:

korpusz	**token**	**százalék**
eredeti MNSZ2	**1 348 000 000**	**100,00**
versek	5 661 000	0,42
UNKNOWN/SKIP	26 825 200	1,99
duplumok	271 217 600	20,12
módosított MNSZ2	**1 044 296 200**	**77,47**

Az adatfeldolgozás menete

Lehetséges igekötők a **Manócskából** (Kalivoda–Vadász–Indig 2018):



Az adatfeldolgozás menete

1. minden ‘igekötő + finit ige’ / UNKNOWN lekérése a korpuszból

2. a közel 178 000 elemű találati lista átnézése

3. a forrásfájl lokális újraelemzése a javított adatokkal
pl. *elfilózok, elfilóztam* $\rightarrow$ *elfilóz*

4. a javított korpuszból a PrevLex előállítása

A PrevLex számokban

kategória	típus	token
összes igekötős ige	54 955	11 959 379
hapaxok	22 043	22 043
UNKNOWN szavak	5 156	26 542
UNKNOWN hapaxok	3 335	3 335

Zipf-eloszlás:

- néhány igekötős ige rendkívül nagy tokengyakoriságú
- a hapaxok ritkák, de sokfélék

Az igekötők morfológiai produktivitása, a produktivitás típusai

A morfológiai produktivitásról

- egy szóalkotási minta akkor produktív, ha tetszőleges számú, szemantikailag transzparens szó alkotható vele egy adott szemantikai tartományban (Kiefer–Ladányi 2000)

- a produktivitásnak vannak fokozatai:
  nem minden affixum egyformán produktív

- a morfológiai produktivitás kvantitatív vizsgálata
  Harald R. Baayen nevéhez köthető (Baayen 2009)

Megvalósult és terjeszkedő produktivitás

Megvalósult produktivitás *(realized productivity)*:

egy affixum részvétele a szóalkotásban a mérés időpontjáig

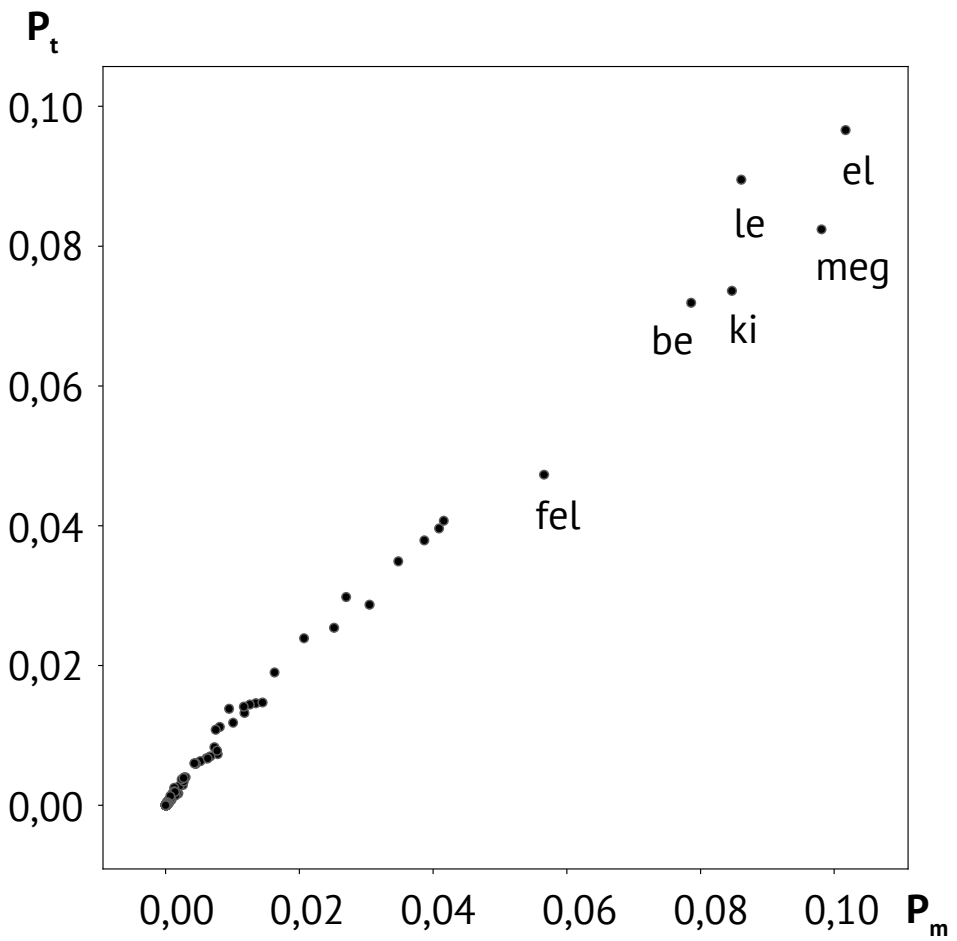
$$P_m = \frac{\text{az affixumot tartalmazó lemmák (típus)}}{\text{a korpusz összes lemmája (típus)}}$$

