Van-e jobb dolog, mint az n-grammok?

Valószínűleg nincs.

Indig Balázs

ELTE BTK Digitális Bölcsészet Tanszék
Digitális Örökség Nemzeti Laboratórium

2022. március 7.

National Laboratory for Digital Heritage

Az előadóról

- 10 éve az n-grammok bűvöletében és még nincs vége :D
- Számtalan eszköz és platform (társ)szerzője
- A Digtális Örökség Nemzeti Laboratórium CTO-ja
- Feketeöves Python programozó, code reviewer
- Nyílt forráskód evangelista :)

Bevezetés

„If A and B have some environments in common and some not (e.g. occultist and lawyer) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments.”

(Harris 1954)

„You shall know a word by the company it keeps.”

(Firth 1957)

Definíció (a puhányoknak)

N-gram: (véges) n (n=1, 2, 3, stb.) elemből álló összefüggő (véges) sorozat

- Nem mondtuk meg, hogy mi az egység!
- Az n-grammok átlapolódhatnak, különböző hosszúságú n-grammok kombinálhatók
- A korpuszbeli gyakoriságuk alapján súlyozhatók (simítással finomítható a súlyozás)
- Adathiány probléma léphet fel (data sparseness), de léteznek trükkök ;)

Főbb fajtái:

- Karakter n-gram (v.ö. szótag és morféma)
  - 3-gram (trigram): **Ka**r, ara, **rak**, akt, kte, **ter**
- Szó n-gram (v.ö. szó, idézet és frázis)
  - 2-gram (bigram): **Szó n-gram**, *n-gram (v.ö.*, ..., *és frázis)*
  - Tokenizálva: **Szó n-gram**, *n-gram (*, *(v.ö.*, ..., **és frázis**, *frázis)*

Közelíti: szózsák (Bag of Words, BoW), subword, szóvektor
(majd egy másik alkalommal ezekről is lesz szó. ;))

Definíció (a kemény mag számára)

- **Pyhton implementáció(k)**: `https://github.com/dlazesz/n-gram-benchmark`
- Markov-elv: A szó teljes bal környezete közelíthető az utolsó pár baloldali szóval

$$\arg\max_{t_1 \ldots t_T} \left[\prod_{i=1}^{T} P(t_i | t_{i-1}, t_{i-2}) Q(w_i | t_{i-1}, t_i) \right] P(t_{T+1} | t_T) \tag{1}$$

- Simítások, közelítések: (Modified) Kneser–Ney smoothing
  - Ezt a legjobb, bár jó Python implemntáció nem nagyon van
  - Chen és Goodman (1999) jól összefoglalja a különféle simításokat
- Durván el lehet veszni az implementációs részletekben
  - logaritmikus valószínűség
  - numerikus stabilitás

Nyelvmodellek

A szöveg (részlet) „magyarosságát” tudja meghatározni, valamint „magyaros” folytatást/kiegészítést tud generálni egy adott szöveghez.

Számos jobbnál jobb implementáció van pár klikk távolságban:

- N-gram - KenLM (Heafield 2011)
- Word2Vec (Mikolov és tsai. 2013) - Gensim (Rehurek és Sojka 2011)
- BERT (Devlin és tsai. 2019) - huBERT (Nemeskey 2021a)
- Majd: GPT-3 (Elkins és Chun 2020)

Ezek vagy könnyen betaníthatók, vagy elérhetőek betanítva. Folyamatosan jönnek újak.

Vajon okosabbak-e, mint egy átlag ember?

Cloze teszt

```
" Be van fejezve a ________ mű, igen.
  A gép forog, ________ alkotó pihen.
  Év-milliókig eljár tengelyén,
  Mig ________ kerékfogát ujítni kell.
  Fel hát, ________ véd-nemtői, fel,
  Kezdjétek végtelen pályátokat.
  Gyönyörködjem ________ egyszer bennetek
  A mint elzúgtok ________ alatt. "
```

1. ábra. **Hiányzó szavak**: az, egy, lábaim, még, nagy, világim

Pont ezen a feladaton tanítjuk a mostani nyelvmodelleket! :)

Köszönetnyilvánítás:
https://www.controlaltachieve.com/2017/04/create-cloze-tests.html

MorphoLogic Kft. Nyelvi játék © 2000

2. ábra. **Megfejtés**: egy-egy

Metajáték

1. Végy egy nyelvi játékot!
   (`https://github.com/ELTE-DH/word-guessing-game`,
   `https://word-concordance-game.herokuapp.com/`)
2. Szerezz egy (valójában két) **tiszta** korpuszt a játékhoz!
3. Alakítsd át a játékot úgy, hogy nyelvmodelleket lehessen tesztelni rajta!
4. Add oda embereknek, hogy játszanak! (Hogy legyen emberi adat is.)
5. Profit! :)

Köszönetnyilvánítás:
Szeretnénk megköszönni az ismeretlen MorphoLogic dolgozónak a játékot. :)

A Korpusz(ok)

A tanítókorpusz a *Webkorpusz 2.0* a *huBERT* miatt (Nemeskey 2021a; Nemeskey 2021b).
Míg a tesztkorpusz a *Webkorpusz 1.0* (Halácsy, Kornai, Németh és tsai. 2004).
(**Nem szabad a tanítóanyagon mérni! :)**)

	Előtte mondatok	Utána mondatok	%	Előtte szavak	Utána szavak	%
Webkorpusz 1.0	42 482 107	13 915 132	32,75	589 080 971	272 544 786	46,26
Webkorpusz 2.0	589 398 448	199 627 778	33,86	9 217 857 283	4 036 428 613	43,78

1. táblázat. A Webkorpusz 1.0 és 2.0 méretei a szűrés előtt és után.

A célspecifikus szűrés a felhasználói élmény növelése érdekében történt.
(Részletek a következő dián. *A számok automatikusan reprodukálhatóak.*)

Köszönetnyilvánítás:
Szeretnénk megköszönni a két korpusz szerzőinek hogy szabadon elérhetővé tették a munkájukat.

A Szűrés

- A 11 szónál rövidebb és 50 szónál hosszabb
- A több mint 25 karakter hosszú szót tartalmazó
- A 3 egymást követő szóban nagy kezdőbetűs szavakat (névelemek) tartalmazó
- A legalább 2 db és 4 karakter hosszú, nagybetűs szavakat (kiabálás) tartalmazó
- A 3 egymást követő, csak nem alfanumerikus karakterekből álló szavakat tartalmazó
- A három egymást követő, csak egybetűs szót (írógép stílus) tartalmazó
- Az összesen három vagy több visszaper (\) karaktert (escapelés) tartalmazó
- A *Unicode Replacement character*-t (U+FFFD) tartalmazó
- A „hullámos ő” (õ) és „hajtott ékezetes ű” (û) (karakterkódolás) betűket tartalmazó
- A *HTML escapelést* (<, >, 〹) tartalmazó

...mondatokat kiszűrtük.

