

Mentális konstrukciók adatorientált azonosítása korpuszban

a Mozaik módszer és a kapcsolódó eljárások segítségével

Indig Balázs

ELTE IK Mesterséges Intelligencia tanszék

2025. március 25.



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Amiről szó lesz...

- Motiváció: Mik a nyelvi minták és a konstrukciók?
- A Mozaik módszer működése
- Kiegészítő stratégiák
- Potenciális alkalmazási területek



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Motiváció: szemlélet

Be conservative in what you generate, be liberal in what you accept.

(Robustness principle, Jon Postel)

A flaw can become entrenched as a de facto standard. Any implementation of the protocol is required to replicate the aberrant behavior, or it is not interoperable.

(Martin Thomson és David Schinazi)



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Motiváció: Generatív eszköztár

- A generatív nyelvi eszköztár kiválóan alkalmas nyelvi produkcióra (lásd nyelvmodellek)
- Szabályozott kimenet állítható elő, ha
 - fel tudjuk sorolni az összes kimenetet
 - vagy ismerjük a szabályokat
- De honnan jönnek a szabályok?
- Mi történik, ha szintaktikailag nem helyes a bemenet?



Motiváció: Konstrukciók

- Szabályok, amik véges darab egymás utáni elemből (n-gram) állnak, amelyek többé kevésbé kötöttek
- Egy elvi spektrumot alkot
 - Mi számít egy elemnek?
 - Mi a kötöttség mértéke?
 - A szabad elemek helyére mit lehet beírni? (pl. szófajok, egész szerkezetek)
- De honnan jönnek a szabályok?
- Mi történik, ha szintaktikailag nem helyes a bemenet?



Motiváció: Szabályok



Szabályok nélkül olyanok lennénk mint az állatok



A klub első szabálya:
senkinek egy szót se, hogy honann jönnek a szabályok



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Motiváció: Korpuszok

- Valóban leírt/elhangzott szövegek tárháza (v.ö. Bábeli könyvtár)
 - Szerkesztett, szerkesztetlen, írott, hangzó, stb.
 - Metaadatok: Nyelv, keletkezés ideje, zsáner, stb.
- **Feltételezés:** "Végtelen" szöveget "elolvasva" meg lehet tanulni (használni) a nyelvet
 - Nagy nyelvmodellek: Challenge accepted! :)
- **Limitációk:** Az emberek percenként 250 szót tudnak olvasni maximum
 - 60 perc 1 óra, 24 óra egy nap, 365 nap egy év...
 - **120 év alatt folyamatosan így olvasva 15,768 milliárd szó**
 - MNSz2 UD (748,55 millió szó), Webkorpusz 2.0 (8,986 milliárd szó), ELTE.DH korpusz (758,99 millió szó)
- Friss eredmény: A magyar szófaji egyértelműsítéshez és szótövesítéshez 125 000 szó elég
 - (Dömötör, Indig és Nemeskey 2025)
- **Mégsincs szükség annyi szövegre?**



Motiváció: Korpuszminták

- A korpuszban elérhető jellmezők alapján (pl. szóalak, szótő, szófaji címke, függőségi relációk)
- Lekérdezéseket definiálunk (pl. CQL (Jakubíček és tsai. 2010))
- A cél:
 - Az adott célnak megfelelő (hasznos) elemek leírása
 - Konstrukciók definiálása, valós példákkal való alátámasztása
 - A talált konstrukciók felett műveletek definiálása
- A lekérdezések átlagos formája: általában egymást követő szavak, hol jobban hol kevésbé specifikálva
 - Valahonnan ismerős ez...
- De honnan jönnek a szabályok?
 - A nyelvész intuíciójából, aki a lekérdezést megírja. De mi van azzal, amire nem gondol?
 - Nem lehetne objektíven, az intuíciótól nem befolyásolva mintákat találni?



