Zipf

Af Hans Degn 12. december 2011 • I kategorien Teknologi •

Zipf er hidtil blevet nævnt én gang i Sprogmuseet. Det er ikke meget i betragtning af, at Zipf er en af sprogforskningens hjørnestene. Det drejer sig om det kvantitative hjørne; det hjørne, der er penge i.

Toldkontrollør Knagsted (1858-1914)

Kommatæling

Toldkontrollør Knagsted samlede på kommaer. Han læste ikke bøger, men talte kommaerne i dem. Her fortæller han om det til overlærer Clausen:

- Hæ, hæ! ... Næ, man skal samle paa Kommaer, skal man! Dem er man nødt til selv at slide for!

- *Samler* du da virkelig paa Kommaer? spurgte Overlæreren vantro.

- Gu' gør jeg saa ja! Her skal du se ... (Knagsted trak sin Lommebog frem og bladede i den) her er de sidste Resultater.

Han pegede ned paa et Blad og læste:

Ewald: Fiskerne. 27,335. Balders Død: 45,860. De brutale Klappere: 39,022 ... Jeg samler jo kun paa *danske* Forfattere, forstaar du. Man skal være Specialist. Og det andet var jo heller ikke til at overkomme! ... Og naar jeg nu er færdig med Ewald, vedblev han – saa tæller jeg ham sammen. Det er i Grunden det morsomste! Holberg gi'er over seks Millioner, du! ... Men man skal være forfærdelig omhyggelig; for det er saa irriterende, naar det ikke stemmer anden Gang!

- Tæller du dem *to* Gange? spurgte Pædagogen med Øjne som Desserttallerkener.

- Hi-ja-a! Og anden Gang bagfra! Clausen faldt i Tanker. Det maatte være et Pokkers rart Arbejde at sidde og pusle med i de lange Vinteraftener.

....

(Clausen tager afsked med sin gæst).

....

Saa gik han tilbage ind i Dagligstuen. Midt paa Gulvet blev han et Øjeblik staaende og lod tankefuldt sine Fingre løbe op gennem Haaret. Derpaa tog han et langt Skridt frem mod Reolen og snappede en Bog ned fra den næstøverste Hylde. Og smilende, halvt forventningsfuldt og halvt genert, satte han sig med Bogen hen ved sit Skrivebord.

Han havde faaet en aldeles uimodstaaelig Lyst til at se efter, hvor mange Udraabstegn der var i Shakespeares Richard den Tredje ...

(Gustav Wied: *Livsens Ondskab* (1899))

Hvis toldkontrolløren eller rettere hans *alter ego* også havde talt antallet af ord og beregnet forholdet mellem antallet af kommaer og antallet af ord, ville han være blevet en af grundlæggerne af den kvantitative lingvistik. Men han overskred ikke den stiplede linje mellem ekcentricitet og genialitet, så det blev han ikke. Det blev som nævnt Zipf.

Jeg ved ikke, hvor Gustav Wied fik ideen til Knagsteds kommahobby. Gustav Wied samlede selv på spadserestokke, men en freudiansk tydning af Knagsteds kommaer, overlærer Clausens udråbstegn og Wieds spadserestokke passer ikke tidsmæssigt, da Freuds store værker først udkom fra 1900 og fremefter. Wied kan vel have foregrebet Freud, men en anden mulighed er mere sandsynlig.

Ordtælling

Ordtælling var ikke en ukendt idé i slutningen af det nittende århundrede. Allerede i 1851 foreslog den engelske matematiker A. de Morgan, at man kunne bruge ordstatistik til at afgøre, om Paulus har skrevet alt det, der i Bibelen er tillagt ham. De Morgan forfulgte ikke selv ideen, men andre gjorde, og nogle gik i gang med at bruge metoden til at finde ud af, hvem der havde skrevet Shakespeares værker, for Shakespeare kunne det jo ikke være. Den såkaldte stylometri tog dermed sin begyndelse.

