Bevezetés a korpuszok csodálatos világába

Nyelv és informatika – Pécs, 2022/23 tavasz
2. óra

Simon Eszter – Vadász Noémi

2023. február 18.

Tartalom

Mi a korpusz?

Mi a korpusz? 1.

Kugler and Tolcsvai Nagy (2000)

„meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi”

Mi a korpusz? 2.

Sinclair (2005)

„a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research”

Mi a korpusz? 3.

A korpusz ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. A szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nemcsak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat). Az MNSZ a mai magyar írott köznyelv általános célú reprezentatív korpusza kíván lenni. Az MNSZ lényegi tulajdonsága, hogy minden szó mellett feltünteti a szótövet, a szófajt és a szó morfológiai elemzését is. A szótő, szófaj és elemzés megállapítása és az elemzések egyértelműsítése automatikus gépi eszközökkel történik. A rendszer megbízhatósága kb. 97,5%-os, így az összes szóalak kb. 2,5%-a hibásan van elemezve. Ennél jobb eredményt csak a kézi elemzés biztosíthatna, ami ekkora méretű anyag esetén megvalósíthatatlan.

(http://corpus.nytud.hu/mnsz/bevezeto_hun.html)

Kulcsszavak

- mennyiség
- nyelvészeti vizsgálatokra alkalmas
- reprezentativitás, a kiválasztás szempontjai
- tárolás módja: elektronikus
- tartalom: szegmentálás, annotáció, metaadatok

Korpusztipológia

A felhasználás módja

- általános: egy nyelv minél hitelesebb reprezentálása, elsősorban a lexikográfusoknak (MNSZ, British National Corpus (BNC))
- speciális (Hong Kong Corpus of Conversational English (HKCCE))

A mintavétel módja

- statikus (Brown, LOB)
- dinamikus (COBUILD 1980 óta, az első korpuszalapú szótár)
- monitor

Modalitás/forrás

- írott
- hangzó (audio) (paasonen_1315.eaf)
- video (http://jelesely.hu/szotar/?dictionary&id=search&search_id=281)
- multimodális (pl. gesztusfelismerés, prozódia, diskurzuselemzés)
- kézzel írott, nyomtatott, eleve elektronikusan keletkezett

Domén/genre/műfaj

- gazdasági rövidhírek
- termékleírások
- szoftverdokumentáció
- szépirodalom
- diákfogalmazások
- tudományos írások
- enciklopédia
- ...

Nyelv

- egynyelvű
- kétnyelvű
- többnyelvű

párhuzamos korpuszok (parallel corpora)
a forrásnyelvi szöveget (S) és annak célnyelvi fordítását (T) tartalmazzák, mondat- vagy bekezdésszinten párhuzamostíva $\rightarrow$ S és T pontos fordítása egymásnak

összevethető korpuszok (comparable corpora)
ha S és T nem pontos fordításai egymásnak, de a mintavétel módját tekintve megegyeznek, akkor beszélünk összevethető korpuszról (McEnery and Xiao, 2007)

Származási hely

US	Brown Corpus
UK	Lancester–Oslo/Bergen Corpus
India	Kolhapur Corpus of Indian English
Ausztrália	Australian Corpus of English
Új-Zéland	Wellington Corpus of Written New Zealand English
Kanada	Corpus of English-Canadian Writing

Származási idő

- szinkrón (MNSz)
- diakrón (Ómagyar Korpusz)

Gold standard korpusz

Követelmények:

- kihalt nyelvek esetében kimerítő, amúgy reprezentatív, de legalábbis kiegyensúlyozott
- nyelvi elemekre van bontva (token, mondat, bekezdés…)
- nyelvi annotáció van minden elemhez rendelve
- az annotáció vagy kézzel készül, vagy kézzel van ellenőrizve egy előre kidolgozott annotációs séma és útmutató alapján
- jellemzően előre meghatározott a méretük

Silver standard korpusz

- maga a korpusz vagy az annotáció automatikusan generált
- kiterjeszthető új szövegekkel és új annotációs szintekkel
- az annotáció megbízhatósága fontos szempont

Főbb kérdések a korpuszépítésnél

Mire jó a korpusz?

- szinkrón nyelvi jelenségek vizsgálatára
- longitudinális nyelvészeti vizsgálatokra
- nyelvtanulásra
- nyelvfeldolgozó eszközök tanítására és tesztelésére
- szótárépítésre
- ...

Korpuszdizájn

Tisztázandó kérdések:

- kik és mire fogják használni a korpuszt
- a nyelvváltozat, amit le szeretnénk fedni
- a műfaj, amit reprezentálni szeretnénk
- a szükséges méret
- a korpusz jövőbeli elérhetősége, használhatósága → copyright kérdések és a szöveggyűjtés nehézségei

Mintavételezés, reprezentativitás

McEnery (2004)
„collected within the boundaries of a *sampling frame* designed to allow the exploration of certain linguistic feature (or set of features) via the data collected”

Hunston (2008)
„*representativeness* is the relationship between the corpus and the body of language it is used to represent”

Reprezentativitás

- a korpusz nem szövegek véletlen halmaza, hanem tudatosan megtervezett gyűjtemény
- a reprezentativitás megvalósítható?
- pl. egy álalános nyelvi korpusz esetén olyan arányban tartalmazzon mindenféle szöveget, amilyen arányban a nyelvhasználatban is előfordulnak (diákszlengtől kezdve a filozófiai értekezéseken át a mikrohullámú sütő használati utasításáig)
- egyetlen nyelvre nézve sem áll rendelkezésünkre pontos statisztika, így a korpuszokat alkotó alkorpuszok százalékos aránya teljességgel önkényes
- kiegyensúlyozott korpusz