A Mérés

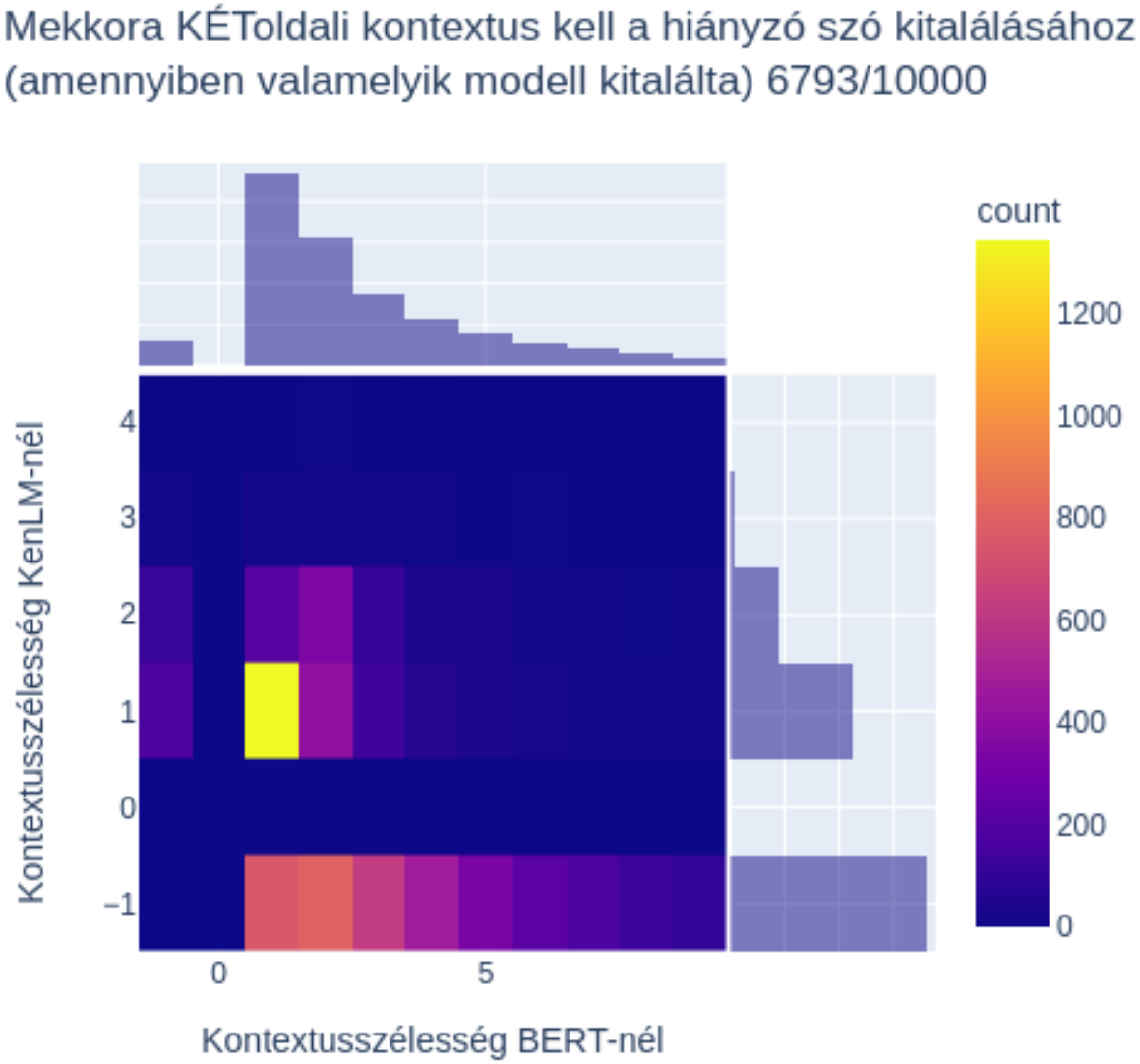
Modellek:

- KenLM (Heafield 2011)
- FastText (Bojanowski és tsai. 2017) – rosszul teljesített, kivettük
- huBERT (Nemeskey 2021a)
- Emberek (ELTE-DHs és NYTKs kollégák 2022) Köszönjük! :)

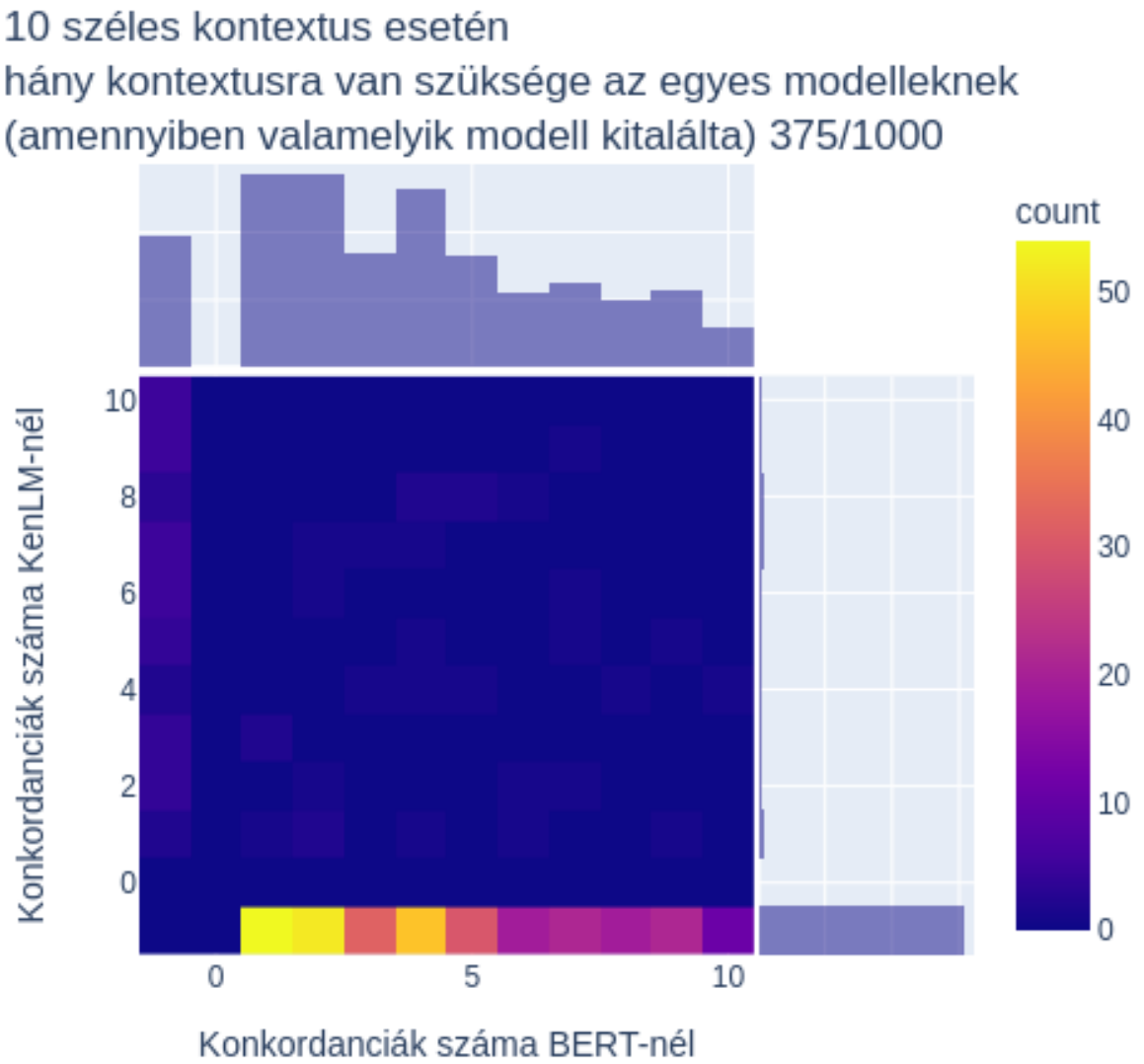
Mérések:

- A modellek teljesítménye **egyetlen bal-, jobb-, kétoldali kontextus** esetén
- A modellek teljesítménye a **kontextus mérete** alapján
- **Több kontextus** esetén javul-e a modell teljesítménye
- Emberi játékossal való összehasonlítás