Det første værk, der angreb et helt sprog med ord-, stavelses- og lydoptællinger i et gigantisk tekstkorpus, var *Häufigkeitswörterbuch der deutschen Sprache*, der blev udgivet af F. W. Kaeding i 1898, året før udgivelsen af *Livsens Ondskab*. Måske er Knagsteds kommasamling

Seneste sprognyheder

6/3	Dansklærer: Tosprogede børn bliver sprogligt forsømt | www.bt.dk
28/2	Flere og flere ordblinde starter på universitetet - Magisterbladet | magisterbladet.dk
22/2	Keeper eller målmand — hvilket ord er bedst? Dansk Sprognævn | dsn.dk
17/12	Snebajer og nytårsskrald: Juleord kan slås op i netordbog - Jubii | www.jubii.dk
24/11	Meet the Last Speaker of a Dying Language | video.nationalgeographic.com
30/6	Irish-language theatre: underground but still alive | www.irishtimes.com
29/6	Celebrating two years of language renewal - Princeton Similkameen Spotlight | www.similkameenspotlight.com
28/6	Linguistics Breakthrough Heralds Machine Translation for Thousands of Rare Languages - MIT Technology Review | www.technologyreview.com

FLERE NYHEDER >>>

Verdens sprog på Sprogmuseet.dk på et større kort

Mere i kategorien 'Teknologi'

SMS – en ny vår for afrikanske språk?
En e-bog er en ny slags bog!
Maskinoversættelse – en sammenligning af to forskellige metoder
IKEAs Body Mass Index når et kritisk punkt

Nye kommentarer

Ruben Schachtenhaufen til Er det forkert at splitte sammensatte ord?
Niels Larsen-Ledet til Er det forkert at splitte sammensatte ord?
william fich til Jødiske efternavne
Jens Michael Kofod-Hansen til Nytårsfortsæt
Jens Michael Kofod-Hansen til Nytårsfortsæt
Herluf Hansen til Ded borrijnholmska måled
Arturo til Hebraisk: Et genoplivet sprog eller et nyt sprog?
Yunus til Two Turkish Loanwords in Swedish
Sonstige til Den Danske Ordbog på nettet
Mads Haupt til Er det forkert at splitte sammensatte ord?

Artikler om


Häufigkeitswörterbuch
der deutschen Sprache.
J. W. Kaeding.
Erster Teil.
Titelbladet til Kaedings Häufigkeitswörtwerbuch fra 1898. Klik for læselige krøllede bogstaver

blot en satire over optælling af tekstelementer.

F. W. Kaeding (1843-1916) var stenograf og engageret i stenografiens udvikling. Formålet med hans værk var at producere statistisk materiale, der kunne bruges til at optimere det Stolzeske stenografisystem. Det enorme statistiske materiale i *Häufigkeitswörtwerbuch*, som er baseret på tekster med tilsammen 11 mil. ord og har beskæftiget hundredevis af optællere, er ikke ledsaget af teoretiske overvejelser.

Om Kaedings værk var til nogen større nytte for stenografien er et åbent spørgsmål, men værket har været af stor betydning for den kvantitative lingvistik. I sin banebrydende bog *The Psychobiology of Language (1935)* angav G. K. Zipf Kaedings værk som en af de vigtigste kilder, han selv havde benyttet. Zipf blev den kvantitative lingvistiks egentlige grundlægger fordi han søgte matematiske formuleringer af de statistiske sammenhænge, han fandt i sine egne ordtællinger, og fordi han søgte at finde årsagerne til disse sammenhænge.


"Zipf belongs among those rare but stimulating men whose failures are more profitable than most men's successes." George A. Miller

Zipfs lov

George Kingsley Zipf (1902-1950) var *university lecturer* ved Harvard Universitetet.

Helliitersglas til Zipf-øl fra Tauberbischofheims Bräuerei til salg på ebay for 5 €

Hans bedstefar var indvandret fra Tauberbischofsheim i Tyskland, hvor familien havde et bryggeri. Efter at have gået på Harvard College tog Zipf i 1926 på studieophold i Tyskland, hvor han fik ideen til at undersøge sprog som et naturfænomen. Han fik ph.d. graden ved Harvard University i 1929 på en afhandling om *Relative Frequency as a Determinant of Phonetic Change* og blev tilknyttet Harvard som underviser i tysk.

Zipfs sprogstatistiske undersøgelser fandt sted et halvt århundrede før nogen opdagede, at der er penge i skidtet. Han overtalte sin hustru til at spare på husholdningspengene, så han kunne få råd til for egen regning at ansætte nogle studentermedhjælpere til at tælle hyppigheden af ord i tekster på forskellige sprog. Resultaterne var nok værd at gå halvsulten i seng for. Han fandt, at der på alle de undersøgte sprog var en simpel sammenhæng mellem et ords hyppighed og dets rang (Definitionen på rang er, at det hyppigst forekommende ord har rang 1, det næsthyppigste rang 2 o.s.v.). Zipfs opdagelse, der nu kaldes Zipfs lov, er, at hyppigheden af et ord er omvendt proportionalt med ordets rang. Det indebærer, at hyppigheden af nummer to i rangordenen er halvdelen af hyppigheden af nummer et. Hyppigheden af nummer tre i rangordenen er en tredjedel af hyppigheden af nummer et o.s.v.. Denne fordeling afspejler, at et forholdsvis lille antal ord bruges hyppigt, mens det store flertal af ord bruges sjældent.