Mintavételezés

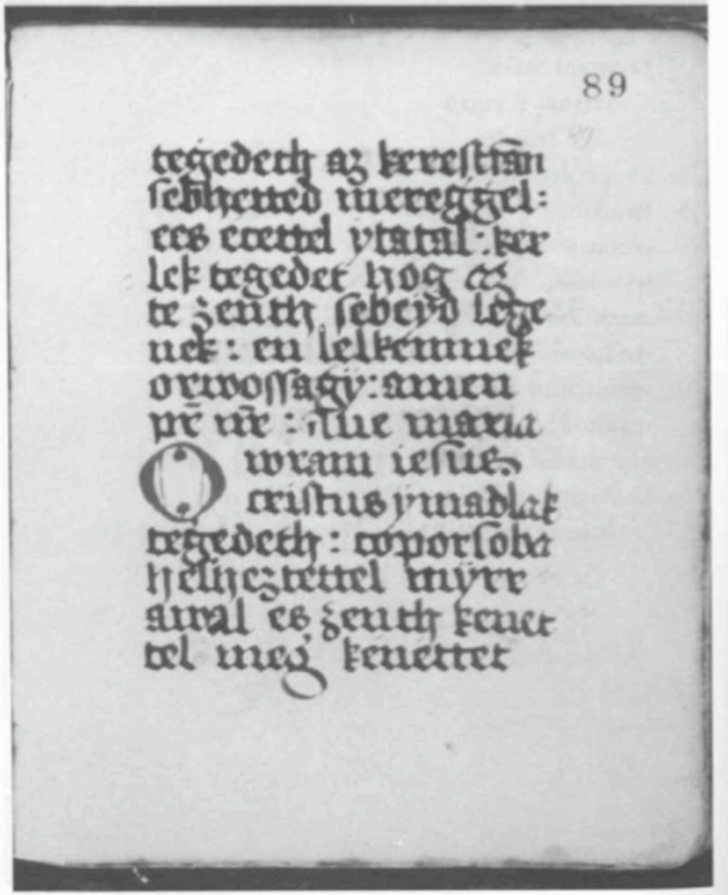
- a kutatás tárgya határozza meg a korpusz összetételét
- minél jobban körülhatárolható a kutatási kérdés, annál könnyebb döntéseket hozni a korpusz tartalmáról
- korai korpuszok: amerikai angol általános korpusz (Brown Corpus) és brit angol általános korpusz (Lancaster-Oslo/Bergen Corpus (LOB))
- mindkettőbe sok, különböző típusú szöveg került bele

Szöveggyűjtés

A szöveg forrása:

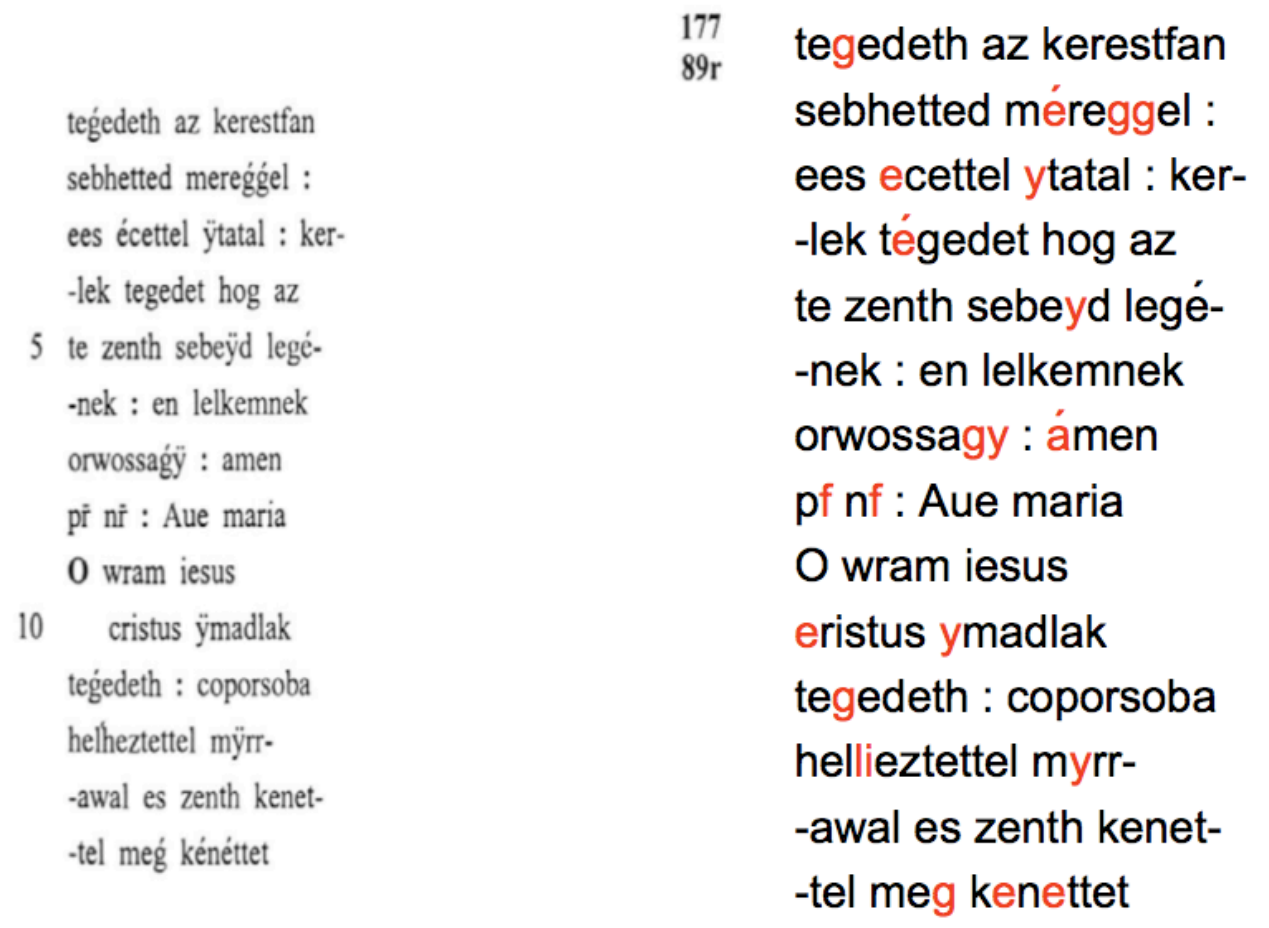
- elektronikus formátum
  - gép által olvasható, strukturált szöveges formátum → XML-parszolás
  - strukturálatlan szöveges formátum → strukturálttá alakítás
  - kép → szöveggé alakítás
- papíralapú formátum → elektronikussá alakítás

Kézzel írott kódexek, nyomtatott kiadások



tegedeth az kerestfan
sebhetted meregĝel :
ees écettel ÿtatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pĭ nĭ : Aue maria
O wram iesus
cristus ÿmadlak
tegedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meg kénéttet

Szkennelés, OCR



tegedeth az kerestfan
sebhetted meregĝel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeyd legé-
-nek : en lelkemnek
orwossagy : amen
pf nf : Aue maria
O wram iesus
eristus ymadlak
tegedeth : coporsoba
hellieztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

177
89r

tegedeth az kerestfan
sebhetted méreggel :
ees ecettel ytatal : ker-
-lek tégedet hog az
te zenth sebeyd legé-
-nek : en lelkemnek
orwossagy : ámen
pf nf : Aue maria
O wram iesus
eristus ymadlak
tegedeth : coporsoba
hellieztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