Terjeszkedő produktivitás *(expanding productivity)*:

feltételezés: ha egy affixum sok hapaxnál megtalálható, egyre inkább növekszik a produktivitása

$$P_t = \frac{\text{az affixumot tartalmazó hapaxok}}{\text{a korpusz összes hapaxa}}$$

Megvalósult és terjeszkedő produktivitás



Lehetséges produktivitás *(potential productivity)*

Mik azok a most még ritka affixumok, amelyek később sok szó képzésében vehetnek részt?

$$P_l = \frac{\text{az affixumot tartalmazó hapaxok}}{\text{az affixumot tartalmazó összes szó (token)}}$$

tokengyakoriság	**igekötő**	**példák**	$\mathbf{P_l}$
	mennybe	száll, visz	0,66
	égbe	megy, emel	0,50
5 <	oldalba	rúg, szúr	0,50
	szarrá	ázik, bombáz	0,44
	szénné	ég, tetovál	0,43

Mitől függhet a produktivitás?

Produktív szóképzési szabályok

Jellemzően produktívabbak azok az igekötők, amelyek névszóból képzett igéhez is kapcsolódhatnak.

A leggyakoribb ‘névszó (N) $\longrightarrow$ ige’ képzési sémák:

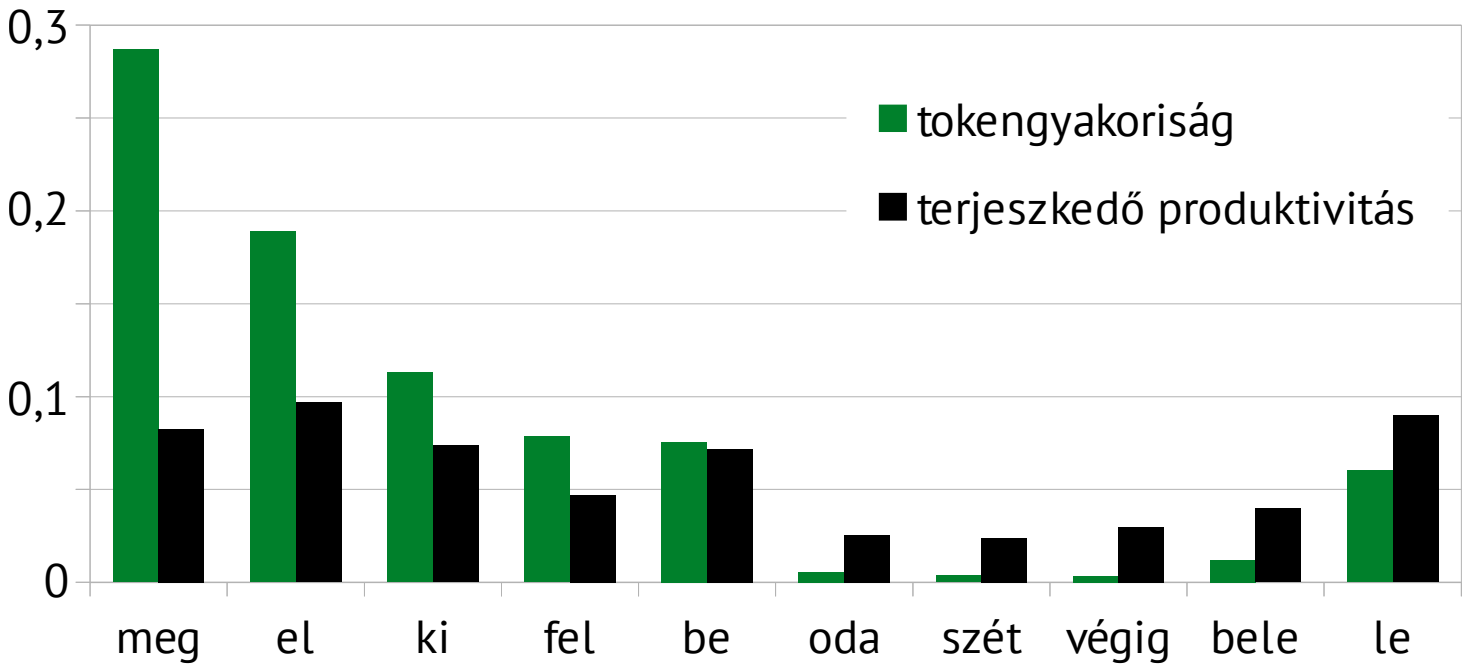
alapjelentés	**sematikus szerk.**	**néhány példa**
N-nel kapcsolatosat csinál	N+(O)z(ik) N+(O)l	*kisfiamozik, testékszerez* *rajzszögel, bokroscsomagol*
N-né változik	N+Ul/sUl N+Odik	*mémesül, szinglisül* *vékonyodik, tahósodik*
N-né változtat	N+ít/(O)sít	*részegít, szálkásít*
N-ként viselkedik	N+kOdik/skOdik	*vandálkodik, jópofáskodik*