Zipfplot

Den nemmeste metode til at afgøre om nogle data følger Zipfs lov er at afsætte dataene i en graf. Hvis man afsætter dem direkte, hyppighed ud ad y-aksen og rang ud ad x-aksen (eller omvendt) får man en krum kurve, som skal være en perfekt hyperbel, hvis dataene opfylder Zipfs lov. Det er svært at se hvor tæt sådan en kurve er på at være en hyperbel, så man bruger et simpelt matematisk kunstgreb og afsætter i stedet logaritmen til hyppigheden mod logaritmen til rangen. Et sådant logaritmisk plot, et zipfplot, forvandler en hyperbel til en ret linie med hældningen -1 (45^o), og det er nemt at se, hvor tæt de afsatte punkter er på at udgøre en sådan linje.

Figuren viser et eksempel på et zipfplot taget fra Zipfs hovedværk *Human Behavior and the Principle of Least Effort* (1947). Kurve A viser data fra James Joyces roman *Ulysses*, der indeholder 260.430 ord, hvoraf 29.899 er forskellige. Kurve B viser data fra en samling af amerikanske avisartikler på 43.989 ord, hvoraf 6002 er forskellige. Kurve C er en ret linie med hældningen -1 indtegnet til sammenligning. Tallet 10000 på y-aksen er placeret forkert. Som det kan ses er de to kurver gode tilnærmelser til den rette linie. Der findes i litteraturen et stort antal zipfplot af data fra andre

aktuelle sprog Alfabeter Anmeldelser arabisk Biblen bogstaver børn Danmark Dansk Dialekter engelsk esperanto Formidling fransk identitet konsonanter Medier modersmål Musik Navne norsk Ord ordbøger ordforråd oversættelse Plansprog religion romanske sprog russisk Sjov skriftsprog sprogdød Sproggeografi sprogkort Sprogpolitik sprogteknologi svensk truede sprog tv tyrkisk tysk Udtale Underholdning video vokaler

Arkiv

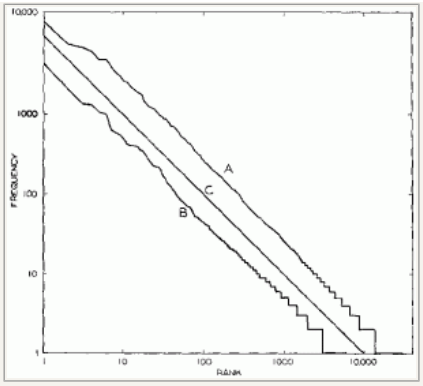
januar 2015
december 2014
november 2014
maj 2014
marts 2014
februar 2014
oktober 2013
august 2013
marts 2013
januar 2013
december 2012
november 2012
oktober 2012
september 2012
juli 2012
juni 2012
maj 2012
april 2012
marts 2012
februar 2012
januar 2012
december 2011
november 2011
oktober 2011
september 2011
august 2011
juli 2011
juni 2011
maj 2011
april 2011
marts 2011
februar 2011
januar 2011
december 2010
november 2010
oktober 2010
september 2010
juni 2010
maj 2010
april 2010

Resources

Ethnologue: Languages of the World
Forvo – All the Words in the World. Pronounced.
LL-Map: Language and Location
Minority Rights Group
Omniglot. Writing Systems and Languages of the World
UNESCO Atlas of the World's Languages in Danger
World Atlas of Linguistic Structures (WALS)

Resurser

Bogstavlyd
Dansk sprognævn
Den danske ordbog
Dialekt.dk
dk.kultur.sprog
Korpus.dk
Nye ord i dansk på nettet (NOID)
Ordbog over det danske sprog
Ordnet. Dansk sprog i ordbøger og korpus
Sproget.dk
Svenska Akademien
ə (Schwa.dk)



Zipfplot af "Ulysses" (A) og amerikanske aviser (B). De opfører sig nogenlunde som den ideelle Zips lov (C).

