Bevezetés a korpuszok csodálatos világába

A számítógépes nyelvészet alapjai – 2022/23 tavasz
4. óra

Simon Eszter

2023. március 20.

Tartalom

1. Mi a korpusz?

2. Korpusztipológia

3. Főbb kérdések a korpuszépítésnél

4. A korpusz mérete

5. Korpuszlekérdezés

6. Néhány hazai korpusz lekérdezőfelülettel

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)
- dinamikus (COBUILD 1980 óta)

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?

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

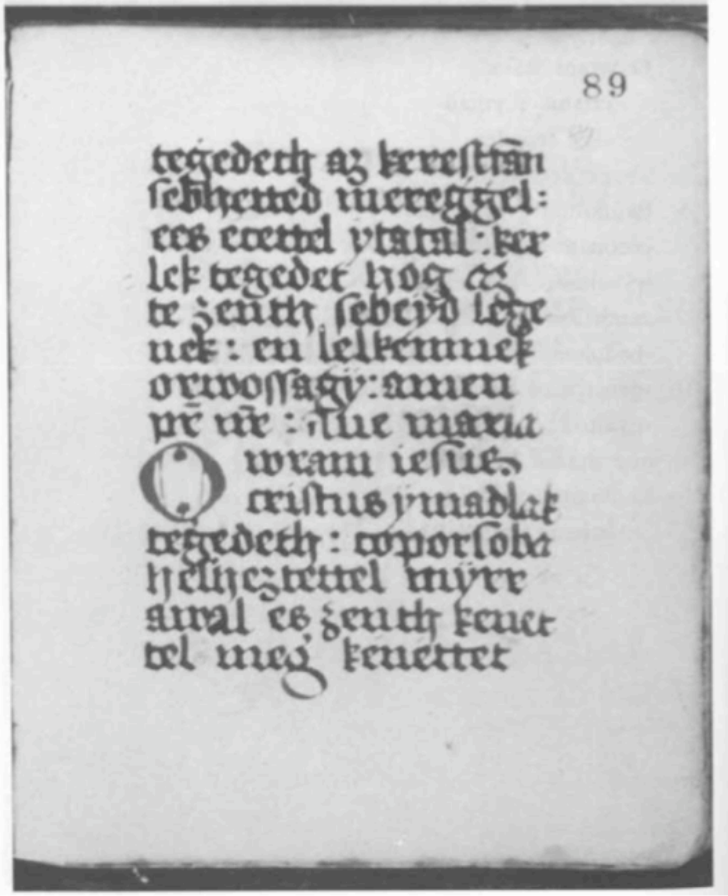
- a kutatás tárgya határozza meg a korpusz összetételét
- minél jobban körülhatárolható a kutatási kérdés, annál könnyebb döntéseket hozni a korpusz tartalmáról
- korai korpuszok: amerikai angol általános korpusz (Brown Corpus) és brit angol általános korpusz (Lancaster-Oslo/Bergen Corpus (LOB))
- mindkettőbe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