Motiváció: Vektoros nyelvmodellek

- Működés nagy vonalakban:
 1. Veszünk egy nagy korpuszt
 2. Véletlenszerűen kijelölünk szavakat
 3. Letakarjuk őket
 4. A modell a kontextus alapján ki kell, hogy találja őket
 5. Megerősítéses tanulással szükség szerint ismételjük
- Ehhez sok dimenziós vektorokat csinál a szavakból (vagy szórészletekből)
- Csökkenteni kell a szótár méretét technikai okokból:
 - A webkorpusz 2.0 szótármérete: 67 285 480 szóalak (45 131 081 lemma)
 - Sok a zaj, ami szisztematikus és gyakori, de a ritkák között is bőven van jó szó
 - Városi legenda: Arany János szókincse: kb. 16 000 szót (kb. 60 000 alakot)
 - Ha minden szó kap azonosítót (4 bájt) „minden szó kapcsolata minden szóval” (szomszédsági mátrix): 15 GB
 - A dupla pontosság 8 bájt lenne (32 bit floating point): 30 GB



Motiváció: Vektoros nyelvmodellek



„Emoji analízis”: South Park S20E09 „Nem vicces” (2016)



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Motiváció: Vektoros nyelvmodellek

Láttam egy _____ -t.
A _____ a fán volt.
A _____ épp evett.
Tegnap egy másik _____ , amit láttam, ami káromgott, de ez nem.

Fókuszáljunk a hiányzó elemre, aztán a kontextusokra.

- A szavak jelentése karakterizálható a környezeteikből
- A vektorok a környezetet kódolák. (Technikai okokból.) De mi lenne, ha nem kellene kódolni?
- A vektorok azért jók, mert lehet rajtuk műveleteket végezni, hogy analogikus relációkat kapjunk
 - Tehát az analogikus relációkra szükségünk van, a vektorokra pedig nem



Motiváció: Vektoros nyelvmodellek



Új játék Szabad a gazda

Tipp Kérek még egy mondatot

...osztódott, és mindegyik részen xxxxxxxx csillagzat szerkesztetett ősz...
...jesztett két szárnya csuklóján xxxxxxxx /Deneb)4)1; eltátott ajkaira, ...
...nagy bólnak négy szegeletében xxxxxxxx edény van elásva, tele pénzze...
...zer is találhatik nyarantszak xxxxxxxx kasban; - végre hogy az anya,...
...lította Miskeit, ez pedig tsak xxxxxxxx Syllabáju szókkal igen rövi...
...vány, és néptelen; imitt-amott xxxxxxxx par [!] szeretseny Familiát...
...sairól, máglyára kárhoztatnád? xxxxxxxx példány ára három forint. De ...
...en fogtunk ülést, mindegyikünk xxxxxxxx ...
...szerezni, mint a multba vetett xxxxxxxx pillantás, mely elénk varázso...

Korábbi tippek
hosszas
mély

Megfejtés: egy-egy

Demó: <https://elte-dh.hu/szojatek/>

A saját implementációnk: Indig és Lévai (2023)

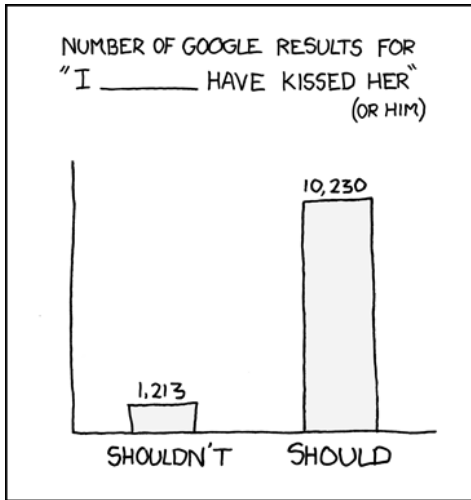


ELTE IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Motiváció: Vektoros nyelvmodellek



Regrets (<https://xkcd.com/458/>)



DEBRECENI
UNIVERSITÁS
IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

A Mozaik módszer működése



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

A Mozaik módszer működése: mozaik példa

| | |
|----------------|--|
| Példány | Egy ideig nem tudtuk eldönteni |
| Szóalak 5-gram | egy ideig nem tudtuk eldönteni |
| Lemma 5-gram | egy idő nem tud eldönteni |
| POS-tag 5-gram | [/Det][Art NDef] [/N][Ter] [/Adv] [/V][Pst.Def.1Pl] [/Inf] |
| Mozaik 5-gram | [/Det][Art NDef] [/N][Ter] nem lemma:tud eldönteni |

- Melyik a jó absztrakciós szint?
- Hogyan találjuk meg?
- Hogyan ellenőrizzük, hogy tényleg jó-e?