Jogi kérdések

- a szerzői jog tulajdonosának előzetes írásbeli beleegyezése nélkül jogellenes mind fénymásolatot, mind pedig elektronikus másolatot készíteni. Manapság ez nem csak teljes művekre, cikkekre, hanem részletekre is vonatkozik.
- EU: az írásművek a szerző halála után 70 évvel válnak szabadon felhasználhatóvá. Ezt megelőzően az írásmű felhasználásához a jogtulajdonos engedélye szükséges.
- a szövegek hasznosíthatóságával kapcsolatban a licenc ad tájékoztatást

A korpusz mérete

Token–Type megkülönböztetés – példa(beszéd)

Mt 13,3-9

„Íme, kiment a magvető vetni. Amint vetett, némely szem az útszélre esett. Jöttek az égi madarak és fölcsipegették. Más mag köves talajba hullott, ahol nem volt neki elég föld. Gyorsan kikelt, mert nem volt mélyen a földben. Amikor azonban forrón tűzött a nap, elszáradt, mert nem volt gyökere. Ismét más szúrós bogáncsok közé esett. Amikor a bogáncsok felnőttek, elfojtották. A többi jó földbe hullott s termést hozott, az egyik százszorosat, a másik hatvanszorosat, a harmadik meg harmincszorosat. Akinek füle van, hallja meg.”

Token–Type megkülönböztetés 1., 2., 3.

11	,
10	.
6	a
3	volt
3	nem
3	az
2	mert
2	meg
2	hullott
2	esett
2	bogáncsok
2	Amikor
1	útszélre
1	és
1	égi
1	Íme

11	,
10	.
7	a
3	volt
3	nem
3	az
2	más
2	mert
2	meg
2	hullott
2	esett
2	bogáncsok
2	amikor
1	útszélre
1	íme
1	és

11	,
10	.
7	a
4	van
3	nem
3	föld
3	az
2	más
2	mert
2	meg
2	hull
2	esik
2	bogáncs
2	amikor
1	ő
1	útszél

Kitekintő: nyelvstatisztika

Nyelvstatisztika

néhány statisztika az angol nyelvről:

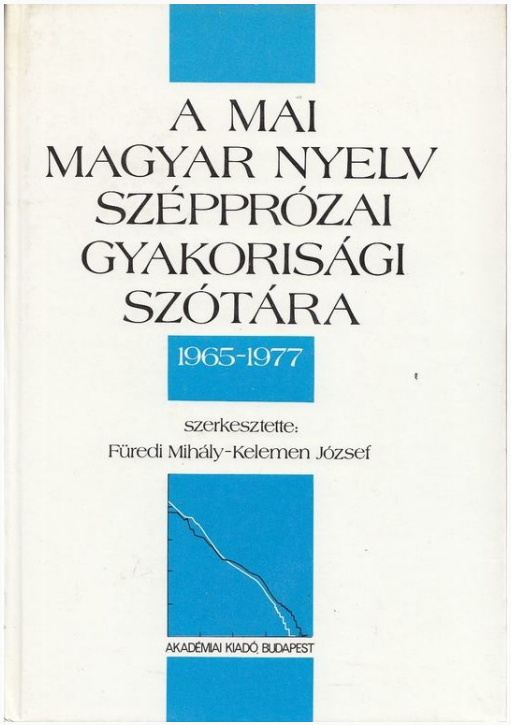
- a *q* betűt majdnem mindig *u* betű követi
- a szöveg kicsit több mint 60%-a mássalhangzó
- a köznapi beszédben használt szótagszerkezetnek kb. az egyharmada CVC szekvencia
- a nyelv 50 leggyakrabban használt szava teszi ki a szövegek 45%-át

Betűgyakoriság

Füredi–Kelemen (1989): a betűk sorrendje előfordulási gyakoriságuk alapján (60-as, 70-es évek szépirodalmán mérve):

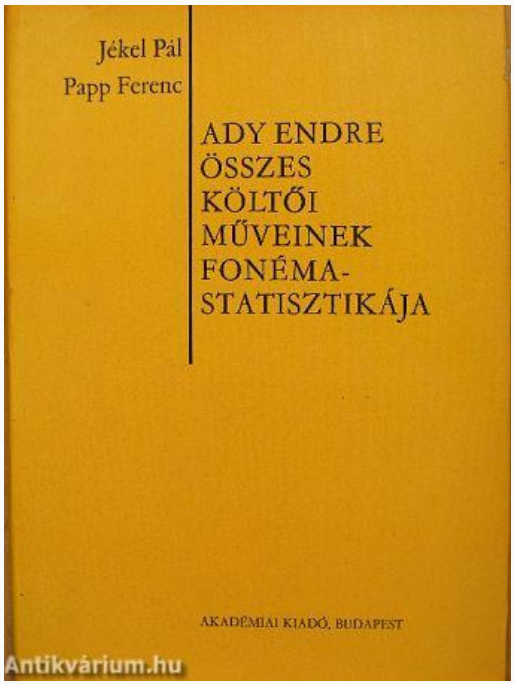
e a t l n s k o m r i g á é d b v h j ö f p u ő ó c ü í ú ű w

Füredi Mihály, Kelemen József (1989): *A mai magyar nyelv szépprózai gyakorisági szótára*. Akadémiai Kiadó, Budapest



Fonémastatisztika

Papp Ferenc, Jékel Pál (1974): *Ady Endre összes költői műveinek fonémastatisztikája*. Akadémiai Kiadó, Budapest