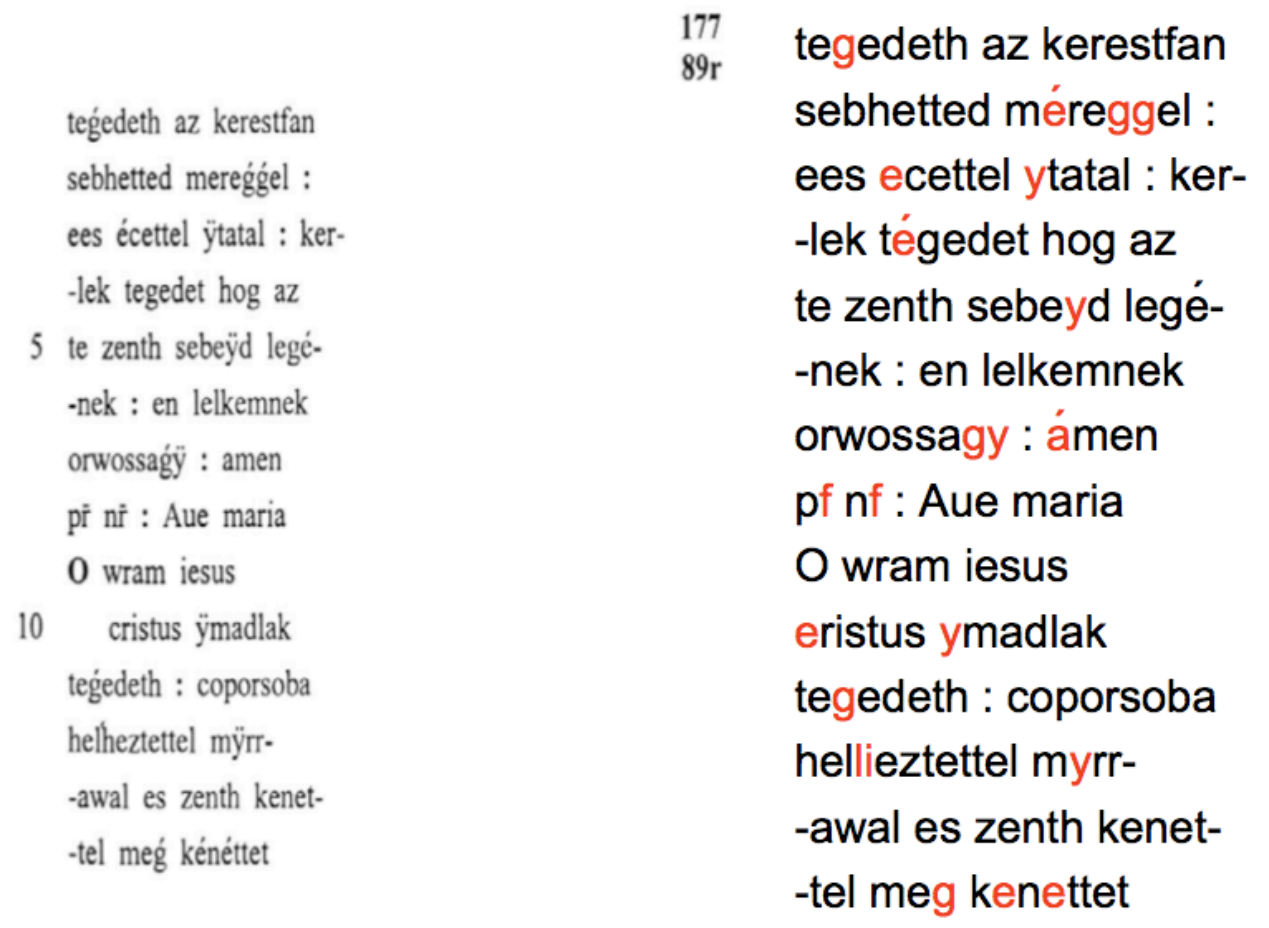
Kézzel írott kódexek, nyomtatott kiadások



177
89r

teǵedeth az kerestfan
sebhetted mereǵǵel :
ees écettel ÿtatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pr̃ nr̃ : Aue maria
O wram iesus
cristus ÿmadlak
teǵedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meǵ kénéttet

Szkennelés, OCR



tegedeth az kerestfan
sebhetted meregggel :
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

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:

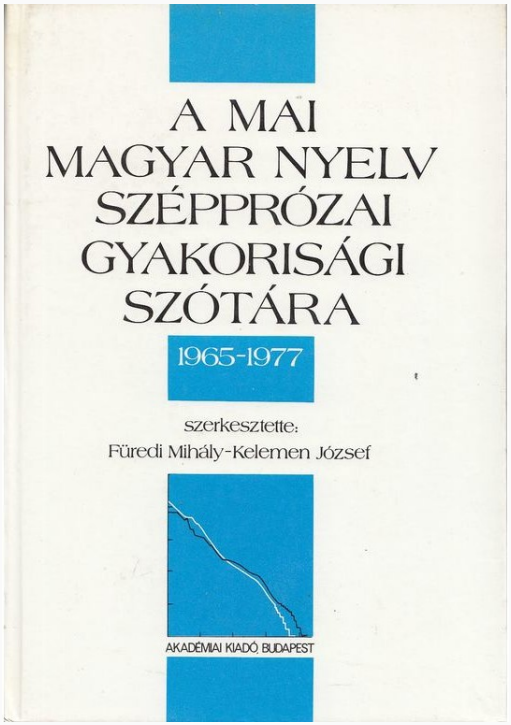
- a *q* betűt majdnem mindig *u* betű követi
- a szöveg kicsit több mint 60%-a mássalhangzó
- a köznapi beszédben használt szótagszerkezetnek kb. az egyharmada CVC szekvencia
- a nyelv 50 leggyakrabban használt szava teszi ki a szövegek 45%-át