A Mozaik módszer működése: mozaik példa

| | | | | | |
|----------------------|-------------|-----------|--------------------|---------------|---------|
| szóalak | Sok | szó | esett | kutyákról | . |
| lemma | sok | szó | esik | kutya | . |
| szófaji címke | [/Num][Nom] | [/N][Nom] | [/V][Pst.NDef.3Sg] | [/N][Pl][Del] | [Punct] |

Esik szó -ról szerkezet (Sass Bálint)

- Adott hosszúság esetén az egyes elemek várt absztrakciós szintjét keressük
- Cél: A maximális elemszámú minta fedése
- Automatikusan visszaírható CQL lekérdezéssé (validálható)
- (A minta elejének és végének megtalálása más részfeladat. → most még manuális)



A Mozaik módszer működése: a megfelelő absztrakciós szint

| példák | mozaikok |
|--------------------------|--|
| A szél lefújta a levelet | <ol style="list-style-type: none">1. [/Det art.Def] szél [Pst.Def.3Sg] [/Det art.Def] [/N][Acc]2. [/Det art.Def] szél lemma:lefúj [/Det art.Def] [/N][Acc]3. [/Det art.Def] [/N][Nom] lefújta [/Det art.Def] [/N][Acc]4. A [/N][Nom] [/V][Pst.Def.3Sg] a [/N][Acc] ... |
| A bíró lefújta a játékot | <ol style="list-style-type: none">1. [/Det art.Def] bíró [Prs.NDef.3Sg] [/Det art.Def] [/N][Acc]2. [/Det art.Def] bíró lemma:lefúj [/Det art.Def] [/N][Acc]3. [/Det art.Def] [/N][Nom] lefújta [/Det art.Def] [/N][Acc]4. A [/N][Nom] [/V][Pst.Def.3Sg] a [/N][Acc] ... |



A Mozaik módszer működése: a megfelelő absztrakciós szint

| | | | | |
|-----|---|--------------------|--------|-----------|
| 46 | | [/V][Prs.NDef.3S] | [/Det] | [/N][Nom] |
| → | 4 | meggyógyul | [/Det] | [/N][Nom] |
| → | 4 | lemma:meggyógyul | [/Det] | [/N][Nom] |
| → | 4 | lemma:meg +van | [/Det] | [/N][Nom] |
| → | 4 | felel | [/Det] | [/N][Nom] |
| → | 3 | mond | [/Det] | [/N][Nom] |
| → | 2 | [/V][Prs.NDef.3Sg] | [/Det] | lábado |
| → | 2 | [/V][Prs.NDef.3Sg] | [/Det] | Isten |
| ... | | | | |

a példa forrása: Bajzát, Indig és Kalivoda (2024)

- Az azonos példákra illeszkedők közül csak a legkevésbé absztrakt marad (meggyógyul vs. lemma:meggyógyul)
- A „példarészhalmozatra” illeszkedő elemeket behúzással jelöljük



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

A Mozaik módszer működése: az algoritmus

- Az algoritmus (Indig, Laki és Prószéky 2016; Indig 2017):
 1. Legenerálunk minden lehetséges esetet
 2. Összeszámoljuk az előfordulásait
 3. Az azonos gyakoriságúakból kiválasztjuk a legkevésbé absztraktot, **mely ugyanazt a mintahalmazt fedi le**
 4. A mintahalmazokat összehasonlítjuk, a részhalmazokat másodlagosnak jelöljük
- Ugyanez működik szósákokra is apró változtatásokkal
- Kód: <https://github.com/bajzattimi/Research-of-infinitive-structures-related-to-the-modal-semantic-domain>
- Érdekesség: <https://github.com/dlazesz/mosaic-n-gram-benchmark>



A Mozaik módszer működése: erősségek és korlátok

Erősségek:

- Nincs szükség szintaxisra (szintaktikai elemzőre)
- A megfelelő absztrakciós szint megtalálásában jó

Korlátok:

- Függ a felhasznált jellemzőktől (pl. címkekészlet finomsága, POS tagger pontossága)
- Csak egész szavakon működik (a szófaji címkét ezért használjuk)
- Az eltérő hosszú n-grammokat nem tudja összehasonlítani

A korlátok kiküszöbölhetők a kiegészítő stratégiákkal

Bővebben: Indig és Bajzát (2024)



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Kiegészítő stratégiák (szerszámoszláda-filozófia)



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Kiegészítő stratégiák: a stratégiák

- Bizonyos absztrakciók fölöslegesek (pl. az „is” szótöve, a határozott névelő szóalakja)
- A szófaji címkék szükség szerint összevonhatóak (a másik irány már nehezebb :D)
 - pl. Kell-e a többesszám jele, a szám/személy, stb.?
 - Egy ellenőrizhető megoldás: Indig és Bajzát (2023)
- A szavakat fel lehet bontani a morfémaikra (thx to emMorph (Novák, Siklósi és Oravecz 2016))
 - be-kap|nyom|tol|csap egy N.ACC
- Bizonyos szavakat össze lehet vonni egy elemmé
 - Szabályalapú főnévcsoport tömörítés esetraggá (Nem NP chunking!)
 - Csak a jól kezelhető esetek (címke alapján)
 - Az eredeti szóalak megmarad mint jellemző, hogy az azonos alakú (gyakori idiomatikus) elemeket felismerjük
- **A szent grál:** A különböző hosszúságú elemek összevethetősége interpretálhatóan (a kollokációk nem jók)



Kiegészítő stratégiák: főnévi csoport tömörítés példa

| N | Frequency | Example | | | | |
|---|-----------|------------------|-------------|-----------------|------------|-----------|
| 3 | 1688 | | | [/N][Acc] | lemma:akar | [/V][Inf] |
| 4 | 1103 | | [/Adj][Nom] | [/N][Acc] | lemma:akar | [/V][Inf] |
| 5 | 1665 | [/Adj][Nom] | [/Adj][Nom] | [/N][Acc] | lemma:akar | [/V][Inf] |
| 5 | 1365 | [/Det Art.Def] | [/Adj][Nom] | [/N][Acc] | lemma:akar | [/V][Inf] |
| 5 | 997 | [/Det Art.Def] | [/N][Nom] | [/N][Poss][Acc] | lemma:akar | [/V][Inf] |

a példa forrása: Indig és Bajzát (2023)

- Mindegyik későbbi eset egyszerűen redukálható az első sorra a nyelvtan megsértése nélkül.
- Az „alanyeset” főnév okoz problémákat (birtokos, névutó, stb.)
- A részes esetű birtokos szabad szórendű, explicit nem kezeljük
- A szófai címkék nem mindig egyértelműek (pl. a terv alapján → Post vs. Supe)



Esettanulmány: főnévi igeneves szerkezetek

(Indig és Bajzát 2023; Bajzát, Indig és Kalivoda 2024; Bajzát és Indig 2024; Indig és Bajzát 2024; Bajzát és Indig 2025)



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

| Alaptag | POS | Minta (MNSz2) | Szűrt (MNSz2) | % (MNSz2) | Minta (Webkorpusz2) | Szűrt (Webkorpusz2) | % (Webkorpusz2) |
|------------|-----|------------------|------------------|--------------|------------------------|------------------------|--------------------|
| Akar | V | 610 836 | 419 324 | 68,65 | 650 000 | 518 123 | 79,71 |
| Bír | V | 22 191 | 15 387 | 69,34 | 179 846 | 112 112 | 62,34 |
| Hajlandó | A | 48 267 | 36 334 | 75,28 | 272 806 | 179 330 | 65,74 |
| Képes | A | 134 843 | 86 833 | 64,40 | 650 000 | 462 001 | 71,08 |
| Képtelen | A | 48 036 | 14 424 | 30,03 | 164 909 | 104 274 | 63,23 |
| Kíván | V | 192 678 | 139 498 | 72,40 | 650 000 | 391 413 | 60,22 |
| Mer | V | 63 729 | 39 177 | 61,47 | 473 966 | 278 887 | 58,84 |
| Szeret(ne) | V | 484 448 | 278 834 | 57,56 | 650 000 | 448 324 | 68,97 |
| Tud | V | 675 000 | 466 863 | 69,16 | 650 000 | 540 175 | 83,10 |