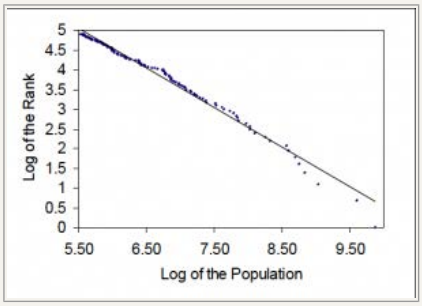
sprog, og de stemmer med få undtagelser med Zipfs lov.

Zipfs programerklæring var, at sproget er et resultat af den biologiske evolution og derfor skal undersøges og beskrives med naturvidenskabelige metoder. Figuren viser, at han var godt på vej. Han havde målt på sit objekt og fundet at resultaterne passede med en simpel matematisk ligning. Men det kiksede, da han ville forklare sin opdagelse. Han mente, at sprogets udvikling er styret af et naturgivet princip, *the principle of least effort*, hvilket oversat til lidt firkantet dansk bliver princippet om mindste anstrengelse. Dette princip, hævdede Zipf, ville minimere summen af anstrengelse hos den talende og den lyttende. Den talendes anstrengelse vil være mindst, hvis han har så få ord som muligt at vælge imellem, egentlig helst kun et, der kan udtrykke alle meninger, mens den lyttendes anstrengelse vil være mindst, hvis der er et ord for hver eneste mening. Disse modstridende tendenser, *unifikation / diversifikation*, vil føre til *vokabulær balance*, som viser sig ved, at sproget adlyder Zipfs lov.

Hvis nogen har troet, at vokabulær balance betød letlæselighed, må *Ulysses* have skuffet dem. Men der er en vægtigere grund til at afvise Zipfs forklaring. Princippet om mindste anstrengelse er ikke nogen naturvidenskabelig forklaring. Zipfs lov kan ikke udledes matematisk af princippet om mindste anstrengelse.

Zipfs lov alle vegne

Når man har opdaget noget så fundamentalt som Zipfs lov, er det første man gør at undersøge, om der er andre, der allerede har gjort den samme opdagelse. Hvis man finder ud af, at det er der, kan man anerkende det eller fortrænge det. Vi ved ikke, hvornår Zipf blev bekendt med, at den omvendte proportionalitet mellem ordhyppighed og rang allerede var blevet observeret i 1916 af J. B. Estoup, der var stenograf ligesom Kaeding og havde samme motiv til at beskæftige sig med den slags ting. Loven kom til at bære Zipfs navn fordi han arbejdede på den i mange år og gjorde den kendt gennem sit forfatterskab. Nogle få forfattere med en højere retfærdighedssans, kalder den for Estoup-Zipfs lov.



Antal indbyggere versus rang for 135 storbyer i USA

Ikke alene var Zipfs lov blevet opdaget før Zipf gjorde det, der var også tidlige rapporter, der viste, at lovens gyldighedsområde strakte sig ud over lingvistikkens grænser. Det mest prominente eksempel var den hyperbolske sammenhæng mellem storbyers indbyggerantal og deres rang, som blev opdaget af F. Auerbach i 1913. Figuren viser et zipfplot af rang *versus* indbyggerantal for amerikanske storbyer i 1991. Overensstemmelsen med den rette linie er forbavsende god. Et andet tidligt eksempel blev rapporteret af V. Pareto, der i 1906 fandt, at der er hyperbolsk sammenhæng mellem virksomheders overskud og deres rang i henhold til overskud (Paretos lov). Paretos lov har også vist sig at gælde for mange andre økonomiske sammenhænge.

Informationsteknologien

Efterhånden viede Zipf mere og mere af sin opmærksomhed til de mange ekstralingvistiske eksempler, der myldrede frem, og han prøvede at se dem alle i lyset af sit *principle of least effort*. Det førte ham ud i omfattende, i samtiden inspirerende, spekulationer om især sociologiske sammenhænge. I modsætning til adskillige andre minimeringsprincipper er *the principle of least effort* ikke holdbart i matematisk-naturvidenskabelig tænkemåde. Man kan stadig se guruer og coaches, der markedsfører sig med henvisning til Zipfs princip.

