DIGITÁLIS BÖLCSÉSZETI KÖZPONT

A Szöveglabor bemutatása

Simon Eszter

2020. szeptember 23.

PIM DBK

Az előadás vázlata

1. Bevezetés

2. Tervezési munkafázisok

Bevezetés

dHUpla alterületek

- forráskiadások
- kritikai kiadások
- born digital
- szöveglabor

Szöveglabor

a munkacsoport:

Kalcsó Gyula
Simon Eszter
Mihály Eszter

szöveglabor beszámoló:

október 7. 10:00

Tervezési munkafázisok

Tervezési munkafázisok

1. Hazai és nemzetközi jó gyakorlatok figyelembe vétele a tervezéskor
2. Szöveglabor funkciók
3. Eszközök a nyelvi feldolgozáshoz
4. A nyelvi feldolgozásra épülő eszközök a szöveglabor fejlesztéséhez
5. Feldolgozható szövegkorpuszok körének meghatározása
6. DHUplába integrálható szövegkorpuszok típusainak definiálása
7. Koherencia létrehozása a többi alterülettel
8. Jogi kérdések

Feldolgozható szövegkorpuszok körének meghatározása

a feldolgozható szövegkorpuszok 3 köre:

1. Saját szövegek a dHUpla 1-3. alterületéről (forráskiadás, born digital, kritikai kiadás)
2. távoli szövegek (DIA, MEK stb., elsősorban konzorciumi partnerek korpuszai)
3. egyéb külső hozott szövegek a szöveglaborba

ami ehhez kell:

- minőségbiztosítás
- formai követelmények
- feltöltőfelület → ellenőrzés → elemzés, ha jó a szöveg & visszajelzés, ha nem jó

Az integrálható szövegkorpuszok típusainak definiálása

- formátum: kép: nem, szöveg: igen
- nyelv: csak magyar → a nem magyar nyelvű részeket kiszűrni és megjelölni
- kor: csak modern → a régi szövegekből legyen normalizált; ha nincs, akkor csak azt lehet vele kezdeni, amit az elemzetlen korpuszon lehet csinálni

Problémák, kihívások, kérdések

- az integrálásra szánt korpusz tartalmaz más annotációkat → a formátumot igazítani kell a dHUpla elvárt bemeneti formátumához → a korpusz tulajdonosa megoldja az átalakítást, vagy lemond az annotációiról
- az integrálásra szánt korpusz már elemezve van, de más elemzővel → újra kell elemezni, különben nem lesz kompatibilis a nyelvi elemzésre épülő további elemzésekkel

Koherencia létrehozása a többi alterülettel

- a többi alterületen is szükség lesz nyelvi feldolgozásra → ezeknél a szövegeknél meg kell határozni, hogy mely ponton mennek át a szöveglabor elemzési lépésein → minél hamarabb, annál jobb
- interoperabilitás: oda-vissza konverzió lehetséges legyen, egységes be-kimeneti formátum
- a DHupla korpuszainak összekötése a névtérrel

Jogi kérdések

milyen a jogi státusza annak a korpusznak, amit feltöltenek a dHUplába és elemeznek a szöveglaborral?

- elvárás: legyen hozzáférhetővé téve a dHUplán a szöveg és a szöveglabor által kiadott elemzések és vizualizációk
- kérdések:
  - a dHUplán közzétett korpusz máshol is publikálható?
  - csak egy adott korpuszra kötődik az együttműködés?