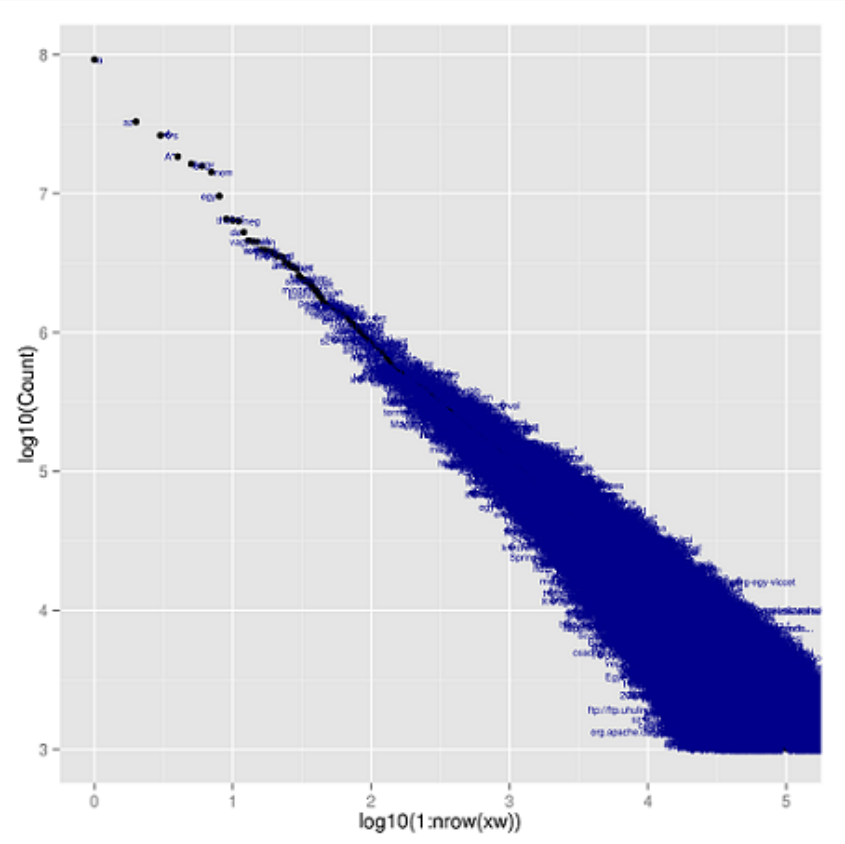
Zipf-törvény

Zipf (1902–1950) amerikai filológus: „Egy szó előfordulási gyakorisága fordítottan arányos a gyakorisági táblában levő rangjával. Így, a leggyakoribb szó közel kétszer gyakoribb, mint a második leggyakoribb szó, és háromszor gyakoribb, mint a harmadik helyen lévő, stb.”

1. számoljuk meg a szavak előfordulását egy szövegben
2. tegyük csökkenő gyakorisági sorrendbe és sorszámozzuk
3. a sorszám szorozva a gyakorisággal állandó

Zipf-görbe

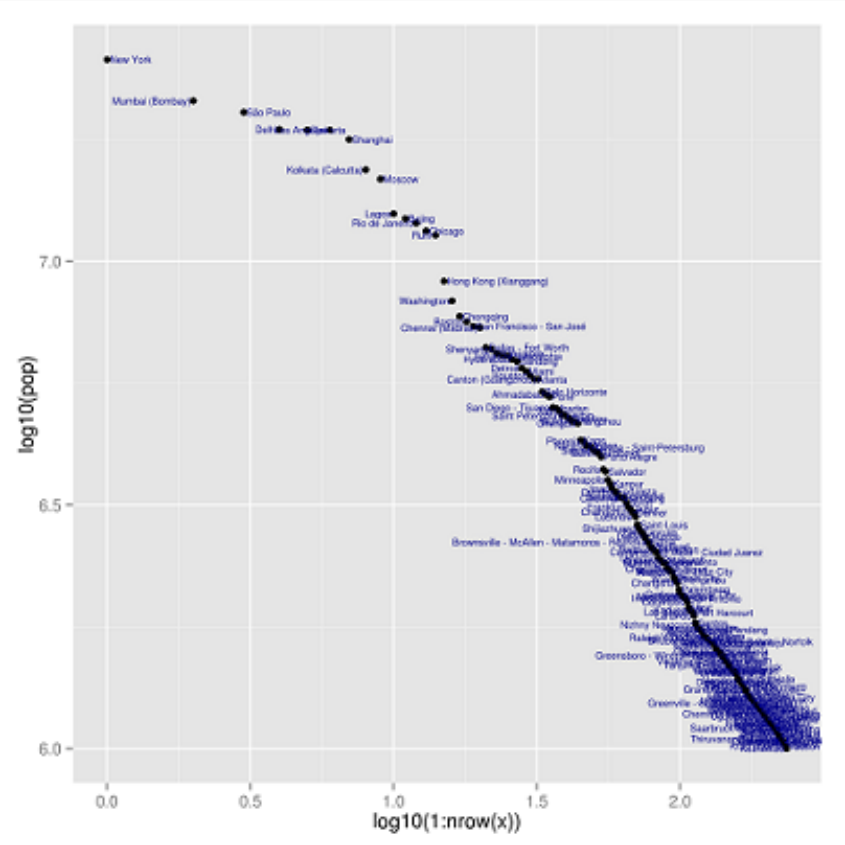
A Magyar Webkorpusz 10 000 leggyakoribb elemét mutatja az alábbi grafikon (a vízszintes tengelyen a frekvenciatáblában elfoglalt pozíciót, a függőlegesen pedig a gyakorisági értéket mutatjuk).



forrás: https://kereses.blog.hu/2013/08/05/szavak_varosok_long_tail_es_a_80_20_szabaly

Zipf-görbe a nyelvtudományon kívül

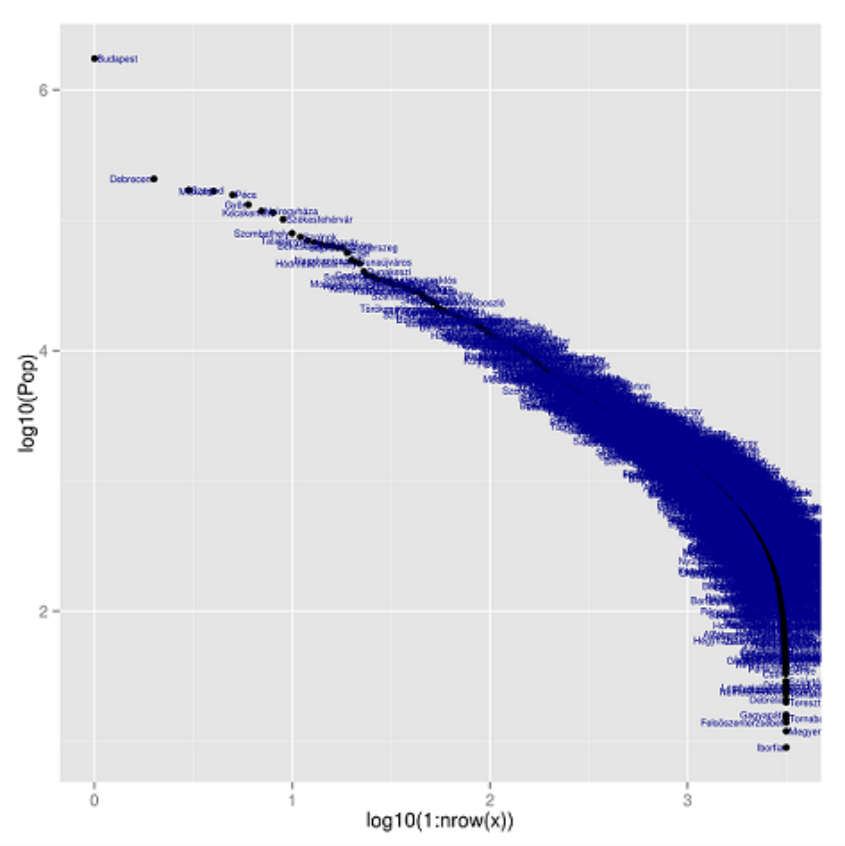
A nagyobb világvárosok lélekszáma és a lakosság szerinti sorrendben elfoglalt pozíció közötti fordított arányosság.



forrás: https://kereses.blog.hu/2013/08/05/szavak_varosok_long_tail_es_a_80_20_szabaly

Zipf-görbe a nyelvtudományon kívül

Ugyanez a magyar városokkal.



forrás: https://kereses.blog.hu/2013/08/05/szavak_varosok_long_tail_es_a_80_20_szabaly

Gyakoriság és szóhossz

Mandelbrot kiegészítése Zipf megfigyeléséhez: Fordított összefüggés van a szó hossza és gyakorisága között.