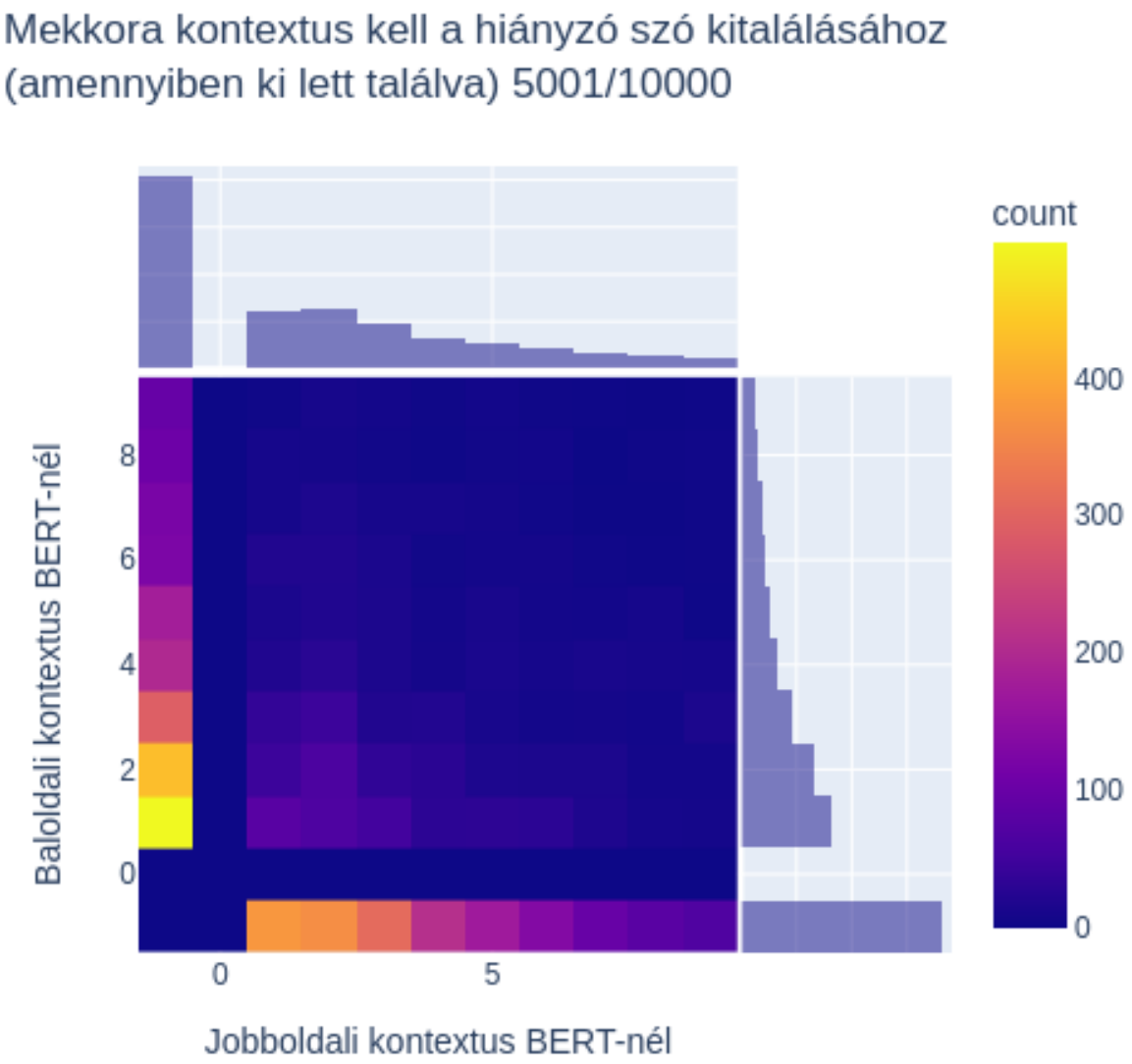
Kétoldali kontextus alapján (szélesség)



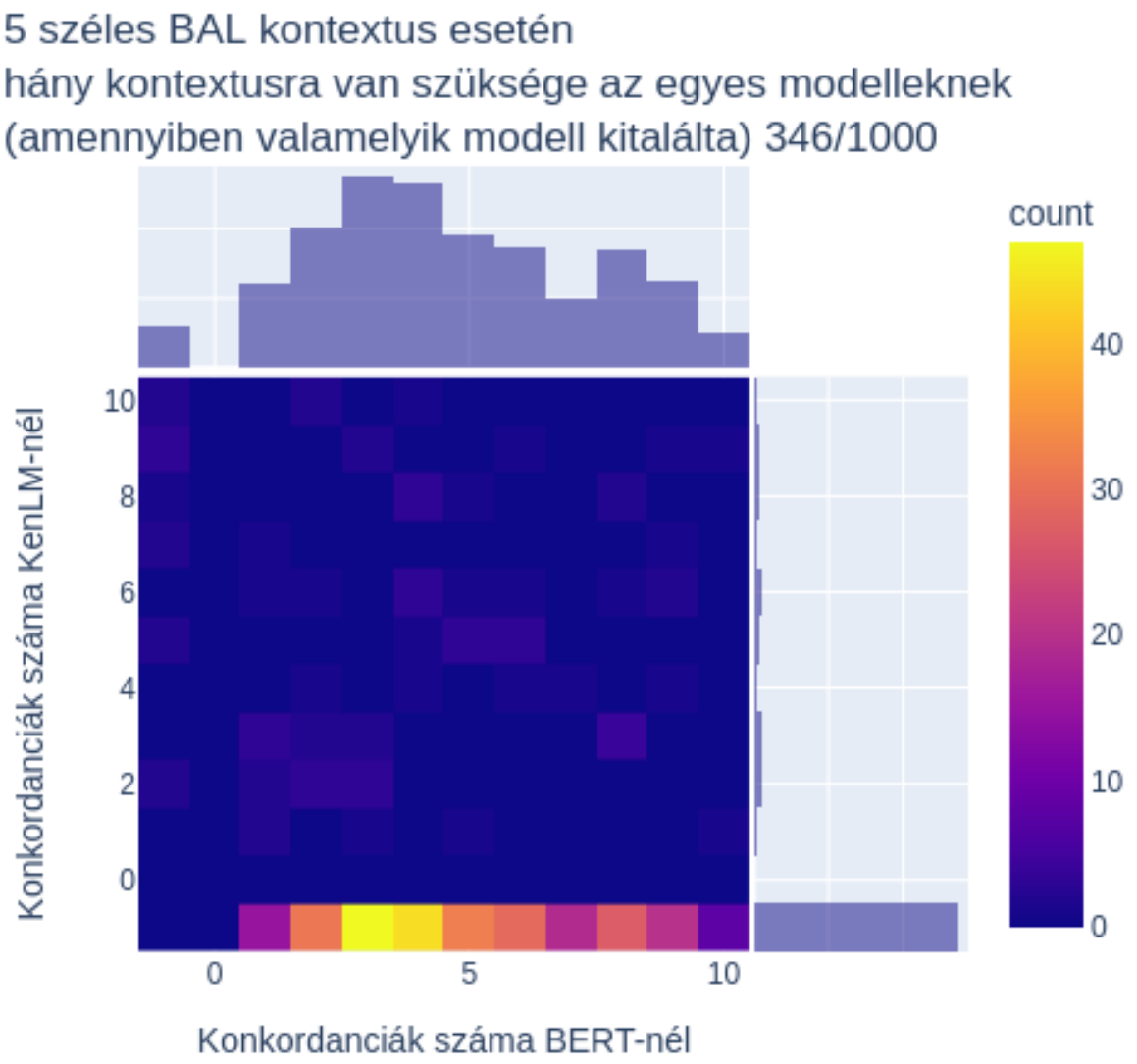
Kétoldali kontextus alapján (számosság)