Hazai és nemzetközi jó gyakorlatok feltérképezése

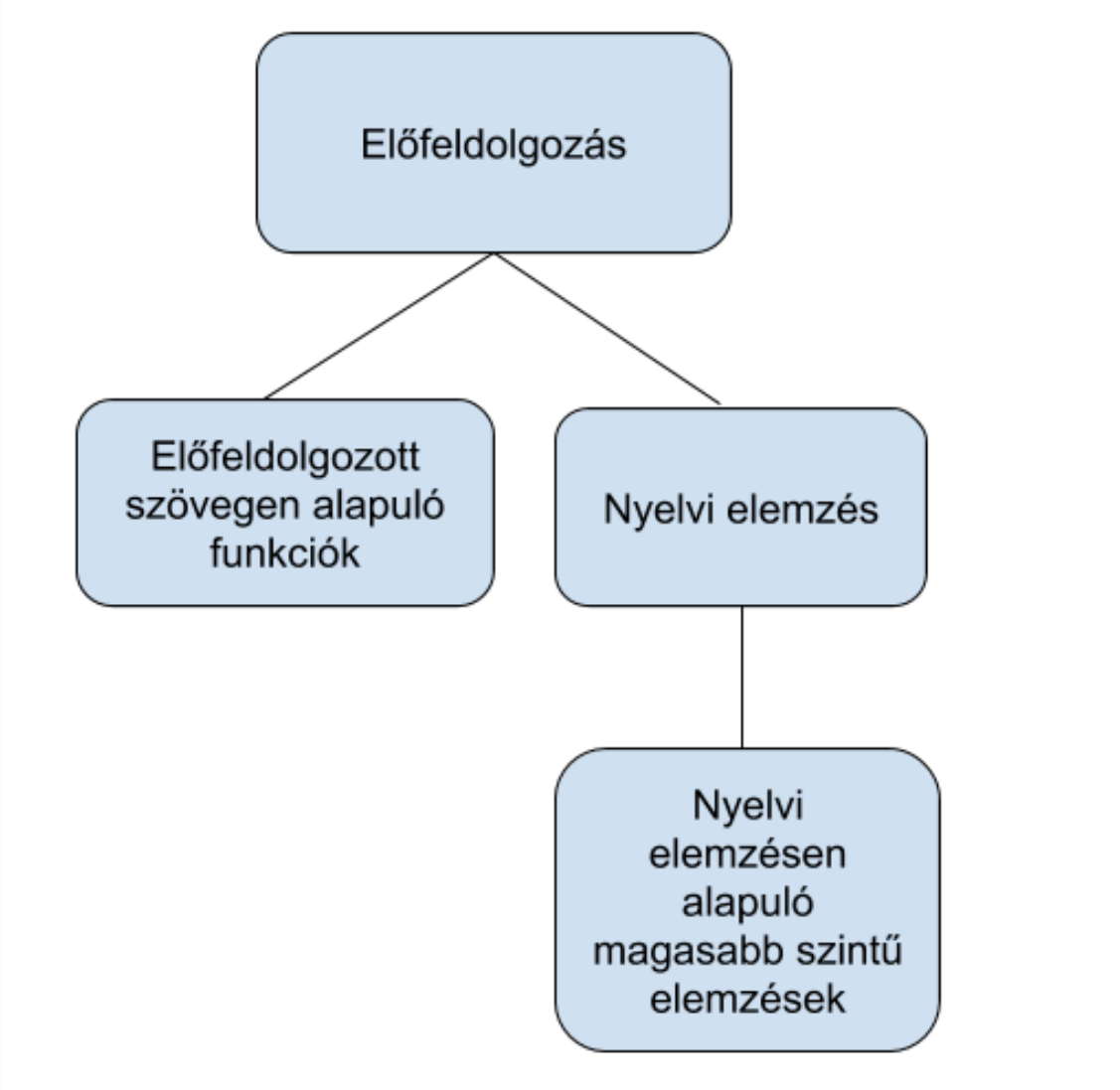
- szakirodalom: kivonatok stilometriai és MSZNY cikkekből
- eszközök
  - nyelvi feldolgozó eszközök
  - nyelvi feldolgozásra épülő eszközök
- best practices: projektek, workflow-k

Szöveglabor funkciók

a szövegfeldolgozásnak 4 szintje van:

1. előfeldolgozás
2. előfeldolgozott szövegen alapuló funkciók
3. nyelvi elemzés
4. nyelvi elemzésen alapuló magasabb szintű elemzések

A szövegfeldolgozás szintjei



Előfeldolgozás

- bemeneti formátum ellenőrzése:
  - UTF-8 karakterkódolás
  - txt
- tokenizálás
- mondatra bontás