- a leggyakrabban használt angol szavak többsége egyszótagú
- ha egy szó gyakorisága növekszik, rövidítünk (magnetofon $\rightarrow$ magnó)
- hatékony kommunikációs elv

… vissza a korpuszokhoz

Az MNSz2 mérete

	magyarországi	*szlovákiai*	*kárpátaljai*	*erdélyi*	*vajdasági*	összesen
sajtó	350,5	11,6	0,7	0,6	1,5	364,8
szépirodalom	77,0	2,3	0,4	0,8	0,2	80,6
tudományos	112,0	3,3	0,7	1,6	0,3	117,9
hivatalos	98,0	0,2	0,3	0,6	0,1>	99,0
személyes	300,3	-	0,4	0,4	0,1>	301,1
beszéltnyelvi	76,2	-	-	-	-	76,2
összesen	1013,9	17,3	2,5	3,9	2,0	1039,7

Egyéb magyar korpuszok méretei

- Hungarian Webcorpus 1,48 milliárd token
- Hungarian Webcorpus 2.0 9 milliárd token
- Szeged Korpusz 1,2 millió token
- Szeged Dependency Treebank 42 ezer token
- NerKor 1 millió token
- SzegedKoref 124 ezer token
- KorKor 25 ezer token

Korpuszannotáció

A nyelvi annotáció

a sztenderd szövegfeldolgozó lépések a modern korpuszoknál nagyjából ugyanazok:

- szegmentálás (tokenizálás, mondatra bontás)
- morfológiai elemzés
- morfoszintaktikai egyértelműsítés

Mi kell az annotációhoz?

- annotációs séma
  - elméleti nyelvészeti alapok lefektetése (pl. mi a tulajdonnév?)
  - címkekészlet
  - az annotáció formátuma (inline vagy standoff)
- annotációs eszköz
- az annotátorok száma $\rightarrow$ annotátorok közötti egyetértés mérése
- annotációs útmutató
- az annotáció minőségének ellenőrzése

Annotációs útmutató

- az útmutatónak egyszerre kell kellően kidolgozottnak és egyszerűnek lennie, hogy az annotátorok számára követhető legyen → ha nem így van, akkor az annotátorok magas hibaszázalékkal fognak dolgozni
- tartalmaznia kell az annotációs feladat leírását, az annotálandó nyelvi elemek felsorolását és példákat arra, hogy mit kell és mit nem kell annotálni
- minél magasabb nyelvi szintre megyünk, minél több szemantika van benne, annál képlékenyebb a feladat → bizonyos nyelvi jelenségek nehezen megfoghatók/formalizálhatók
- ha az útmutató nem elég egzakt, akkor az annotátorok elkezdik követni az intuíciójukat → a nem teljesen egyértelmű esetekben ez problémákat okozhat

NER annotációs útmutatók

- MUC-7 Named Entity Task Definition (Chinchor, 1997)
- ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (Linguistic Data Consortium, 2008)
- Hunner project proposal és útmutató
- NYTK-NerKor útmutatók

Az annotáció formátuma

inline (XML)

```
<s><w>Ez</w> <w>egy</w> <w>mondat</w> <c> .</c>
<s><w>Meg</w> <w>a</w> <w>második</w> <c> .</c>
```

standoff

Ez
egy
mondat
.

Meg
a
második
.

XML

EXtensible Markup Language
egyfajta jelölőnyelv (markup language) → vannak más hasonlók:
YAML, JSON, MD

Előnyei:

- mind ember, mind gép számára olvasható formátum
- támogatja a Unicode-ot
- szabványos és platformfüggetlen
- képes a legtöbb általános számítástudományi adatstruktúra ábrázolására

Hátrányai:

- szintaxisa elég bőbeszédű és részben redundáns
- nagyobb tárolási költség
- nincs lehetőség a dokumentum egyes részeinek közvetlen elérésére
- átfedő adatstruktúrák modellezése nehéz/lehetetlen

Standoff annotáció

- az eredeti dokumentumok sima szöveg fájlok maradnak
- az annotációk nem szövegközi tagek, hanem egy külső fájlban jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét kap a szövegrészlet
- az annotálást teljesen különválasztjuk a használt feldolgozó eszközöktől, és közben minden információt megtartunk
- az átfedő és beágyazott annotáció is könnyen kezelhető

Beágyazott és átfedő annotáció

Beágyazott annotáció

```
<LOC><PERSON>Kossuth Lajos</PERSON>utca</LOC>
```

Átfedő annotáció

```
a Kossuth Lajos és a Petőfi Sándor utca sarkán
```

BIE1

[...] közölte Wolf László, az OTP Bank vezérigazgató-helyettese az MTI érdeklődésére.

közölte	O
Wolf	B-PER
László	E-PER
,	O
az	O
OTP	B-ORG
Bank	E-ORG
vezérigazgató-helyettese	O
az	O
MTI	1-ORG
érdeklődésére	O
.	O

BIE1

A szállásunk egy Balaton melletti kis üdülőfaluban, Zamárdiban volt.