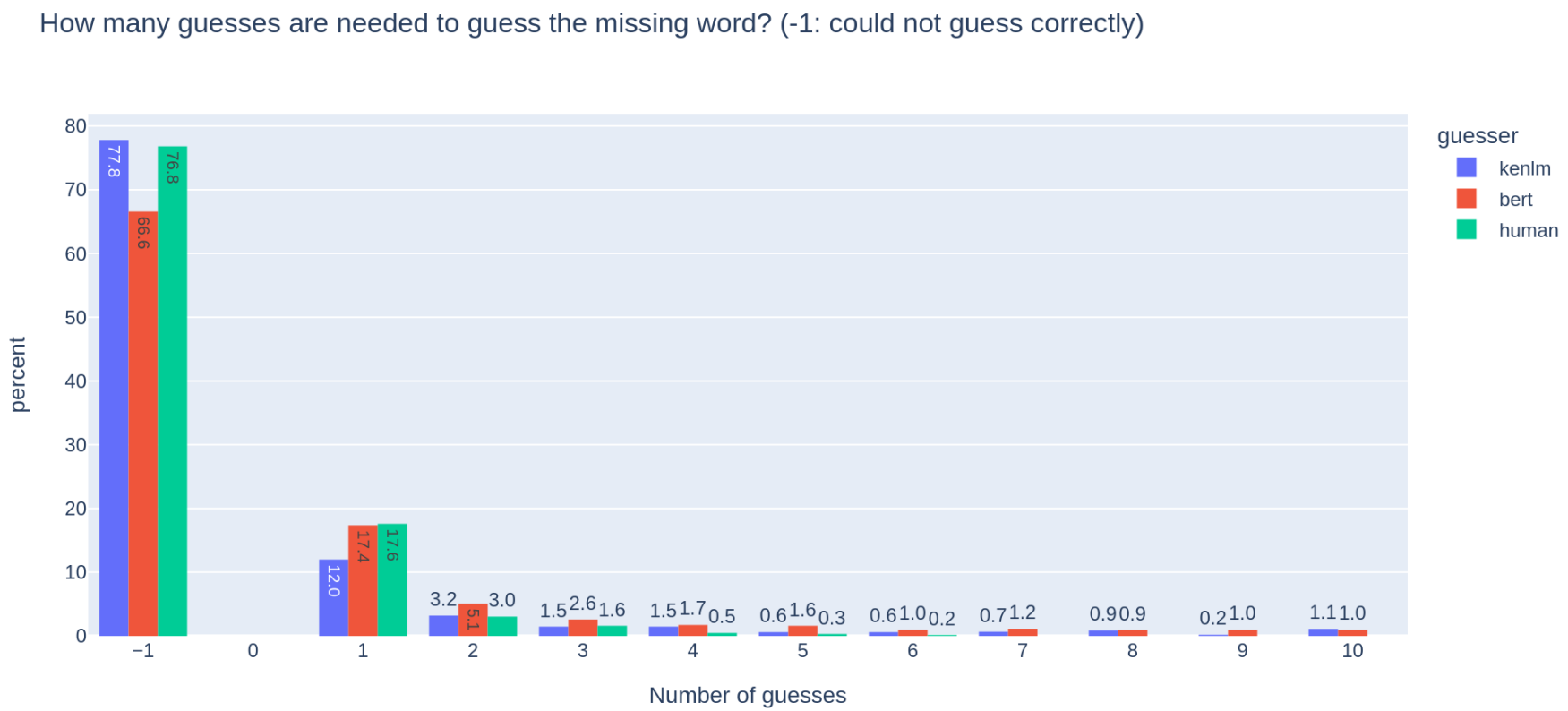
Egyoldali kontextus alapján (BERT, jobb vs. bal oldal)



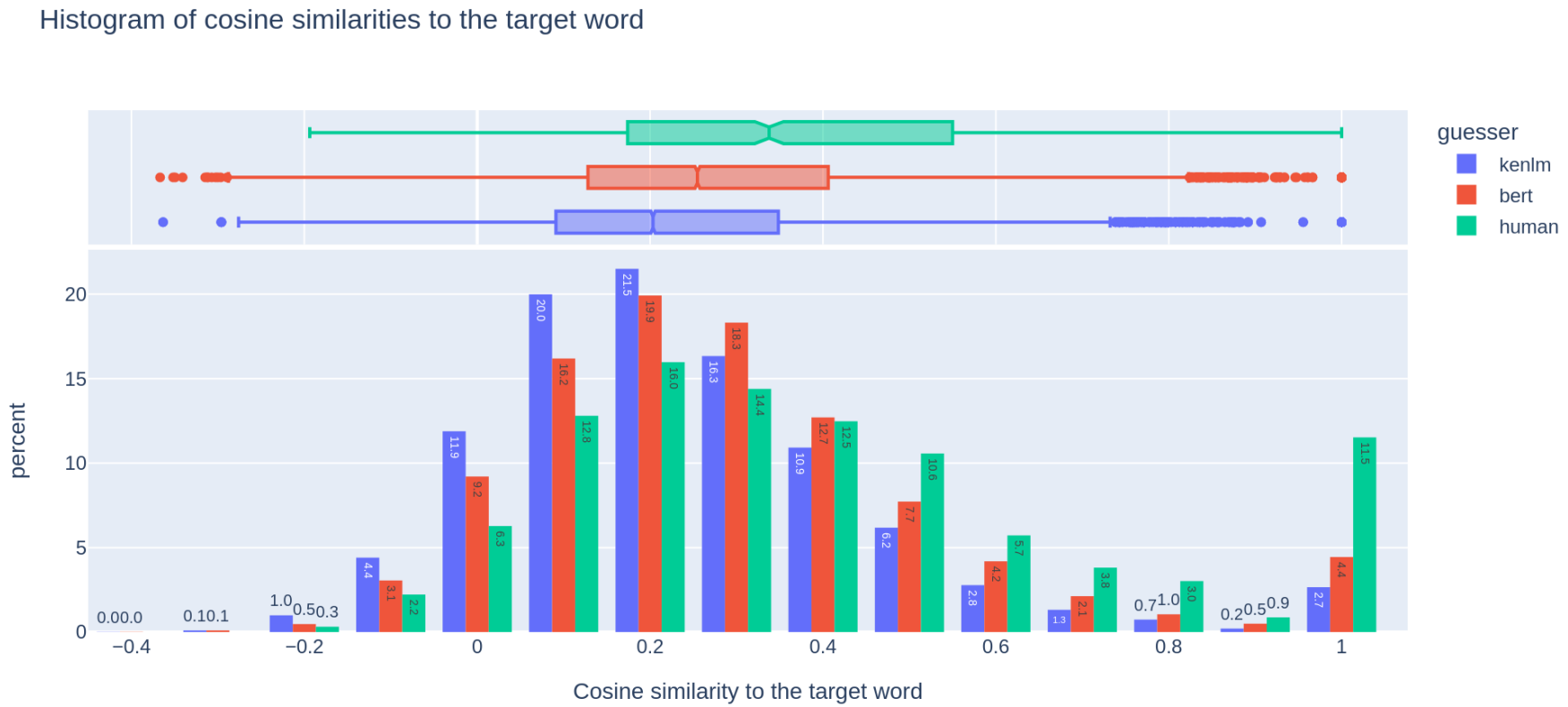
Egyoldali kontextus alapján (bal oldal, számosság)



Emberi kiértékelés (Hány példa kellett?)



Emberi kiértékelés (Mennyire hasonlít a rossz tipp?)