De, der i Zipfs levetid var toneangivende i lingvistikken, tog ingen videre notits af ham. Hans gennembrud kom lang tid efter hans død, da der opstod en teknologi, der havde brug for hans resultater. Det var informationsteknologien, hvis hovedbeskæftigelse er at gøre noget ved tekster som for eksempel at oplagre dem og finde dem igen, når nogen spørger. Den sidstnævnte aktivitet hedder at google. I informationsteknologien bekymrer man sig ikke om de klassiske lingvistiske dyder eller ordenes betydning, men man går meget op i at lagre teksterne på en sådan måde, at de kan genfindes på kortest mulig tid, de hyppigst rekvirerede hurtigst. Her er sprogstatistik det vigtigste teoretiske værktøj.

For at illustrere Zipfs betydning for informationsteknologien har jeg brugt Google til at tælle hvor mange dokumenter der er kommet på nettet inden for det seneste år, hvor udtrykket “Zipf's law” forekommer sammen med et af informationsteknologiens engelske fagudtryk (Tabel 1). Til sammenligning har jeg også optalt sammentræf af “Zipf's law” og fagudtryk fra energiteknologien (Tabel 2).

marts 2010
februar 2010
januar 2010
december 2009
november 2009
oktober 2009
september 2009
august 2009
juli 2009
juni 2009
maj 2009
april 2009
marts 2009

Tabel 1
Sammentræf af udtrykket "zipf's law" og fagudtryk fra **informationsteknologien** optalt i Google (seneste år).

internet	423
information retrieval	317
search engine	288
machine learning	259
data mining	247
artificial intelligence	211
information technology	199
proxy	178
cache	158
n-gram	150
machine translation	130
cryptograhy	101

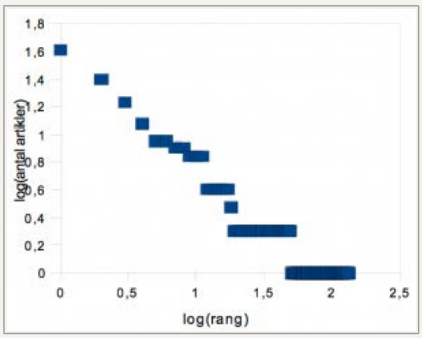
Tabel 2
Sammentræf af udtrykket "zipf's law" og fagudtryk fra **energiteknologien** optalt i Google (seneste år).

power consumption	32
power grid	19
solar energy	12
oil industry	5
fuel cell	5
wind power	5
electric car	5
biofuel	3
energy storage	3
hybrid car	2
wave energy	1
energy crop	0

Af tabel 1 kan man se, at der hvert år dukker flere tusinde dokumenter op på nettet, hvor Zipfs lov indgår i diskussioner af snart sagt alle aspekter af informationsteknologien. Af tabel 2. fremgår det, at det samme ikke er tilfældet med energiteknologien. Når Zipfs lov har så stor betydning for en af vor tids største pengemaskiner, Internettet, så kan det ikke være helt dumt at hoppe på den vogn og muligvis heller ikke kedeligt.

Man skal ikke tro, at den oprindelige udgave af Zipfs lov, der angår sammenhængen mellem ordhyppighed og rang, er den eneste, der har anvendelse i informationsteknologien. Zipfs lov gælder for mange vigtige aspekter af internettrafikken såsom antal besøg pr måned på et website og antal sider et website består af. Disse størrelser er begge omvendt proportionale med rang, hvilket har betydning for søgemaskinernes strategi.

Man skal heller ikke tro, at informationsteknologien er den eneste, der gør flittigt brug af Zipfs lov. Som allerede nævnt gælder Zipfs lov for storbyers indbyggerantal og koncerners indtjening. Det betyder, at loven indgår i det teoretiske grundlag for planlægning af byudvikling og infrastruktur samt økonomiske prognoser og investeringsstrategier.



Zipfplot af Sprogmuseets forfattere rangordnet efter antal artikler. Som man ser gælder Zipfs lov også her.

Hvad er forklaringen?

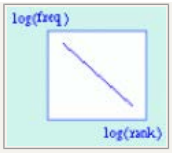
Hvis der er nogen, der ikke er overbevist om, at Zipfs lov er alle steder, kan følgende eksempel måske hjælpe dem. Sprogmuseet har en forfatteroversigt, der er en rangordnet liste over forfatterne til alle hidtidige indlæg. Den er lige til at lave et zipfplot af. Figuren til venstre viser logaritmen til antallet af en forfatters indlæg i Sprogmuseet som funktion af logaritmen til forfatterens rang. Punkterne ligger pænt på en ret linie med hældningen -1, som Zipfs lov foreskriver.