A	B-NP
szállásunk	E-NP
egy	B-NP
Balaton	I-NP
melletti	I-NP
kis	I-NP
üdülőfaluban	I-NP
,	I-NP
Zamárdiban	E-NP
volt	O
.	O

Irodalom i

Javasolt olvasmányok:

- O'Keeffe and McCarthy (2010)
- Lüdeling and Kytö (2008)
- Szirmai (2005)

A kézi annotáció minősége

Az annotáció minősége

- a kézzel annotált korpuszokat tanító- vagy kiértékelőanyagként használják felügyelt gépi tanulással működő eszközök számára
- felügyelt gépi tanulással működő rendszerek sikeressége a tanítóanyag minőségén múlik
- csak olyan feladatokat lehet felügyelt gépi tanulással megoldani, amelyeket az ember is képes elvégezni
- csak olyan nyelvi jelenségekhez tudunk kézi annotációt készíteni, amelyeket eléggé megértettünk ahhoz, hogy pontosan le tudjuk írni őket
- megbízható az annotáció, ha a jelenségek leírását több annotátor is hasonlóképpen megértette és ez alapján hasonlóképpen kódolják az egyes jelenségeket
- a feladatleírásnak tehát érthetőnek kell lennie az annotátorok számára, akik ideális esetben egyetértenek az egyes jelenségek címkézésében

Az annotátorok közötti egyetértés

- a cél a minél magasabb annotátorok közötti egyetértés
- minél egyszerűbben leírható nyelvi jelenség annotálásáról van szó, annál könnyebb magas annotátorok közötti egyetértést elérni, a nyelvi jelenség összetettségével az egyetértés mértéke is könnyen csökken
- mitől lehet alacsony?
  - a feladat megfogalmazása nem egyértelmű vagy nem teljes
  - az annotátoroknak túl sok kategóriát kell kezelniük
  - átláthatatlan felületen kell dolgozniuk

Az annotátorok közötti egyetértés

- az annotátorok (vagy kódolók), amikor kategóriákat rendelnek egyes elemekhez, szubjektív döntéseket hoznak
- ha az annotátorok egyetértenek az egyes elemekhez rendelt kategóriákban, akkor az adat megbízható, és ha a kódolók következetesen hasonló eredményt produkálnak, akkor hasonlóképpen értették meg a feladatot és az annotálási útmutatót, ezért a továbbiakban is hasonló eredményeket várhatunk tőlük
- *megfigyelt egyetértés*: azt mutatja meg, hogy az esetek hány százalékában értett egyet a két kódoló
- DE! nem elég, ha két kódoló egyetért, hiszen mindketten tévedhetnek is
- a címkék számának csökkentésével növekszik a megfigyelt egyetértés, ráadásul nem érzékeny az egyes címkék eltérő gyakoriságára
- megoldás: valószínűség-korrigált együtthatók, amelyek számolnak a véletlen eseményekkel is

Az annotátorok közötti egyetértés: koefficiensek

Különböző mérőszámok az egyetértésre:

- megfigyelt egyetértés
- S (Bennett, Alpert és Goldstein 1954): minden kategória ugyanolyan valószínű, a kategóriák között egyenletes eloszlást feltételez
- π (Scott, 1955): kategóriánként eltérő, de kódolók között megegyező eloszlás
- κ (Cohen, 1960): kategóriánként és kódolónként eltérő eloszlás, ez már kezeli az elfogultságot
- α (Krippendorff, 1980): nem csak az egyetértést vizsgálja, hanem az egyet nem értés különböző fokozatait

Az egyetértés erőssége

Landis and Koch (1977)

κ	strength of agreement
<0.00	poor
0.00 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 1.00	almost perfect

Példák

Tulajdonnév-felismerés

hunNERwiki korpusz (Simon and Nemeskey, 2012):

- $\kappa = 0,967$
- F-mérték: 92,94%

Szeged NER korpusz (Szarvas et al., 2006):

- egyetértési arány: 99,6%

Metaforikus kifejezések felismerése (Babarczy et al., 2010)

egyetértési arány:

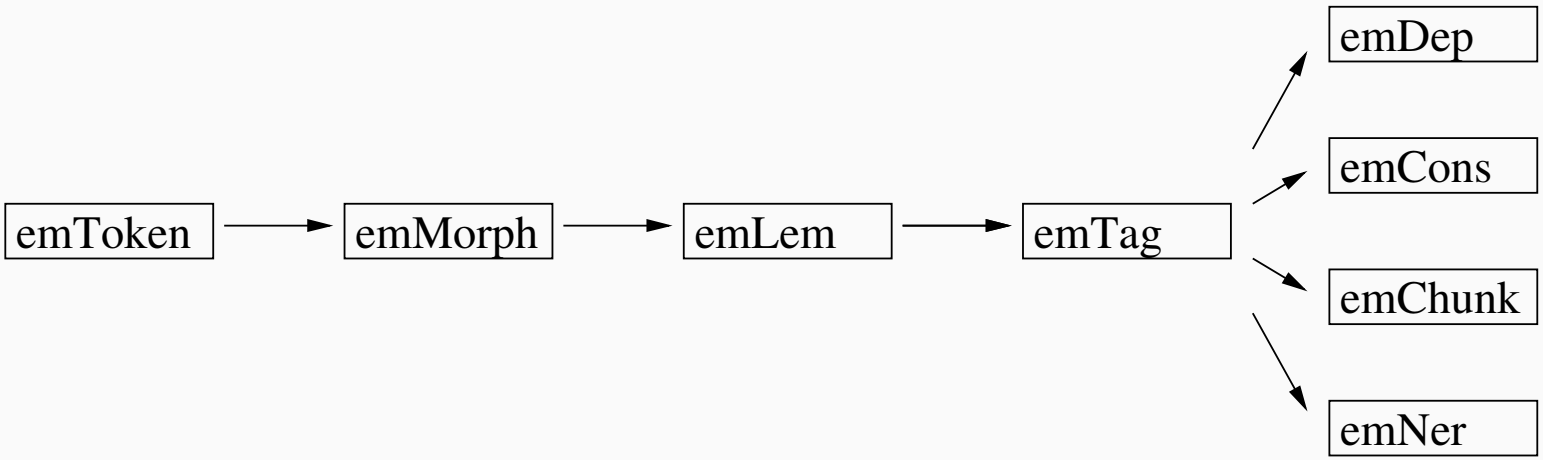
- 1. körben: 17%
- 2. körben: 48%

Szövegfeldolgozási és annotációs szintek

Alapszintű szövegfeldolgozási szintek

- mondatrabontás és tokenizálás
- morfológiai elemzés
- sekély szintaktikai elemzés
- mély szintaktikai elemzés
- tulajdonnév-felismerés
- ...

Egymásra épülő annotációk



Mondatszegmentálás

Mittelholcz (2017)

- Minden mondat.
- Mondathatárok azonosítása.
- Pontos problémák:
  - Rövidítések (*du. 5-kor*).
  - Római számok (*V. László*).
  - Sorszámok (*10. éve, hogy …*).
- Egyéb nehézségek:
  - Idézeten belüli mondatok.
  - Zárójelen belüli mondatok.