Összefoglalva

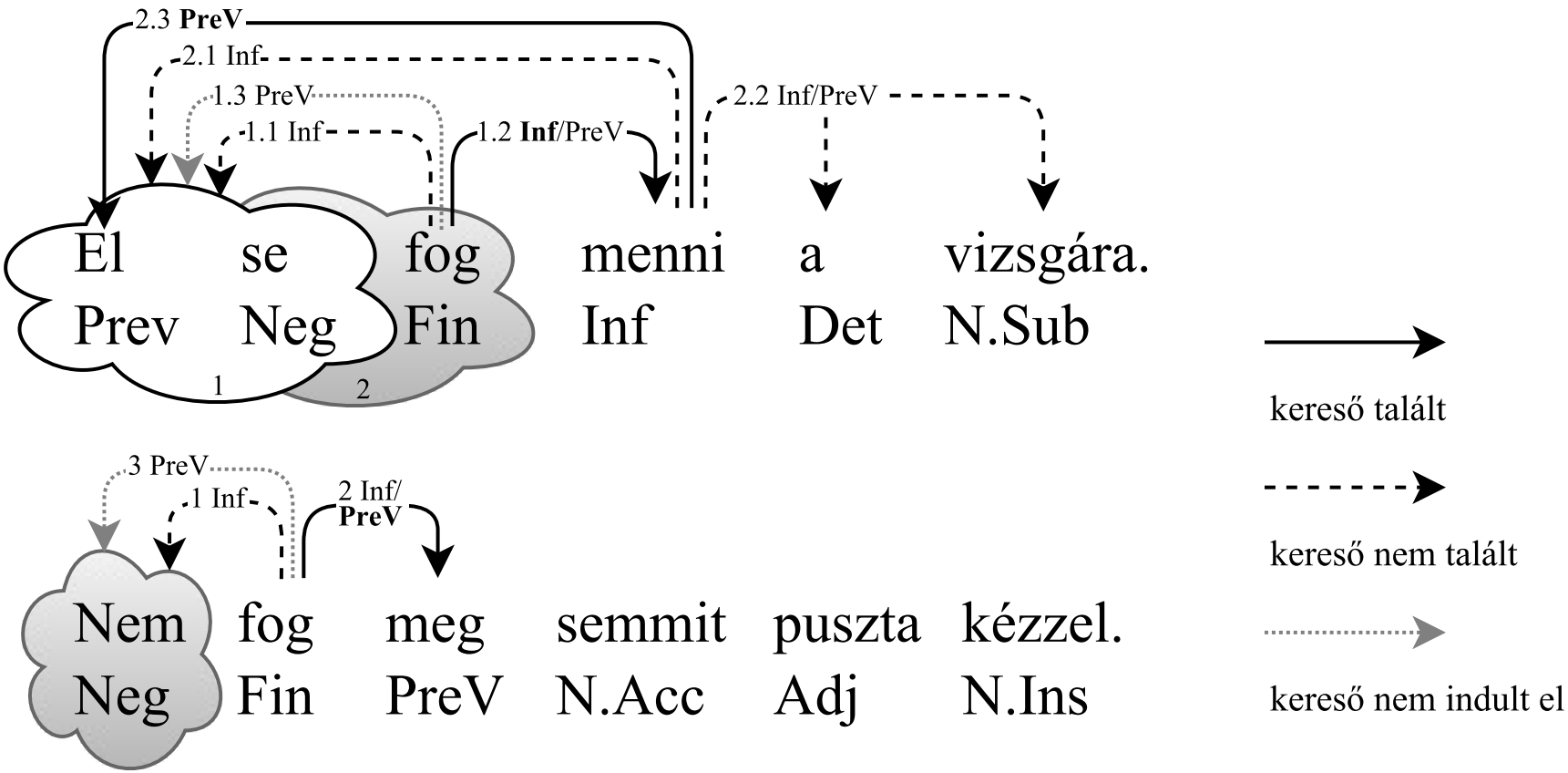
- Létrehoztunk:
  - Egy egyedül vagy gép ellen játszható Nyelvi Játékot (új, független kódbázis, LGPL 3.0)
  - Ami tesztkörnyezetként is használható `#pszicholingvisztika`
  - Egy a játék elvére épülő platformot, melybe könnyen befűzhetők új nyelvmodellek
  - Szabadon elérhető, automatikusan reprodukálható `#ReplicationCrisis` `#OpenScience`
- Kimutattuk:
  - Nem meglepő módon a BERT jobban teljesít a KenLM-nél :'(A **részletek** a legfontosabbak!
  - A BERT láthatóan képes többféle módon is kihasználni a hosszú relációkat
  - Nem elég csupán a bal vagy jobb kontextus, a szavak **kitalálhatóság szempontjából két majdnem diszjunkt halmazt alkotnak**
  - A kiértékelések túl *binárisak* (kvantitatívak) – láthatóan az emberek hasonló arányt érnek el a KenLM-hez képest, de *szemantikailag* jobbak a tippek

AnaGramma (Prószéky és Indig 2015)

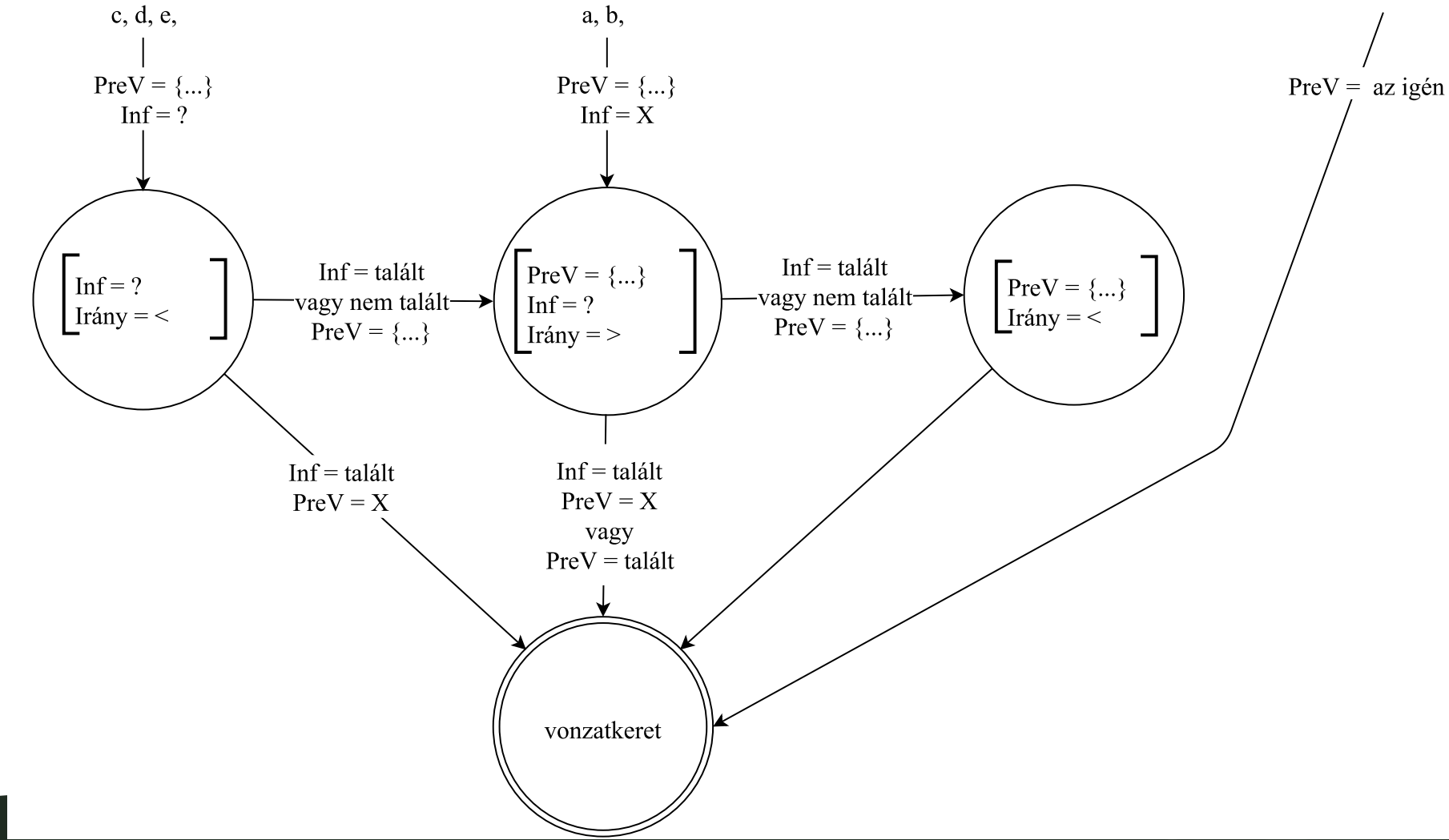
- Balról jobbra szavanként haladva (ahogy a beszéd elhangzik)
- Egy minimális három szó széles ablakkal jobb oldali irányban
- A szöveg többi részéről nem tudva (mivel még nem hangzott el)
- Kereslet-kínálat elven kezelve a szavak kapcsolatait
- Az összes lépést egyszerre végrehajtva (tokenizálás, szótövesítés, függőségi elemzés)
- Így kiküszöbölve a pipeline hatást
- „Ahogy az emberi elemző is csinálja…”

Tulajdonképpen véges állapotú automaták szó n-grammokon!