- MNSz2 (SUM): 2 097 149 → 1 496 674 (71.37%)
- Webkorpusz2(SUM): 4 341 527 → 3 034 639 (69.90%)



Esettanulmány: főnévi igeneves szerkezetek (Indig és Bajzát 2024)

- A Magyar Nemzeti Szövegtár 2.0 (Oravecz, Váradi és Sass 2014) és a Webkorpusz 2.0 (Nemeskey 2020) közismert korpuszok, de számos különbség van köztük, amit igyekeztünk egységesíteni
 - Mindkettő elérhető korpuszlekérdező felületről → A minták mérete felülről korlátos
 - Eltérő a cimkekészletük → Újraelemezve e-magyarral
 - Zaj, duplikáció, stb. → Előszűrés (lásd előző dia)
- A részes esetű birtokosok ritkák → Vajon hol lehetnek, ha nem itt? (Egyelőre nem foglalkozunk velük)
 - MNSz2: a példák 0,36%-a (11 típus)
 - Webkorpusz2: a példák 0.59%-a (27 típus)
- Az egyedi típusok száma jelentősen lecsökkent (25-ös küszöbérték) → **Nyelvészeti értelemben kezelhető**
 - MNSz2: a példák 54.43%-ra (2593 típus)
 - Webkorpusz2: a példák 60.83%-ra (5573 típus)
- Az $N \geq 5$ példák az egész mennyiség 25.92%-át (MNSz2) és 27.38%-át (Webkorpusz2) teszik ki
 - Az egyedi mintáknak pedig a 47 és 44%-át



A leggyakoribb jelöltek hosszváltozása MNSz2(M) és Webkorpusz2 (W)

| | N = 3 | | N = 4 | | N = 5 | | N = 6 | | N = 7 | |
|----------|--------|--------|--------|--------|-------|-------|-------|-------|-------|-------|
| | M | W | M | W | M | W | M | W | M | W |
| Akar | 103,26 | 103,06 | 117,22 | 109,87 | 52,79 | 49,05 | 33,33 | 34,67 | 0 | 0,99 |
| Bír | 100 | 100 | 115,38 | 122,78 | 57,14 | 63,73 | 0 | 41,67 | 0 | 65,28 |
| Hajlandó | 100 | 112,50 | 200 | 163,41 | 87,10 | 60,55 | 85,71 | 76,47 | 0 | 5,00 |
| Képes | 108,33 | 103,92 | 268,42 | 229,63 | 63,33 | 47,64 | 5,26 | 21,95 | 0 | 3,26 |
| Képtelen | 125,00 | 107,14 | 150 | 116,67 | 25,00 | 52,94 | 0 | 5,88 | - | 0 |
| Kíván | 148,15 | 116,95 | 225,49 | 217,56 | 23,76 | 29,64 | 29,41 | 37,82 | 0 | 0,83 |
| Mer | 100 | 110,94 | 107,41 | 107,83 | 74,42 | 68,07 | 0 | 30,77 | - | 0 |
| Szeret | 111,32 | 103,40 | 119,21 | 119,29 | 28,92 | 34,34 | 6,25 | 10,80 | 0 | 0 |
| Tud | 111,25 | 109,52 | 137,65 | 137,46 | 53,92 | 50,47 | 40,25 | 39,34 | 2,04 | 0,88 |

Az eredeti eloszláshoz képest (%)



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

| TOP 4-grammok (Webkorpusz2) | Gyak. | kíván | mer | bír | hajlandó | képes | képtelen |
|---------------------------------|-------|-------|-----|-----|----------|-------|----------|
| [/Cnj] [/N][Acc] L [/V][Inf] | 18950 | M W | M W | 0 W | M W | M W | 0 W |
| [/N][Nom] nem L [/V][Inf] | 16290 | M W | M W | M W | M W | M W | 0 0 |
| [/N][Nom] [/N][Acc] L [/V][Inf] | 16130 | M W | 0 W | 0 W | 0 W | M W | 0 0 |
| [/Cnj] nem L [/V][Inf] | 15768 | M W | M W | M W | M W | M W | 0 0 |
| [/N][Acc] nem L [/V][Inf] | 13869 | M W | M W | M W | M W | M W | 0 0 |
| [/N][Acc] [/Prev] L [/V][Inf] | 9242 | M W | 0 W | 0 W | 0 0 | 0 0 | 0 0 |
| [/Cnj] [/Prev] L [/V][Inf] | 8663 | M W | M W | M W | 0 0 | 0 0 | 0 0 |
| nem L [/V][Inf] [/N][Acc] | 7127 | M W | M W | M W | M W | M W | 0 0 |
| [/Cnj] L [/V][Inf] [/N][Acc] | 6738 | 0 W | M W | 0 W | M W | M W | M W |
| [/N][Nom] [/Prev] L [/V][Inf] | 5100 | M W | M W | 0 W | 0 0 | 0 0 | 0 0 |