Tokenizálás

- Detokenizálhatóság és elválasztás (és az *-e* partikula).
- Szóalkotó karakterek, szónemalkotó karakterek, és amik köztük vannak:
  - Zárójelek, idézőjelek, aposztrófok kezelése.
  - Rövidítések végén lévő pont vs. mondatvégi pont.
- Számok (space-szel tagolt számok, mértékegységek, képletek, dátumok).
- Informatikai kifejezések (URL, elérési út, emailcím).
- Smiley-k és emoji-k.

Morfológiai elemzés

tokenszintű elemzés → nem lát se előre, se hátra → no kontextus → többértelműség

kerekesszék

kerek/ADJ+esszé/NOUN<PLUR>

kerekes/ADJ+szék/NOUN

kerék/NOUN[ATTRIB]/ADJ+szék/NOUN

kerek/ADJ[ATTRIB]/ADJ+szék/NOUN

kerék/NOUN[ATTRIB]/ADJ+szék/NOUN

kerek/ADJ[ATTRIB]/ADJ+szék/NOUN

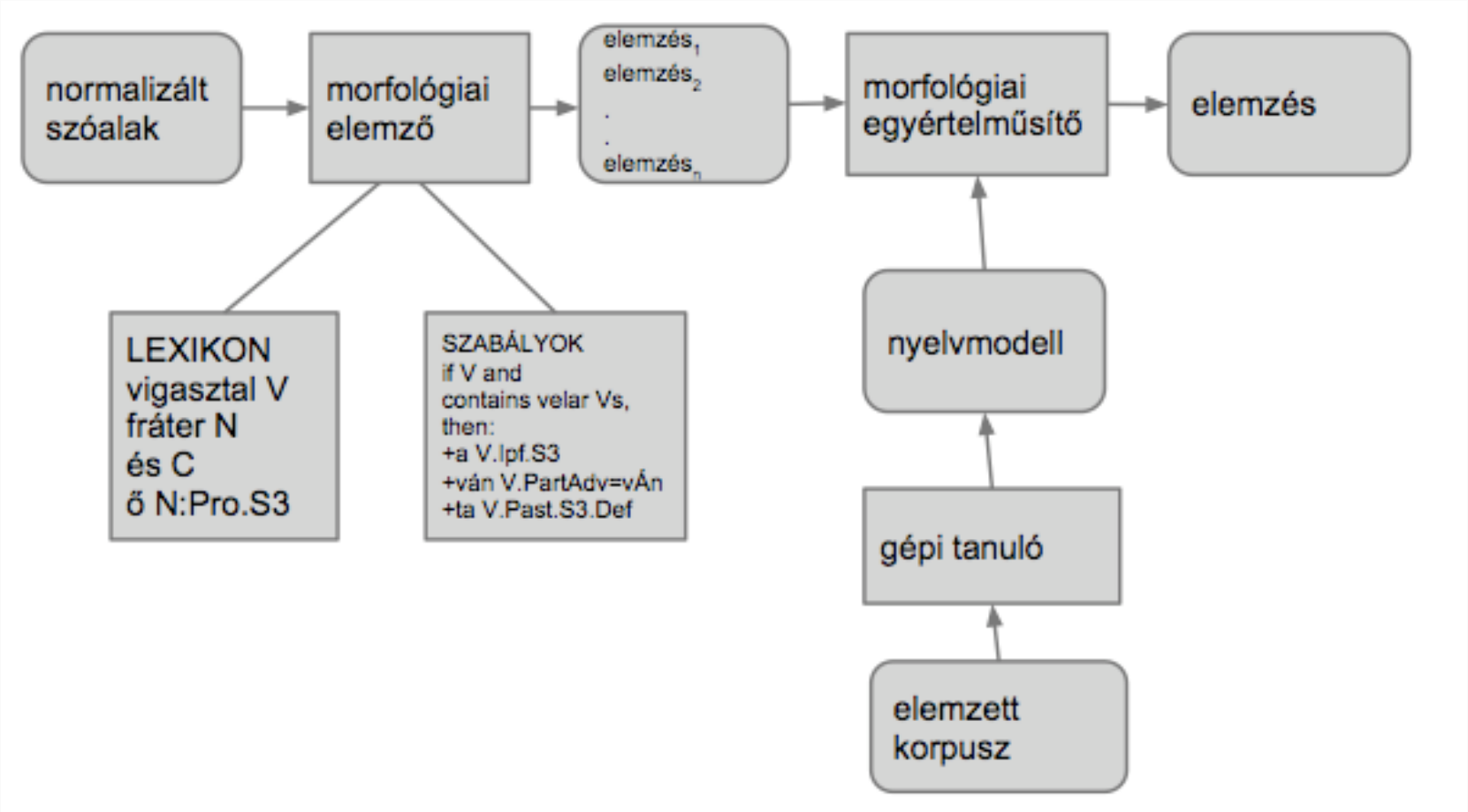
falucska

```
fa [/N] + luc[/N] + ska[/N] + [Nom]
fa[/N] + lucsok[/N]=lucsk + a[Poss.3Sg] + [Nom]
falu[/N] + cska[_Dim:cskA/N] + [Nom]
falucsok[/N]=falucsk + a[Poss.3Sg] + [Nom]
falucska[/N] + [Nom]
```