Előfeldolgozott szövegen alapuló funkciók

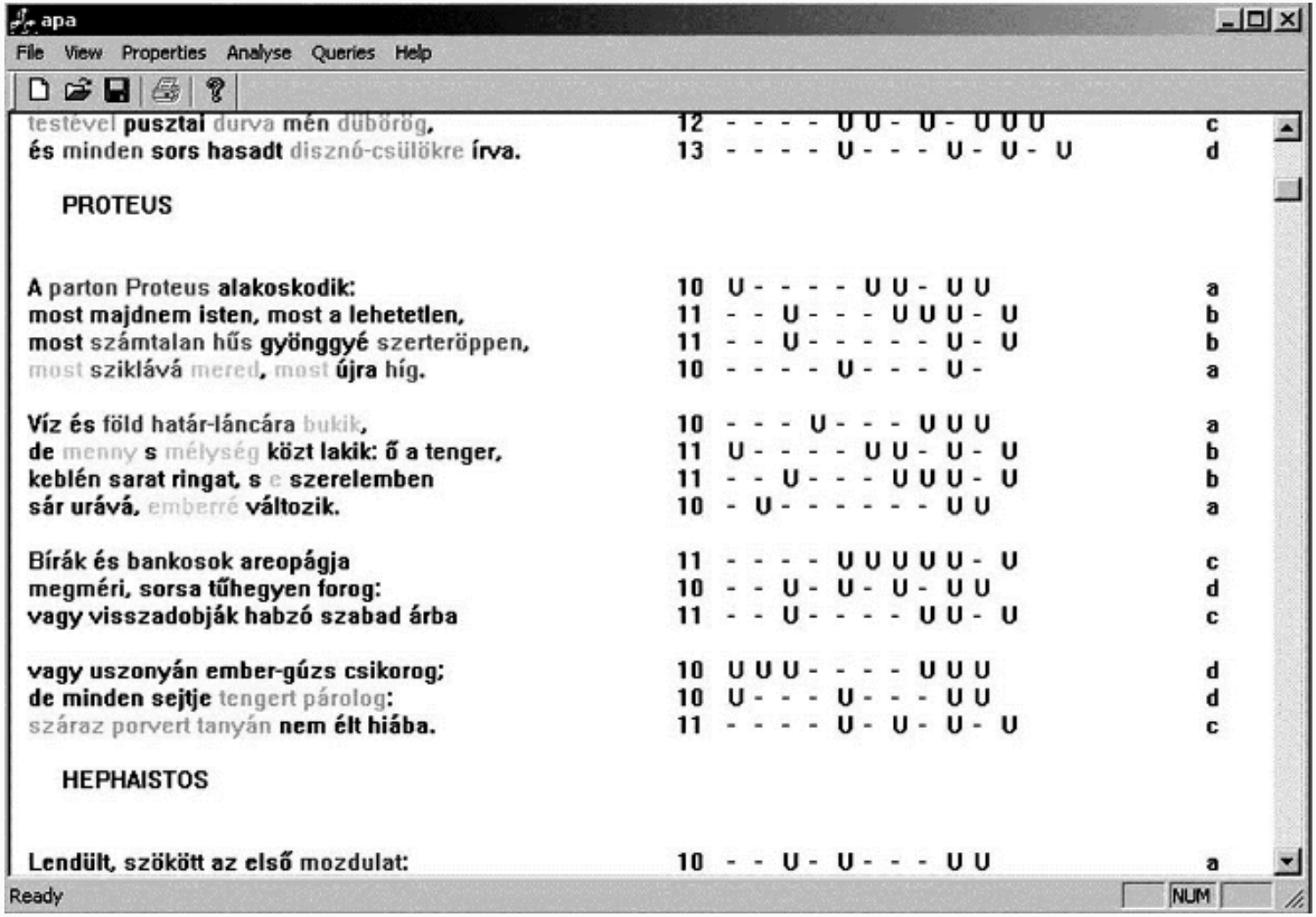
- w*ldcardos keresés
- [*rR*]eguláris kifejezése*k*+el való keresés
- kollokációk keresése (akár wildcardosan)
- konkordancia készítése
- szógyakoriság
- szókészlet listázása
- szó n-gramok $\rightarrow$ gyakorisági lista
- központozás vizsgálata
- magánhangzók–mássalhangzók arányának vizsgálata
- szavak és mondatok hosszának eloszlása
- az egyes szavak hány karakterből állnak
- az egyes mondatok hány szóból állnak
- verstani elemzés: metrika, rím, alliteráció automatikus detektálása