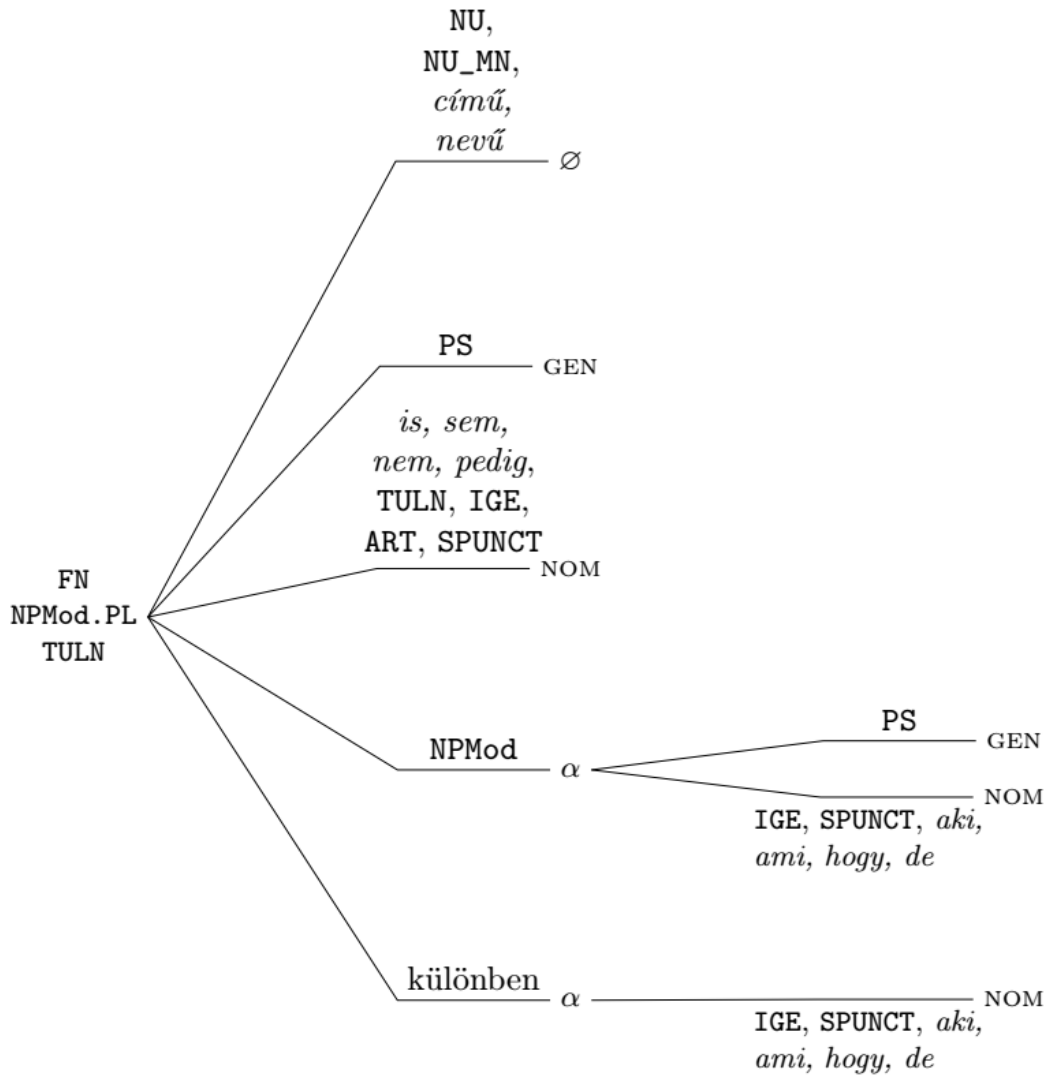
Az igekötő maximum n-gram távolságban van (Indig, Vadász és Kalivoda 2016)



Igekötő keresés az ablakban (a VFrame eljárás) (Vadász, Kalivoda és Indig 2017)



Nominativuszi 0 egyértelműsítése (Nom-or-What) (Ligeti-Nagy és tsai. 2018)



National Laboratory for Digital Heritage

Mozaik n-grammok (Indig, Laki és Prószéky 2016)

Definíció:
Olyan n-gram, ahol az egyes elemek specificitása eltérő (pl. szó, lemma, szófaji címke)

- Premissza: Minden tokennek van lemmája és szófaji címkéje és gyakorisága
- Vannak jellemzők, amik bár ugyanahhoz a konstrukcióhoz tartoznak, variálhatóak
- Az eredeti ötlet Sass Bálinttól származik:
  - Pl. van olyan trigram minta, hogy `lemma:esik szó [DEL]`
  - Ennek a variációi egységesen kezelendők
  - De hogy lehetne ilyeneket találni anélkül, hogy explicit rájuk keresnénk a korpuszban?
- Három kérdést kell „csak” megválaszolni tudni:
  - Hol a eleje és a vége a mintának?
  - Az egyes elemek melyik típusba sorolandók a maximális fedéshez?
  - Érdekes-e a minta, amit találtunk? (A „vessző, hogy a/az” a fenti kettő kritériumnak megfelel)
- Lásd még: Siklósi és Novák (2016a) és Siklósi és Novák (2016b) ugyanarról a konferenciáról (`word2vec` alapon)

Mozaik n-grammok (Indig, Laki és Prószéky 2016)

Vizsgált minta					Gyakoriság
[FN][NOM]	**[FN|NM][ACC]**	**lemma:mond**	**,**	**[KOT]**	**11918**
úr	azt	mondta	,	hogy	906
úr	azt	mondja	,	hogy	304
törvény	azt	mondja	,	hogy	176
miniszterelnök	azt	mondta	,	hogy	168
miniszter	azt	mondta	,	hogy	158
asszony	azt	mondta	,	hogy	126
államtitkár	azt	mondta	,	hogy	118
ember	azt	mondja	,	hogy	117
kormány	azt	mondja	,	hogy	108
gábor	azt	mondta	,	hogy	104
istván	azt	mondta	,	hogy	102
	viktor, lászló, péter, ferenc …				
túlzás	azt	mondani	,	hogy	86

Szekvenciális címkézés

- Tanítóanyag: szó-címke párok sorozatai (mondatokon belül)
- Minden szóhoz egy címkét kell rendelni egy meghatározott címkekészletből
- A címke sorozatok n-gramjait vesszük (**és még pár trükköt**)
- Általában zárójelezési probléma (és a címkékhez keressük a szót nem fordítva!)
  - Névelem felismerés (NER)
  - Maximális főnévi csoport meghatározása (NP chunking)
  - *Szentiment elemzés
  - *Szófaji egyértelműsítés (POS tagging) Ez nem zárójelezés!
- IOB címkézés variációi és lexikalizációjuk (Indig és Endrédy 2016; Indig 2017)
  - A szóeloszlás nem egyenletes (Zipf görbe) -> a címke-szó párok eloszlása sem
  - Átmenetileg a több címke jobb (v.ö. vektorreprezentációk)
  - A gyakori szavak címkéinek finomítása egyértelműsíti a ritka szavak címkéinek környezetét
  - Lásd még vektortér reprezentációk

Szekvenciális címkézés reprezentációi

szó	IOB1	IOB2	IOE1	IOE2	IOBES
These	I	B	I	E	S
include	O	O	O	O	O
,	O	O	O	O	O
among	O	O	O	O	O
other	I	B	I	I	B
parts	I	I	I	E	E
,	O	O	O	O	O
each	I	B	I	I	B
jetliner	I	I	E	E	E
’s	B	B	I	I	B
two	I	I	I	I	O
major	I	I	I	I	O
bulkheads	I	I	I	E	E
,	O	O	O	O	O

Szekvenciális címkézés reprezentációi (lexikalizáció)

	Lexikalizálatlan		**Lexikalizált**			
	Eredeti forma		***teljes***		***részleges (csak gyakori)***	
szó	**POS**	**IOB címke**	**POS**	**IOB címke**	**POS**	**IOB címke**
Rockwell	NNP	B-NP	NNP	NNP+B-NP	NNP	B-NP
said	VBD	O	VBD	O	VBD	O
the	DT	B-NP	the+DT	the+DT+B-NP	the+DT	the+DT+B-NP
agreement	NN	I-NP	NN	NN+I-NP	NN	I-NP

2. táblázat. IOB címke lexikalizáció. Molina és Pla (2002) **teljes** lexikalizációja alapján kialakított **részleges forma** (Indig és Endrédy 2016). A gyakori szavak címkéi megkapják a szót magát és a POS tagjét. A ritkák pedig csak a POS taget (teljes) vagy semmit (részleges).