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



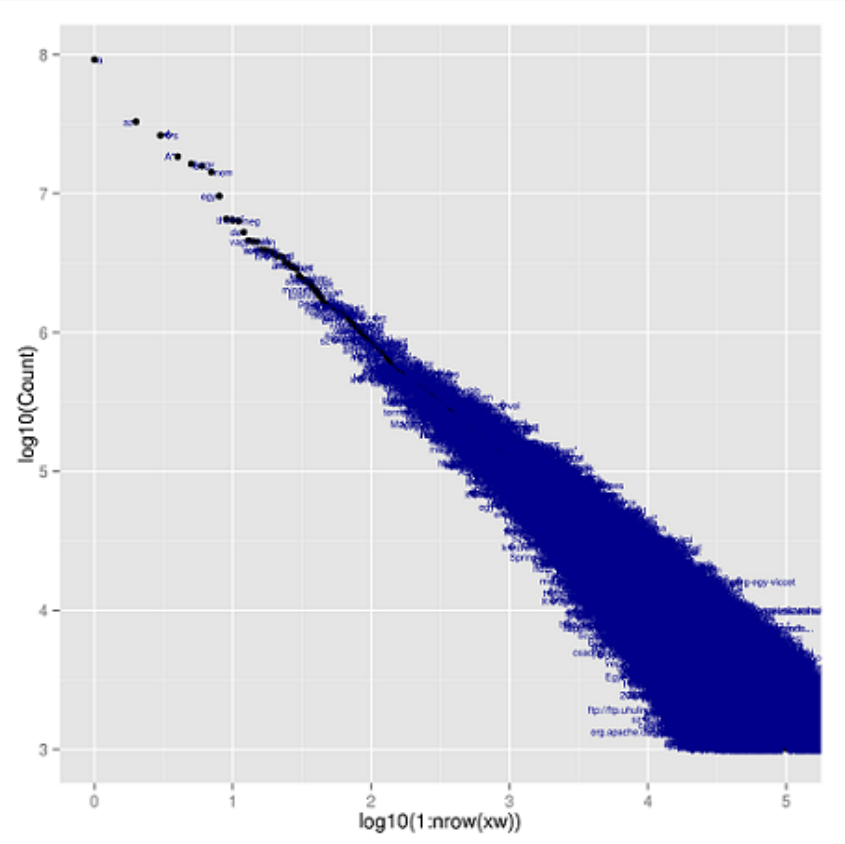
Zipf-törvény

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

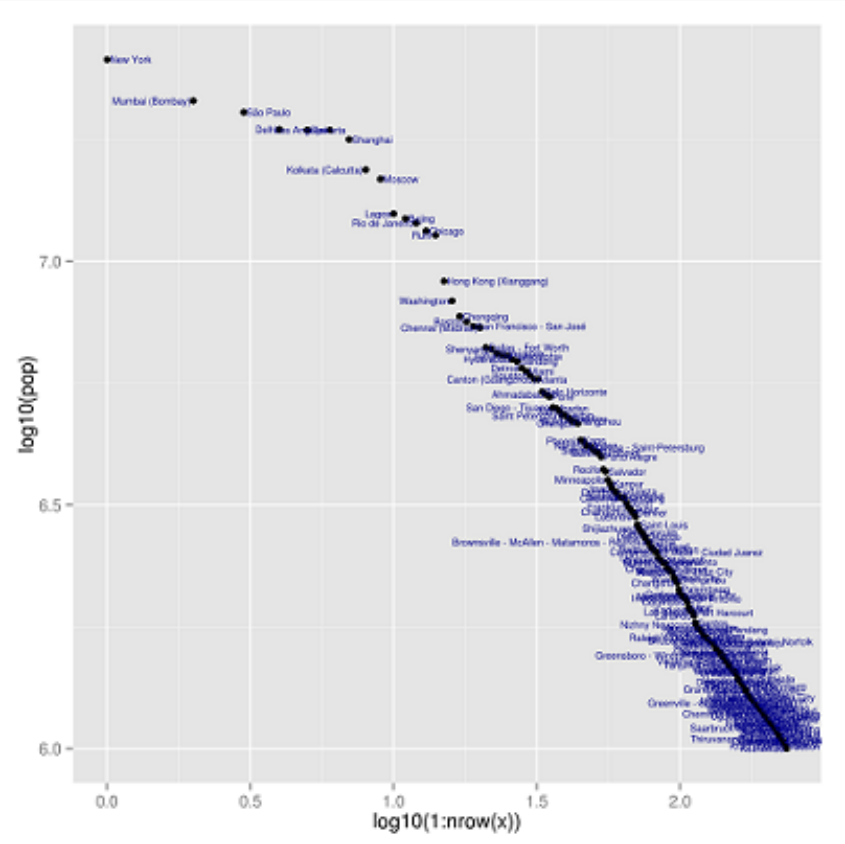
A Magyar Webkorpusz 10 000 leggyakoribb elemét mutatja az alábbi grafikon (a vízszintes tengelyen a frekvenciatáblában elfoglalt pozíciót, a függőlegesen pedig a gyakorisági értéket mutatjuk).