Stílusregiszter

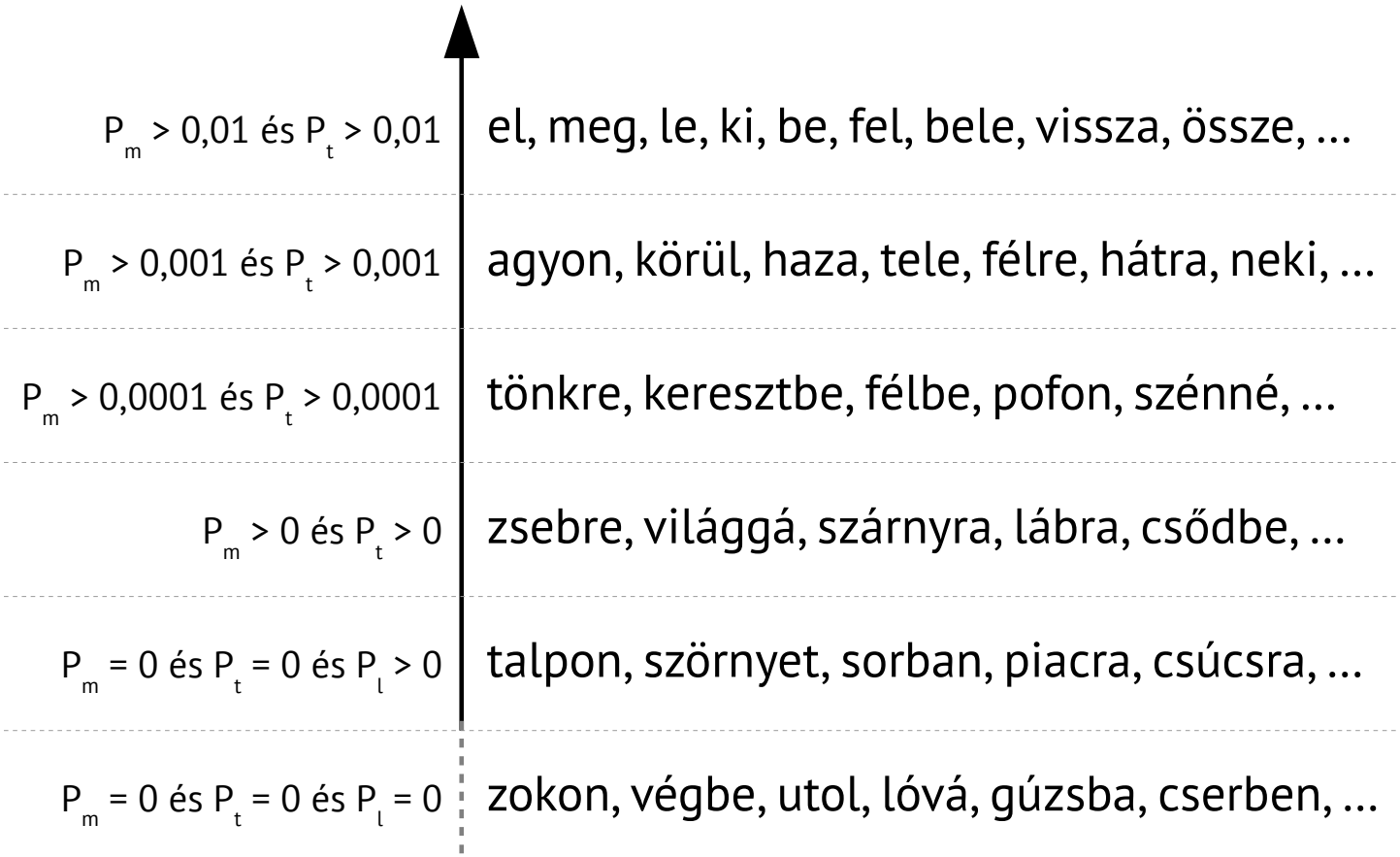
	személyes	beszélt	szépirod.	sajtó	tud.	hivatalos
MNSZ	**29,0**	**7,3**	**7,8**	**35,1**	**11,3**	**9,5**
el	34,6	6,1	28,6	18,2	9,6	2,9
meg	31,0	7,4	31,3	14,9	12,6	2,8
le	45,2	6,9	17,6	20,3	7,0	3,0
ki	33,0	6,5	29,7	17,4	10,4	3,0
be	41,7	8,2	22,0	16,2	8,8	3,1
fel	34,6	6,6	24,5	19,0	13,7	1,6
át	27,5	6,8	28,2	23,1	11,2	3,2
bele	34,0	6,2	33,1	18,0	6,0	2,7
vissza	27,7	7,2	33,2	20,5	7,2	4,2
össze	32,7	6,9	31,6	17,0	9,9	1,9