Ezzel 1%-al javítható az angol főnévi csoport keresés SOTA. (v.ö. vektorreprezentáció)

A szófaji egyértelműsítés n-grammokkal

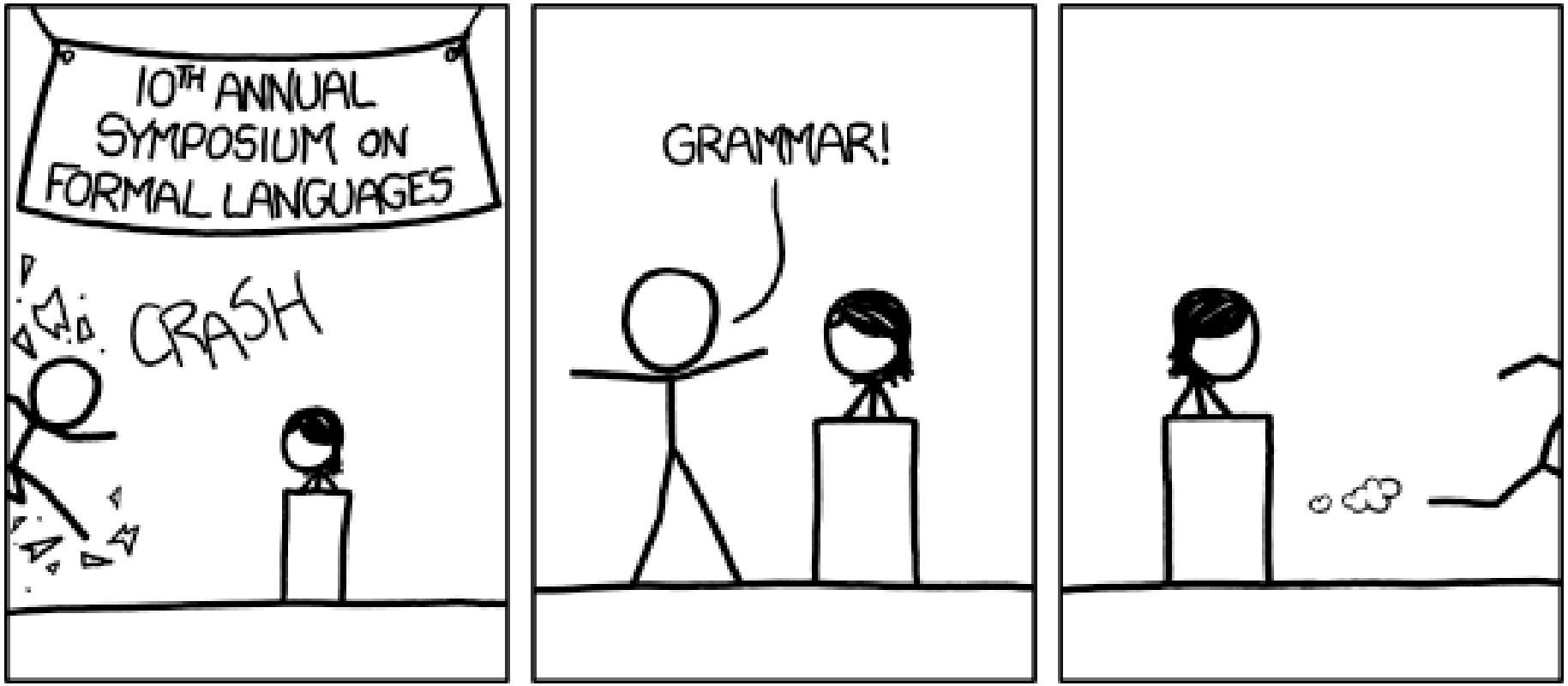
- Tanítóanyag: szó-címke páros angolra, magyarra a szótő is ideértendő!
- A szavakra tippelni valamit a végződésük alapján (suffix guesser) pl.
  - Hány betű vágható le a szó végéről?
  - Mit kell hozzáadni a végéhez, hogy megkapjuk a lemmát?
  - Hány betű vágható le a szó elejéről?
  - Mit kell hozzáadni az elejéhez, hogy megkapjuk a lemmát?
  - Ezt a négy jellemzőt rendeljük a címkékhez
- Adott egy szabályalapú morfológiai elemző mint segítség
  - Szabályalapú elemző: A gyakori szavak rendhagyóbbak (kevesen vannak sokszor)
  - Statisztika: A ritka szavak szabályosabbak (sokféle szó egy kaptafára húzható)
- **A trükkök fontosabbak a modellnél!** (Halácsy, Kornai és Oravecz 2007)
  - Kisbetűs és nagybetűs szavak külön (Az UTF-8-ban nem minden kis XOR nagybetűs)
  - Mondat elején kisbetűsítünk (A végén extra elem, jelzi a végét.)
  - A számokat reguláris kifejezésekkel entitásokra cseréljük (szótárméret csökkentés)

Morfológiai elemző és az n-grammok

- `emMorph` (Novák 2014; Novák, Siklósi és Oravecz 2016; Novák 2003)
- A beépített lexikonban található szavakra egy vagy több elemzést ad
- A lexikonban nem található szavakra (Out Of Vocabulary, OOV) pedig semmit
- Egyszerűen tesztelhető: `https://emmorph.herokuapp.com/`
- Morféma szintű felbontás: Felszíni, mély alakok és címkék sorozatai
  - Nem mindig triviális: *él|elme, *per|zselés, de tele|fon!
  - Mi lesz a képzőkkel? (adósságának -> ad?)
- Szótövesítés: A felbontás egyszerűsítése szótővé és szófaji címke pár(okk)á
  - Destruktív folyamat, nem reverzibilis :(
  - Sokszor önkényes: alatti: [/Adj][Nom], mosott: [/V][Pst.NDef.3Sg] vagy [/Adj][Nom]
  - Pedig tudja az igazságot! :)

A túlzott engedékenység tesztelhető random karakter n-gramokkal (fuzzing)

XKCD 1090



N-grammok további érdekes alkalmazásai

- A stilometria jelenleg legjobban magyarázható jellemzője (`stylo`)
  - A Zipf görbe bal harmada nyelv azonosításra jó,
  - a középső szerző azonosításra,
  - a jobb pedig téma azonosításra és kulcsszavazásra (v.ö. TF-IDF)
- A helyesírásnál az elütések jól javíthatóak vele (nem kell ismerni az adott szót hozzá)
- A betűcserék 11%-al lassítják csak az olvasást (Typoglycemia) (Rayner és tsai. 2006)
  - `https://www.mrc-cbu.cam.ac.uk/people/matt.davis/cmabridge/`
- Keresőknél (régen): Regex matching with trigram index
  - `https://swtch.com/~rsc/regexp/regexp4.html`
- Ebben is n-grammok vannak: Webaratás projekt a DH Lab-ban :)
  - (Többek között) HTML tag bigramok segítségével normalizáljuk a HTML-eket TEI XML-be
  - Korpuszkeresőkben mint a Sketch Engine: `https://sketchengine.elte-dh.hu/`
  - Összefoglaló előadás: `https://videotorium.hu/hu/recordings/35075`

Kitekintés (az n-grammok világából): Az Aszaló

A tipikus munkafolyamat:

- Beszerezzük a korpuszt (pl. webaratás)
- Nyelvileg elemezzük (pl. e-magyar/emtsv, HuSpaCy, Stanza, UDPipe)
- Valami saját elemzési réteget is hozzáadunk
- Kiraknánk egy jól hozzáférhető keresőrendszerbe, ha nem lenne bonyolult...

Aszaló to the rescue! `#DemocratiseScience`

- `https://github.com/dlazesz/aszalo`
- Tetszőleges TSV-ből könnyen konfigurálható webes keresőfelület (SQLite alapon)
- Ritka adat támogatással
- Példa a PrevCons adatbázis (Kalivoda 2021): `https://aszalo.herokuapp.com/`

Kitekintés (az n-grammok világából): e-magyar (emtsv)

„Itt ez a sok szuper eszköz, de hogy tudnám őket (együtt) használni?”