A leggyakoribb mozaik 4-grammok a Webkorpusból
 (Az *akar, tud* is *szeret(ne)* mindkét korpuszban minden mintánál szerepel)



| | 3-gram | | | | 4-gram | | | | 5-gram | | | | 6-gram | | | |
|-----------------|--------|-----|----|------|--------|------|----|------|--------|------|-----|------|--------|-----|----|-------|
| | Ö | J | H | % | Ö | J | H | % | Ö | J | H | % | Ö | J | H | % |
| <i>akar</i> | 96 | 94 | 2 | 2,08 | 414 | 405 | 9 | 2,17 | 492 | 478 | 14 | 2,85 | 72 | 60 | 12 | 16,67 |
| <i>bír</i> | 22 | 22 | 0 | 0,00 | 97 | 97 | 0 | 0,00 | 130 | 120 | 10 | 7,69 | 0 | | | |
| <i>hajlandó</i> | 18 | 18 | 0 | 0,00 | 67 | 67 | 0 | 0,00 | 131 | 126 | 5 | 3,82 | 53 | 51 | 2 | 3,77 |
| <i>képes</i> | 49 | 48 | 1 | 2,04 | 250 | 244 | 6 | 2,40 | 242 | 229 | 13 | 5,37 | 41 | 31 | 10 | 24,39 |
| <i>képtelen</i> | 27 | 26 | 1 | 3,70 | 63 | 61 | 2 | 3,17 | 36 | 33 | 3 | 8,33 | 3 | 3 | 0 | 0,00 |
| <i>kíván</i> | 62 | 57 | 5 | 8,06 | 295 | 287 | 8 | 2,71 | 251 | 234 | 17 | 6,77 | 41 | 31 | 10 | 24,39 |
| <i>mer</i> | 70 | 67 | 3 | 4,29 | 248 | 241 | 7 | 2,82 | 339 | 321 | 18 | 5,31 | 24 | 18 | 6 | 25,00 |
| <i>szeret</i> | 143 | 140 | 3 | 2,10 | 406 | 395 | 11 | 2,71 | 221 | 213 | 8 | 3,62 | 19 | 14 | 5 | 26,32 |
| <i>tud</i> | 89 | 86 | 3 | 3,37 | 402 | 394 | 8 | 1,99 | 536 | 503 | 33 | 6,16 | 72 | 60 | 12 | 16,67 |
| Összesen | 576 | 558 | 18 | 3,13 | 2242 | 2191 | 51 | 2,27 | 2378 | 2257 | 121 | 5,09 | 325 | 268 | 57 | 17,54 |

A mozaik n-gram mintázatok által azonosított típusok megoszlása.

Ö: Összes mintázattípus száma az adott hosszúságon

J: Jó, vagyis valódi mintázat a korpuszból

H: Hibás mintázat



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Esettanulmány: főnévi igeneves szerkezetek (Indig és Bajzát 2024)

Összegzés:

- A módszer teljesítette a hozzá fűzött reményeket: kezelhető mennyiségű, jó minőségű mintát kaptunk
 - Fejből nem tudtuk volna felsorolni és rangsorolni őket
 - Főleg triviális hibák vannak csak (pl. névutóval kezdődő vagy névelővel végződő minta)
- A következő lépés: a kapott minták validálása emberek bevonásával (folyamatban)
 - A nyelvi játék segítségével összehasonlítani az emberi tudást a nyelvmodellekével specifikus eseteken
- Nyitott kérdések:
 - Alacsonyabb küszöbnél mi történne?
 - Előbb a főnévicsoport tömörítés és utána az ablak kivágása a mondatból és végül mozaik
 - Jobb, átgondoltabb főnévicsoport tömörítés (fontos a visszaállíthatóság a CQL konverzió miatt)
 - Az ablak kivágását átgondolni (idézőjelek, zárójelek, központoszási hibák stb.)
- Az eljárás kiterjesztése
 - Igeköötös szerkezetek, Aszaló (<http://github.com/dlazesz/aszalo>)
 - Magyar Szerkezettár (Sass és tsai. 2025)





Potenciális alkalmazási területek



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Potenciális alkalmazási területek

- Metaadatok tükrében
 - Zsánerspecifikus konstrukciók (elfogadták az absztraktunk, Indig és Bajzát (2025))
 - Diakrón konstrukciók (pl. a mai nyelv összevetése a Történeti Magánéleti Korpusssal (Bajzát, Indig és Kalivoda 2024))
- Szófelbontás
 - Finomabb szélek a kitölthető elemekhez (pl. el- [csúszda/piac/rizsa] -(V)zta az időt, Kalivoda (2022))
 - Disztributív szóosztályok a környezet figyelembevételével
 - szinonimák/antonimák/asszociációk/analógiák (v.ö. király -> királynő)
- Szóösszevonás
 - Mondatvázak, diskurzív szerkezet (pl. korlátozott domainen (hírek), egyszerűbb esszészövegműfajok (nyelvvizsga))
 - Névelem felismerés, körülírások, koreferenciák (Az USA elnöke -> Donald Trump, A kéretlen techmessiás -> Elon Musk)
 - Kombinálható mondat címkézéssel (szentiment elemzés) → Sajtófigyelés



Potenciális alkalmazási területek

- Konstrukciók hálózatai
 - Kérdés formalizálás állításból (kérdés válasz rendszerek, chatbotok)
 - Tagadás, „következtetés” (pl. Ha A ad B-nek C-t, akkor B kapott C-t A-tól és C átkerült A-tól B-hez)
- Olyan esetek, ahol nincs (jó) szintatikai elemző
 - pl. Történeti Magánéleti Korpusz (Dömötör, Gugán és tsai. 2017; Novák, Gugán és tsai. 2018)
 - Gold standard korpusz szintaktikai elemzéshez
- Tanulói korpusz
 - Hibaminták (Bari és tsai. 2021)
 - Használt/Negligált szerkezetek
 - Megtanulandó stratégiák/konstrukciók (pl. panaszlevél írásához)
- Nyelvmodellek ellenőrzése (megértés, produkció, fordítás)
 - Gépi fordításnál: forráshűség, tükörfordíthatóság



Összefoglalás

- Korpuszból elő tudunk állítani konstrukciókat
 - Adatpálható, félautomatikus módszerrel
 - Pontosan
 - Interpretálhatóan, visszakereshetően
- Számos alkalmazási lehetőséggel
 - Meglévő és új módszerek és erőforrások kifejlesztéséhez
- Számos nyitott és pontosításra váró kérdés van
 - Például, hogy hogyan lehetne becsatorázni más (elméleti) kutatásokba?
 - Hogyan tovább? Pontosítás vs. szélesítés



Mentális
konstrukciók



Korpuszminták



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Köszönöm a XXXXXXXXXX!