Gyakoriság

Nincs szoros összefüggés a produktivitás és a gyakoriság között:



Következtetések



Összefoglalás

Ami elhangzott:

- új lexikai erőforrás: PrevLex
- az igekötők morfológiai produktivitásának meghatározása
- az igekötő-kategória mint kontinuum

Kitekintés:

- a PrevLex-szel bővíthetők a morfológiai elemzők lexikonjai
- így csökkenthető az UNKNOWN elemzések száma
- ez jobb lexikont, pontosabb nyelvmodelleket eredményez

Hivatkozások

Baayen, H. (2009): Corpus linguistics in morphology: morphological productivity. In Lüdeling, A. – Kytö, M. (szerk.): Corpus Linguistics. An international handbook, Berlin, Mouton De Gruyter.

Kalivoda, Á. – Vadász, N. – Indig, B. (2018): Manócska: A Unified Verb Frame Database for Hungarian. In Sojka, P., et al. (szerk.): TSD 2018, Brno, Csehország, Springer-Verlag.

Kiefer, F. – Ladányi, M. (2000): A szóképzés. In Kiefer, F. (szerk.): Strukturális magyar nyelvtan 3., Morfológia, Budapest, Akadémiai Kiadó.

Kornai, A. – Nemeskey, D. M. – Recski, G. (2016): Detecting Optional Arguments of Verbs. In Calzolari, N., et al. (szerk.): LREC 2016, Portorož, Szlovénia, ELRA.

Oravecz, Cs. – Váradi, T. – Sass, B. (2014): The Hungarian Gigaword Corpus. In Calzolari, N., et al. (szerk.): LREC 2014, Reykjavik, Izland, ELRA.

Prószéky, G. – Tihanyi, L. – Ugray, G. (2004): Moose: a robust high-performance parser and generator. In Hutchins, J. (szerk.): EAMT 2004, La Valletta, Málta, Foundation for International Studies.

Sass, B. (2015): 28 millió szintaktikailag elemzett mondat és 500 000 igei szerkezet. In Tanács, A., et al. (szerk.): MSZNY 2015, Szeged, Szegedi Tudományegyetem Informatikai Intézet.