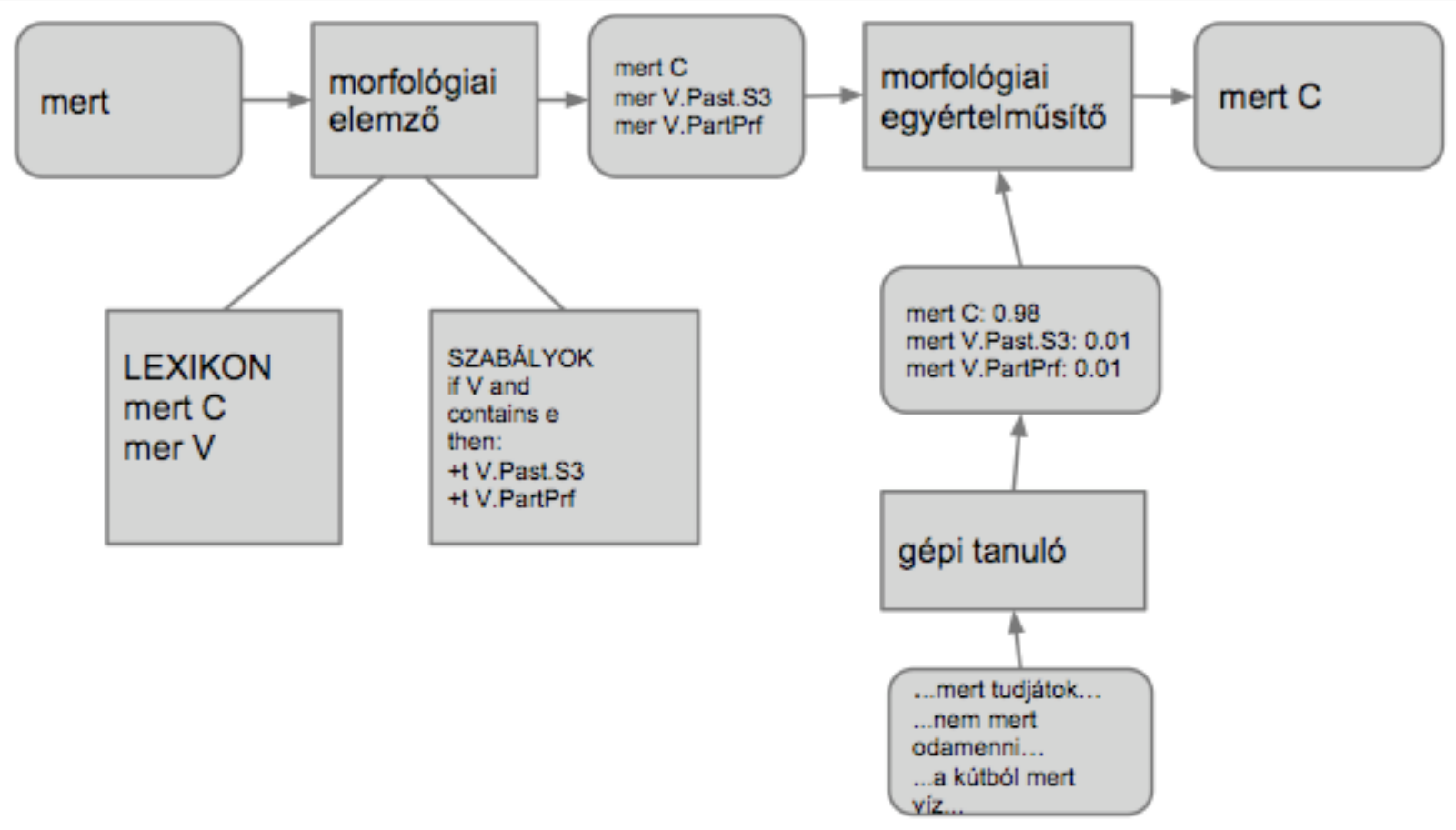
Mit tartalmazhat a kimenet?

- morfoszintaktikai információk
- jelentésre vonatkozó információk
- hangalakra vonatkozó információk (allomorfia)
- szófajkód
- lemma
- morfológiai szegmentumok

Morfológiai egyértelműsítés 1.



Morfológiai egyértelműsítés 2.



Nézzük meg az e-magyart!

Tulajdonnév-felismerés

Named Entity Recognition (NER)

2 lépésből áll:

1. a nevek lokalizálása strukturálatlan szövegben
2. a megtalált elemek besorolása előre definiált névosztályokba
   - *Person, Location, Organization, Date, Time, Money, Percent, Measure* (MUC)
   - *Person, Location, Organization, Miscellaneous* (CoNLL)

Kérdések, nehézségek

- a tulajdonnevek definiálása problémás
- egymásba ágyazott nevek és kompozicionalitás
- van-e a tulajdonnévnek jelentése?
- a tulajdonnevek a szintaxis szempontjából oszthatatlan nyelvi egységek
- nem lehet belülről módosítani őket
- a ragok mindig az NP-t alkotó tulajdonnév végére kerülnek
- a tulajdonnevek alaki sérthetetlenségének elve
- metonimikusan viselkedő tulajdonnevek
- eltérő annotációs sémák $\rightarrow$ még a statisztikai alapú rendszereket is nehéz átvinni egyik korpuszról a másikra, vagy egyik műfajról a másikra

Sekély szintaktikai elemzés

chunking

[Immár] [negyedik éve] [a Manchester United]
[a világ leggazdagabb csapata] [bevétel szerint].

1. minden frázis megtalálása egy mondatban
2. maximális NP-k megtalálása
3. alap NP-k megtalálása

Összetevős és függőségi elemzés

Összetevős elemzés

A mondatok összetevős szerkezeti elemzése azt tárja fel, hogy a szavak egymással kombinálódva milyen kifejezéseket alkotnak, illetve hogyan állnak össze egy mondattá.

Függőségi elemzés

A függőségi elemzés a mondatok szerkezeti egységei közötti függőségi viszonyokat (pl. alany, tárgy, jelző) tárja fel.

Összetevős és függőségi szintaktikai elemző

- kétféle elméleti keret szerint
- függőségi elemzés: Bohnet parser alapján
- összetevős elemzés: Berkeley parser alapján
- tanító adat: Szeged (Dependencia) Treebank
- bemenet: morfológiai egyértelműsítő kimenete
- kimenet: CoNLL formátum (függőségi elemzés), Berkeley kimeneti formátuma

Hogyan működik az elemző?

Irodalom

Hivatkozások

Babarczy, A., Bencze, I., Fekete, I., and Simon, E. (2010). The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-based Analysis. In Bel, N., Daille, B., and Vasiljevs, A., editors, *Proceedings of the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, pages 31–36, Malta.

Hunston, S. (2008). Collection strategies and design decisions. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, pages 154–167. Walter de Gruyter, Berlin.

Kugler, N. and Tolcsvai Nagy, G., editors (2000). *Nyelvi fogalmak kisszótára*. Korona, Budapest.

Irodalom ii

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Lüdeling, A. and Kytö, M., editors (2008). *Corpus Linguistics. An International Handbook*. Walter de Gruyter, Berlin.

McEnery, A. and Xiao, R. (2007). *Parallel and comparable corpora: What are they up to?* Translating Europe. Multilingual Matters.

McEnery, T. (2004). Corpus Linguistics. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 448–463. Oxford University Press, New York.

Mittelholcz, I. (2017). ***emToken***: Unicode-képes tokenizáló magyar nyelvre. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 61–69, Szeged.

O'Keeffe, A. and McCarthy, M., editors (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge, London and New York.

Irodalom iii

Simon, E. and Nemeskey, D. M. (2012). Automatically generated NE tagged corpora for English and Hungarian. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 38–46, Jeju, Korea. Association for Computational Linguistics.

Sinclair, J. (2005). Corpus and Text – Basic Principles. In Wynne, M., editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 1–16. Oxbow Books, Oxford.

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., and Csirik, J. (2006). A highly accurate Named Entity corpus for Hungarian. In *Electronic Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Szirmai, M. (2005). *Bevezetés a korpusznyelvészetbe. A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában*. Tinta Könyvkiadó, Budapest.