Példák

http://clara.nytud.hu/mnsz2-dev/index.html

- reguláris kifejezésekkel való keresés
- kollokációk keresése
- konkordancia készítése
- szógyakoriság
- szókészlet listázása
- CV-váz

Példák



Lesi Zoltán: Automatikus formai verselemzés.
In: Alkalmazott Nyelvtudomány VIII. évf. 1-2. szám 2008. 197-208.o.

Nyelvi elemzés

- morfológiai elemzés és egyértelműsítés
- tulajdonnév-felismerés
- NP chunking
- szintaktikai elemzés
  - összetevős elemzés
  - függőségi elemzés

Nyelvi elemzésre épülő magasabb szintű elemzések 1.

morfológiai információn alapuló funkciók:

a morfológiai elemző és egyértelműsítő kimenetén alapulnak

- szófaji gyakoriság
- szófajok aránya
- toldalékgyakoriság
- szógyakoriság lemmatizált szövegen
- keresések:
  - regexes keresés → bizonyos szófajú/tövű/ragozású szavak keresése
  - kollokációk keresése → bizonyos szófajú/tövű/ragozású szavak együttállása
  - konkordancia készítése → lemmatizált szövegen
  - szókészlet listázása → lemmatizált szövegen

Példák

4. Táblázat: A domének hasonlósága a szófajok eloszlása terén.

	Lenin-mesék	Lenin előtt	Lenin után	szovjet mesék	Mátyás-mesék	Rákosi-mesék	szocreál irodalom
Lenin-mesék		0,993	0,991	0,9976	0,9939	0,9951	0,9865
Lenin előtt	0,993		0,9953	0,9942	0,993	0,9917	0,9868
Lenin után	0,991	0,9953		0,9944	0,995	0,9919	0,9864
szovjet mesék	0,9976	0,9942	0,9944		0,9939	0,9939	0,9853
Mátyás-mesék	0,9939	0,993	0,995	0,9939		0,9927	0,9853
Rákosi-mesék	0,9951	0,9917	0,9919	0,9939	0,9927		0,9914
szocreál irodalom	0,9865	0,9868	0,9864	0,9853	0,9853	0,9914	

Horváth Csilla: “S azóta jól élnek, és vidám dalokat énekelnek Leninről és Sztálinról.” Szovjet propagandamesék műfaji azonosításának kísérlete. In: MSZNY 2018. 331-339.o.

Nyelvi elemzésre épülő magasabb szintű elemzések 2.