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

Zipf-görbe a nyelvtudományon kívül

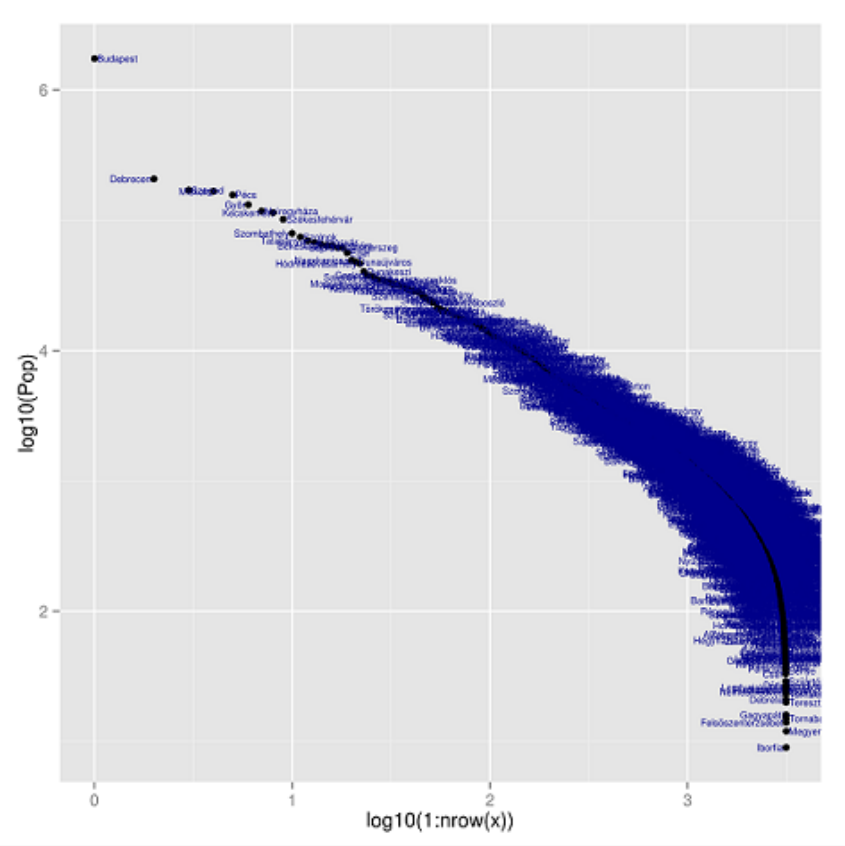
A nagyobb világvárosok lélekszáma és a lakosság szerinti sorrendben elfoglalt pozíció közötti fordított arányosság.



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

Korpuszlekérdezés

Terminológia

- nyers (*raw*) korpusz: szóalakok
- annotált korpusz: valamiféle annotáció kapcsolódik a tokenekhez vagy tokenszekvenciákhoz
- konkordancia: a lekrdezés eredménye
- találat (*hit*): kwic + kontextus
- kwic: *keyword in context*: a lekérdezett kifejezés
- ablak (*window*): a tokensorozat egy szakasza, amit kezdő- és végpozícióval definiálunk a kwic-hez képest (pl. -1...2: 4 szó, a kwic, egy szó balra és két szó jobba tőle)
- lekérdezőrendszer (*corpus query system, CQS*): a rendszer, amin keresztül a júzer konkordanciákat nyerhet ki lekérdezésekkel (pl. NoSketchEngine (NoSke), Emdros)
- formális korpuszlekérdező nyelv (*formal query language*) (pl. CQL, MQL)

Reguláris kifejezések

joker-karakter: `.`

mennyiségjelzők: `? + * { }`

választás: `| []`

csoportosítás: `( )`

Példák:

- `a|b*`
- `gr(a|e)y, gr[ae]y`
- `b[aeiou]bble`
- `colou?r`
- `go+gle`

Néhány hazai korpusz lekérdezőfelülettel

Magyar Nemzeti Szövegtár (MNSZ)

- 1998-ban kezdték építeni
- összetétele: http://clara.nytud.hu/mnsz2-dev/stat.html
- MNSZ2 1,04 milliárd szövegszó
- hat stílusréteg, öt regionális nyelvváltozat
- 76 millió szavas beszéltnyelvi (rádiós) alkorpusz felolvasott szöveges tartalommal és spontán beszéddel
- személyes alkorpusz: fórum (57,9 millió szó), közösségi (243,2 millió szó)

MNSZ2

- 1,04 milliárd szövegszó (1,348 milliárd token)
- használatához hozzáférést kell igényelni
- morfoszintaktikai kódok:
  http://clara.nytud.hu/mnsz2-dev/msd.html

https://www.youtube.com/@magyarnemzetiszovegtar

Mazsola

- igék és argumentumaik közvetlen vizsgálata
- kollokációk a vonzatkeretek feltárásához
- 2006-2009 között fejlesztették az MNSZ anyagát felhasználva
- gyakorisági vonzatkeret-szótár alapjául szolgált (Magyar igei szerkezetek: http://www.tintakiado.hu/book_view.php?id=286)
- anyanyelvi nevelés, magyar mint idegen nyelv
- lexikográfia
- nyelvészeti kutatások (gyakoriság, szemantikai osztályozás bővítménykeret alapján, szemantikai szelekció stb.)
- http://corpus.nytud.hu/mazsola/s/mazsola_hun.html

Magyar Történeti Szövegtár

- jelenleg 30 millió szövegszót tartalmaz
- eredetileg: 1772 és 2000 között keletkezett különböző műfajú és stílusú szövegek gyűjteménye
- 2015-ben 3 millió szövegszónyi 2001 és 2010 közötti szemelvénnyel egészült ki, megtartva a regiszterek arányait
- részletes bibliográfiai adatokat is megjelenít az egyes találatok mellett
- gyakorisági listák, a lekérdezések szűrése, kollokációk keresése
- a Unicode szerinti karakterkódolás miatt a 18. században még gyakori régi grafémák is eredeti formájukban jelennek meg
- http://clara.nytud.hu/mtsz/

Ómagyar Korpusz (MGTSZ)

- 2009-2013, 2015-2018
- annotált korpusz, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) és néhány középmagyar kori (1526–1772) szövegemléket, és amely nyelvészetileg releváns kérdésekre tud választ szolgáltatni
- 3,2 millió szövegszó: 47 ómagyar kódex, 24 rövidebb ómagyar szövegemlék, 244 misszilis és 5 középmagyar kori bibliafordítás
- http://omagyarkorpusz.nytud.hu/hu-descr.html

Párhuzamos Bibliaolvasó

- Párhuzamos Bibliakorpusz:
  - bibliafordítások a magyar nyelv különböző korszakaiból
  - egyéb uráli nyelvek bibliafordításai
  - King James Bible az angol nyelvű glosszázáshoz
- https://parallelbible.nytud.hu/

Történeti magánéleti korpusz (TMK)

- az ó- és középmagyar kor magánéleti nyelvi regiszteréhez legközelebb álló műfajokat tartalmazza
- 1772 előtti magánlevelek és peres eljárások jegyzőkönyvei
- történeti morfológiai és szociolingvisztikai, történeti mondattani, pragmatikai és lexikológiai kutatásokhoz
- 8.6 millió karakter (magyar nyelvű rész: 7,68 millió karakter, 1 millió 112 ezer elemzett szövegszó) (2020 szeptemberi adat)
- https://tmk.nytud.hu/3/
- Emdros, MQL
- útmutató a kereséshez: https://tmk.nytud.hu/utmutato.php

Házi feladat

Házi feladat

- Válassz ki egy nyelvi jelenséget!
- Végezz korpuszalapú viszgálatot arra a nyelvi jelenségre vonatkozóan!
- Írd le, hogy melyik korpuszon mit csináltál, mi volt a hipotézis, és mire jutottál!

Ez túl bonyi, de érdekes:
https://telex.hu/tudomany/2023/03/05/nem-tudom

Irodalom

Hivatkozások

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.

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.

Irodalom ii

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.