ELTE | IK
INFORMATIKAI KAR








MESTERSÉGES
INTELLIGENCIA
TANSZÉK

Hivatkozások I

-  Bajzát Tímea és Balázs Indig, „A mer, a bír és a szeret + főnévi igenév szerkezetek korpuszvezérelt azonosítása a forma–funkció oldaláról (Kézirat)”, Grammatikai kutatások és alkalmazások, 2025.
-  — „Személyjelölési mintázatok feltárása mozaik n-gram alapon három segédige + főnévi igenévi típus konstrukcióiban”, Személyjelölés és grammatika, 2024, 85–120. old.
-  Bajzát Tímea, Balázs Indig és Ágnes Kalivoda, „A fatens felelt pedig...” – A Történeti Magánéleti Korpusz igei szerkezeteinek mozaik n-gram alapú feldolgozása”, XX. Magyar Számítógépes Nyelvészeti Konferencia, 2024, 43. old.
-  Bari Diána Éva és tsai., „Mein Oma nicht surfing in die Internet ...” Eine Pilotstudie zum sprachlichen Management ungarischer Prüfungskandidaten und deren Bewertung”, ZEITSCHRIFT FÜR MITTELEUROPAISCHE GE 7 (2021), 61–77. old., ISSN: 2192-3043.
-  Dömötör Adrienne, Katalin Gugán és tsai., „Kiútkeresés a morfológiai labirintusból : korpuszépítés ó- és középmagyar kori magánéleti szövegekből”, NYELVTUDOMÁNYI KÖZLEMÉNYEK 113 (2017), 87–114. old., ISSN: 0029-6791, DOI: 10.15776/NYK.2017.113.3.








Hivatkozások II

-  Dömötör Andrea, Balázs Indig és Dávid Márk Nemeskey, „A méret a lényeg? Morfológiailag annotált korpuszok összehasonlító kiértékelése”, XXI. Magyar Számítógépes Nyelvészeti Konferencia, 2025, 219–230. old.
-  Indig Balázs, „Mosaic n-grams: Avoiding combinatorial explosion in corpus pattern mining for agglutinative languages”, Human Language Technologies as a Challenge for Computer Science and Linguistics, 2017, 147–151. old.
-  Indig Balázs és Tímea Bajzát, A human-centered analytical linguistic framework in the age of generative language. Paper will be presented at Workshop “Corpus linguistics 2040, Mannheim, 2025.
-  — „Bags and Mosaics: Semi-automatic Identification of Auxiliary Verbal Constructions for Agglutinative Languages”, Human Language Technologies as a Challenge for Computer Science and Linguistics – 2023, 2023, 111–116. old.
-  — „Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian”, Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, szerk. Mika Härmäläinen és tsai., Helsinki, Finland: Association for Computational Linguistics, 2024. nov., 96–103. old., URL: <https://aclanthology.org/2024.iwclul-1.12/>.



Hivatkozások III

-  Indig Balázs, László János Laki és Gábor Prószéky, „Mozaik nyelvmodell az AnaGramma elemzőhöz”, XII. Magyar 2016, 260–270. old.
-  Indig Balázs és Dániel Lévai, „I’m Smarter than the Average BERT! – Testing Language Models Against Humans in a Word Guessing Game”, Human Language Technologies as a Challenge for Computer Science and Ling 2023, 106–110. old.
-  Jakubíček Miloš és tsai., „Fast Syntactic Searching in Very Large Corpora for Many Languages”, PACLIC (2010), 741–47. old.
-  Kalivoda Ágnes, „PrevDistro: An open-access dataset of Hungarian preverb constructions”, Acta Linguistica Acad 69.4 (2022), 549–563. old., DOI: 10 . 1556 / 2062 . 2022 . 00578, URL: <https://akjournals.com/view/journals/2062/69/4/article-p549.xml>.
-  Nemeskey Dávid Márk, „Natural language processing methods for language modeling”, dissz., Doctoral School of informatics, Eötvös Loránd University, Faculty of Faculty of Informatics, 2020.



Hivatkozások IV



Novák Attila, Katalin Gugán és tsai., „Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence”, Lang. Resour. Eval. 52.1 (2018. márc.), 1–28. old., ISSN: 1574-020X, DOI: 10.1007/s10579-017-9393-8, URL: <https://doi.org/10.1007/s10579-017-9393-8>.



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), szerk. Nicoletta Calzolari és tsai., Portorož, Slovenia: European Language Resources Association (ELRA), 2016. máj., 1315–1322. old., URL: <https://aclanthology.org/L16-1209/>.



Oravecz Csaba, Tamás Váradi és Bálint Sass, „The Hungarian Gigaword Corpus”, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. máj., 1719–1723. old., URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf.



Sass Bálint és tsai., „Magyar szerkezetár demó”, XXI. Magyar Számítógépes Nyelvészeti Konferencia, 2025, 247–256. old.



ELTE | IK
INFORMATIKAI KAR



MESTERSÉGES
INTELLIGENCIA
TANSZÉK