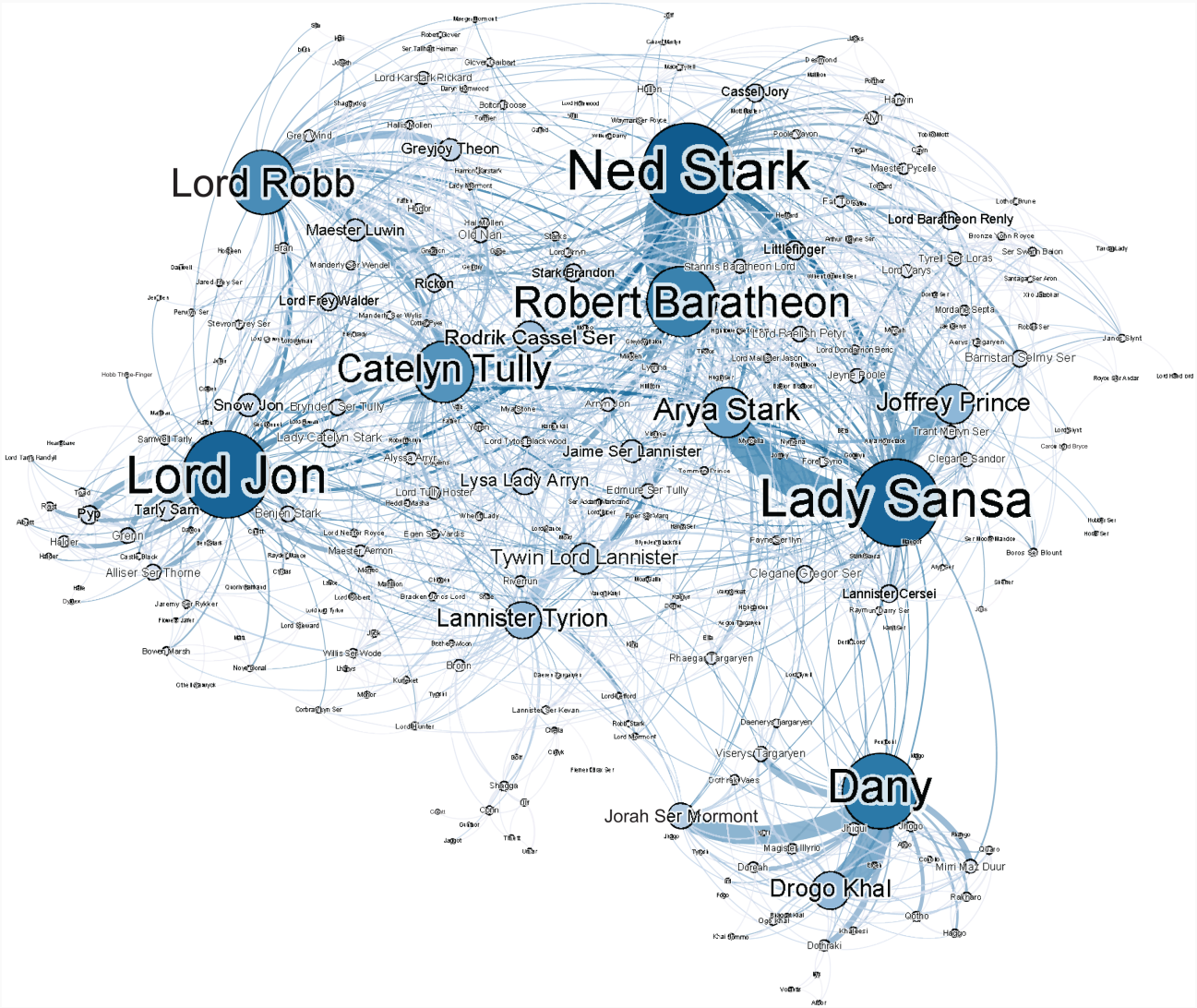
szintaktikai információn alapuló funkciók:

mondatszerkezeti statisztika

tulajdonnév-felismerésen alapuló funkciók:

- tulajdonnév-statisztika
- egy dokumentumban előforduló személyek kapcsolati hálója