Så må vi endnu en gang stille spørgsmålet: Hvorfor opfylder dette sæt data og rigtig mange andre, der ikke har noget med hinanden at gøre, Zipfs lov? Zipf gav ikke selv en plausibel forklaring på, hvorfor loven gjaldt for ordhyppighed, og mange andre har forgæves forsøgt at forklare den i andre sammenhænge. Det kan tyde på at forklaringen ikke skal søges i det felt, hvor dataene kommer fra, men i selve statistikken. Der kan være tale om et statistisk fænomen forklædt som en naturlov. Den tanke er blevet forfulgt ved hjælp af *monkey-typing*.


Chimpanse i færd med tilfældigvis at komme til at skrive Gustav Wieds samlede værker (og enhver anden tekst), bare den får tid nok. (Foto: New York Zoological Society 1907)

Scenariet er, at man sætter en abe foran en skrivemaskine og lader den trykke på tangenterne efter forgodtbefindende i al evighed. Når aben har trykket på tangenterne i enhver mulig rækkefølge, indeholder dens *oeuvre* alt, hvad mennesker har skrevet til dato og alt, hvad de endnu ikke har skrevet, f.eks. alle Shakespeares værker, og alle de sonetter, skuespil og breve til Anne Hathaway, som han havde planlagt, men ikke fik tid til at skrive. Forsøget kan speedes lidt op ved, at man lader en virtuel abe i en computer udføre en tilfældig rækkefølge af anslag. Det er naturligvis blevet gjort mange gange. Ingen tekster frembragt på den måde tåler sammenligning med Shakespeare, men nogen siger, at de adlyder Zipfs lov, og nogen siger det modsatte. Så er vi lige vidt.

Til allersidst stiller vi et lumsk spørgsmål: Er det muligt for mennesker at udmanøvrere Zipfs lov? Tag for eksempel Zipfs lov anvendt på personlig indkomst (Paretos lov). Den indebærer, at 20% af befolkningen tjener 80% af pengene. Hvis man ad politisk eller anden vej vil udjævne menneskers indkomst, er man altså oppe imod noget, der ligner



Zipf "logo"

en naturlov. Det er måske derfor, at det aldrig rigtigt er lykkedes for nogen fra Robin Hood og fremover?

.

Reference

Wentian Lis bibliografi over 800+ videnskabelige artikler mm om Zipfs lov og beslægtede lovmæssigheder

***Hans Degn**, cand. scient.*

Læs også:

1. Klichéer på alle hylder I kvantitativ lingvistik analyserer man tekster ved at optælle, hvor hyppigt udvalgte ord og vendinger forekommer. Af de indsamlede tal forsøger man så at drage konklusioner ved brug af mere...
2. Klingon og mutsun Lt. Worf, en tam klingon fra Star Trek. Her ved årsskiftet indledte DR HD en kavalkade med 10 af de 11 Star Trek-spillefilm. De bliver vist lørdage kl 20.50. Foreløbig...
3. Ord med fordømmelsespotentiale I et tidligere indlæg her i Sprogmuseet har jeg antydet, at ordet sprogsyn ligesom ordet menneskesyn har fordømmelsespotentiale. Betegnelsen opfandt jeg til lejligheden og den er vist selvforklarende. Jeg overvejede...
4. Pædagogsprog – eller hvad? Når man googler ordet pædagogsprog finder man ingenting substantielt, kun enkelte spredte, ofte negative bemærkninger. Ikke så meget som en avisartikel. Søgning i bibliotekernes database giver heller intet resultat. Ikke...

Tagget med: hyppig, Ord, ordstatistik, rang, Shakespeare, statistik, stenografi, zipf

1 kommentar

Ruben Schachtenhaufen
12. december 2011 • 21:49

Tak for artiklen! Jeg blev tilfældigvis opmærksom på Zipfs lov første gang for kun tre dage siden da jeg ville finde ud af noget om hapakslegomena – ord der optræder netop én gang i en tekst eller et korpus. Intuitivt tænker man at måske at hapakslegomena er temmelig sjældne, men helt i overensstemmelse med Zipfs lov viser det sig at omtrent halvdelen af ordene (efter type) i et stort korpus er hapakslegomena.

Svar

Skriv en kommentar

Navn (kræves)

E-mail (kræves)

Hjemmeside

Send mig en e-mail når der kommer flere kommentarer.