- A pipeline-ok elvárt tulajdonságai az xtsv példáján (Indig, Sass és Mittelholcz 2020)
  - A magja 1000 sor Python kód (rendkívül sokoldalú felhasználást tesz lehetővé)
- Egy öszvér fejléces függőleges TSV formátumot használ
  - Hasonló a Sketch Engine-ben látotthoz (vertical formátum)
  - És a CoNLL-U plus formátumhoz (A CoNLL-U szabvány továbbfejlesztése, terjedőben van)
- Az emtsv jelenleg 21 különböző modul több mint 30 féle alkalamazásánál tart
  - Gondolj egy feladatra és vagy van rá modul, vagy nemsokára lesz! :)
  - Folyamatosan dolgozunk az egész rendszer és új modulok fejlesztésén
  - `https://github.com/nytud/emtsv`
- Nem a leggyorsabb, (egyesek szerint) nem a legokosabb, de a miénk és működik! :)
  - `http://emtsv.elte-dh.hu:5000/`
- Saját modulok könnyen integrálhatók, így az egyéni kutatások nem vesznek el

Összefoglalás

- Az n-gram modellek a 2000-es évek mesterséges neurális hálói…
  - …és a 2020-as évek hipszter szakálla :D
- Bár mára majdnem kihaltak, a trükkök tovább élnek az új modellekben
- Sokkal jobban magyarázhatóak és rugalmasabbak,
- kevesebb erőforrást (gép és tanítóanyag) igényelnek,
- cserébe néhol kicsit butábbak mint a jelenlegi modellek.
- Még mindig tartogatnak meglepetéseket! :)
- Pont ezért van piacuk (nem fontos mi van a modell alatt, ha működik)
  - UNKP, szakgoldozat, PhD konzultálást, code review-t vállalok
  - Gyakornoki és junior pozíciók nyitottak a DH Labban az érdeklődőknek

Vizsgált minta				Gyakoriság
lemma:köszön	**a**	**lemma:figyelem**	**.**	**14582**
köszönöm	a	figyelmüket	.	7654
köszönöm	a	figyelmet	.	6762
köszönöm	a	figyelmét	.	142
köszönjük	a	figyelmüket	.	32
köszönjük	a	figyelmet	.	12
köszöni	a	figyelmüket	.	5
köszönöm	a	figyelmünket	.	3
köszönöm	a	figyelmeteket	.	2
köszöni	a	figyelmet	.	1
köszönjük	a	figyelmét	.	1

Köszönöm a XXXXXXXXX!

https://dh-lab.hu/
https://elte-dh.hu/
https://github.com/elte-dh
https://zenodo.org/communities/elte-dh

https://word-concordance-game.herokuapp.com/
https://sketchengine.elte-dh.hu/
https://aszalo.herokuapp.com/
https://emmorph.herokuapp.com/
http://emtsv.elte-dh.hu:5000/

Hivatkozások I

- Bojanowski Piotr és tsai., „Enriching Word Vectors with Subword Information”, *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. old., doi: 10.1162/tacl_a_00051, url: https://aclanthology.org/Q17-1010.
- Chen Stanley F. és Joshua Goodman, „An empirical study of smoothing techniques for language modeling”, *Computer Speech & Language* 13.4 (1999), 359–394. old., issn: 0885-2308, doi: https://doi.org/10.1006/csla.1999.0128, url: https://www.sciencedirect.com/science/article/pii/S088523089990
- Devlin Jacob és tsai., „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019. jún., 4171–4186. old., doi: 10.18653/v1/N19-1423, url: https://aclanthology.org/N19-1423.

Hivatkozások II

- Elkins Katherine és Jon Chun, „Can GPT-3 Pass a Writer's Turing Test?", *Journal of Cultural Analytics* 1.1 (2020), 17212. old.
- Firth J., „A Synopsis of Linguistic Theory 1930-1955", *Studies in Linguistic Analysis*, reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow., Philological Society, Oxford, 1957.
- Halácsy Péter, András Kornai, László Németh és tsai., „Creating Open Language Resources for Hungarian", *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal: European Language Resources Association (ELRA), 2004. máj., url: `http://www.lrec-conf.org/proceedings/lrec2004/pdf/525.pdf`.
- Halácsy Péter, András Kornai és Csaba Oravecz, „Poster paper: HunPos – an open source trigram tagger", *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic: Association for Computational Linguistics, 2007. jún., 209–212. old., url: `https://aclanthology.org/P07-2053`.

Hivatkozások III

- Harris Zellig S., „Distributional Structure”, *WORD* 10.2-3 (1954), 146–162. old., doi: 10.1080/00437956.1954.11659520.
- Heafield Kenneth, „KenLM: Faster and Smaller Language Model Queries”, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland: Association for Computational Linguistics, 2011. júl., 187–197. old., url: https://aclanthology.org/W11-2123.
- Indig Balázs, „Less is more, more or less... Finding the optimal threshold for lexicalization in chunking”, *Computación y Sistemas* 21.4 (2017), 637–646. old.
- Indig Balázs és István Endrédy, „Gut, besser, chunker–selecting the best models for text chunking with voting”, *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2016, 409–423. old.
- Indig Balázs, László János Laki és Gábor Prószéky, „Mozaik nyelvmodell az AnaGramma elemzőhöz”, *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2016)* (2016).

Hivatkozások IV

- Indig Balázs, Bálint Sass és Iván Mittelholcz, „The xtsv Framework and the Twelve Virtues of Pipelines”, English, *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2020. máj., 7044–7052. old., isbn: 979-10-95546-34-4, url: https://www.aclweb.org/anthology/2020.lrec-1.871.
- Indig Balázs, Noémi Vadász és Ágnes Kalivoda, „Decreasing Entropy: How Wide to Open the Window?”, *International Conference on Theory and Practice of Natural Computing*, Springer, 2016, 137–148. old.
- Kalivoda Ágnes, „Az igekötők produktív kapcsolódási mintái”, *Argumentum* 17 (2021), 56–82. old., doi: https://doi.org/10.34103/ARGUMENTUM/2021/4.
- Ligeti-Nagy Noémi és tsai., „Nulla vagy semmi? Esetegyértelműsítés az ablakban”, *Vincze Veronika (szerk.): XIV. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, Informatikai Intézet* (2018), 25–37. old.
- Mikolov Tomas és tsai., „Efficient Estimation of Word Representations in Vector Space”, *Proceedings of Workshop at ICLR* 2013 (2013. jan.).

Hivatkozások V

- Molina Antonio és Ferran Pla, „Shallow parsing using specialized hmms”, *The Journal of Machine Learning Research* 2 (2002), 595–613. old.
- Nemeskey Dávid Márk, „Introducing huBERT”, *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, 2021, 3–14. old.
- — „„Natural language processing methods for language modeling”, dissz., Doctoral School of informatics, Eötvös Loránd University, Faculty of Faculty of Informatics, 2021.
- Novák Attila, „A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation”, angol, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, szerk. Nicoletta Calzolari és tsai., Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, isbn: 978-2-9517408-8-4.
- — „„Milyen a jó Humor? [What is good Humor like?]”, *I. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged: SZTE, 2003, 138–144. old.

Hivatkozások VI

- Novák Attila, Borbála Siklósi és Csaba Oravecz, „A New Integrated Open-source Morphological Analyzer for Hungarian”, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia: European Language Resources Association (ELRA), 2016. máj., 1315–1322. old., url: https://aclanthology.org/L16-1209.
- Prószéky Gábor és Balázs Indig, „Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel”, *Alkalmazott nyelvtudomány* 15.1-2 (2015), 29–44. old.
- Rayner Keith és tsai., „Raeding Wrods With Jubmled Lettres: There Is a Cost”, *Psychological Science* 17.3 (2006), PMID: 16507057, 192–193. old., doi: 10.1111/j.1467-9280.2006.01684.x, eprint: https://doi.org/10.1111/j.1467-9280.2006.01684.x, url: https://doi.org/10.1111/j.1467-9280.2006.01684.x.
- Rehurek Radim és Petr Sojka, „Gensim–python framework for vector space modelling”, *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

Hivatkozások VII

- Siklósi Borbála és Attila Novák, „Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra”, *Tanács Attila, Varga Viktor, Vincze Veronika (szerk.). XII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, Informatikai Intézet* (2016), 3–14. old.
- — „„Közeli rokonunk, az autó”, *Tanács Attila, Varga Viktor, Vincze Veronika (szerk.). XII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, Informatikai Intézet* (2016).
- Vadász Noémi, Ágnes Kalivoda és Balázs Indig, „Ablak által világosan–Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével”, *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)* (2017), 26–27. old.