Példák



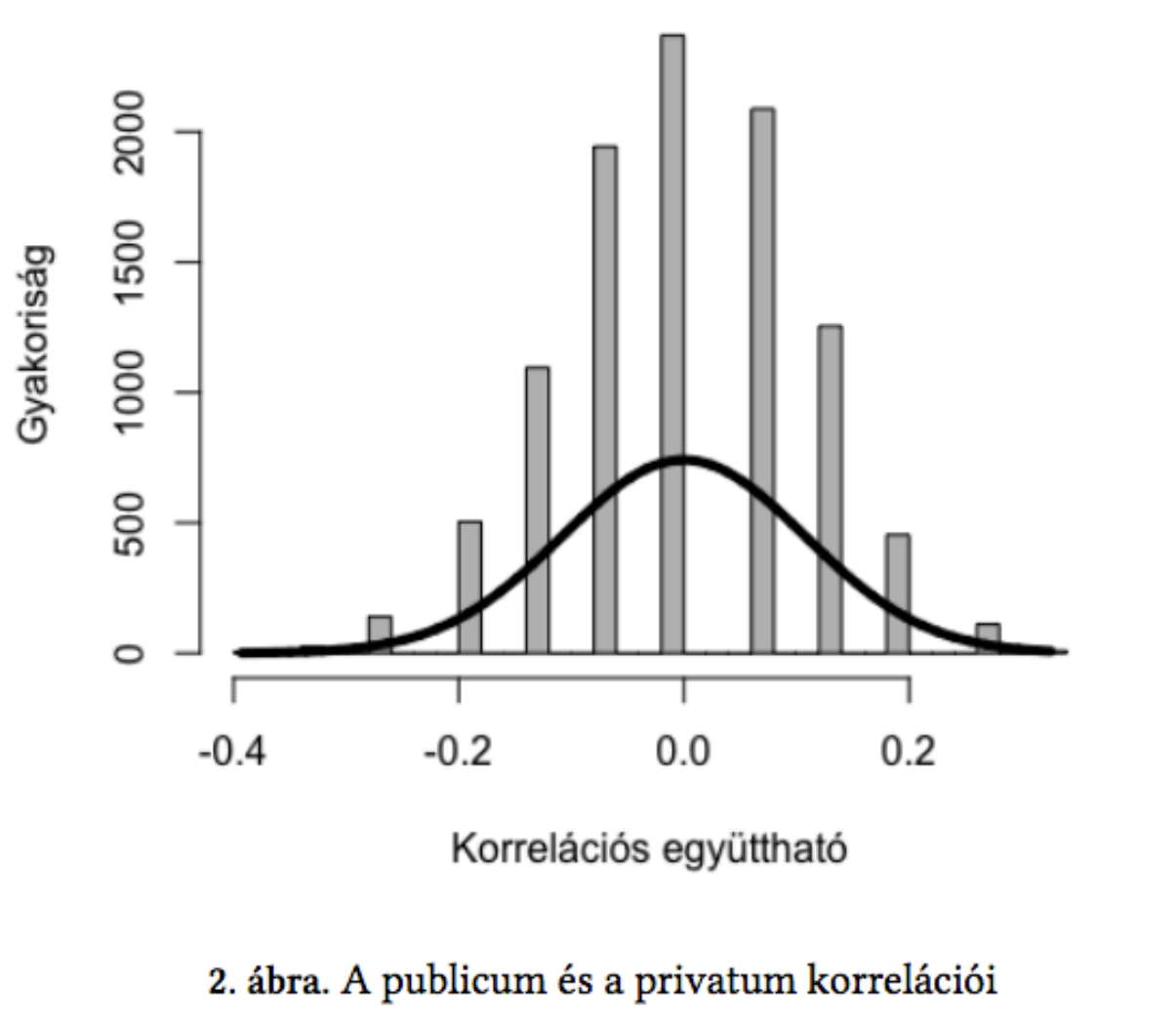
DOI: 10.7717/peerj-cs.189/fig-1

Nyelvi elemzésre épülő magasabb szintű elemzések 3.

emóciógyakoriság

- nyolc kategóriás emóciószótáron alapul: öröm, bánat, düh, szeretet, félelem, undor, meglepetés, feszültség → a szótár az ezeket az emóciókat megjelenítő szavakat tartalmazza
- kell hozzá lemmatizálás
- bemenetként az egyes tokenek lemmáját kéri, és azt összehasonlítja a szótár elemeivel: szimpla listaelem-illesztés
- az egyes emóciókhoz tartozó szövegbeli előfordulásokat leszámoljuk, belőlük gyakorisági listát készítünk

Példák



2. ábra. A publicum és a privatum korrelációi

Labádi Gergely: Az olvasó gép. In: Digitális Bölcsészet 1 (2018)

Eszközök a nyelvi feldolgozáshoz

eszközök:

- *e-magyar*
- Magyarlánc 3.0
- UDPipe
- huspaCy
- stanza

kérdések:

- gyorsaság → offline vagy realtime elemzés?
- továbbfejleszthetőség → modularitás, nyílt forráskód
- fenntarthatóság
- kívánt funkciók
- pontosság

ezt a kívánt funkcionalitások listája fogja meghatározni

kérdések:

- meglevő eszközök adaptálása vagy új eszközök fejlesztése?
- interoperabilitás: átjárhatóság az egyes eszközök között → egységes be-kimeneti formátum, konverzió
- integrálhatóság: ha nem nyílt forráskódú, akkor hogyan integráljuk a láncba?

stilometria → Kalcsó Gyula előadása

Köszönöm a figyelmet!

simon.eszterke@gmail.com