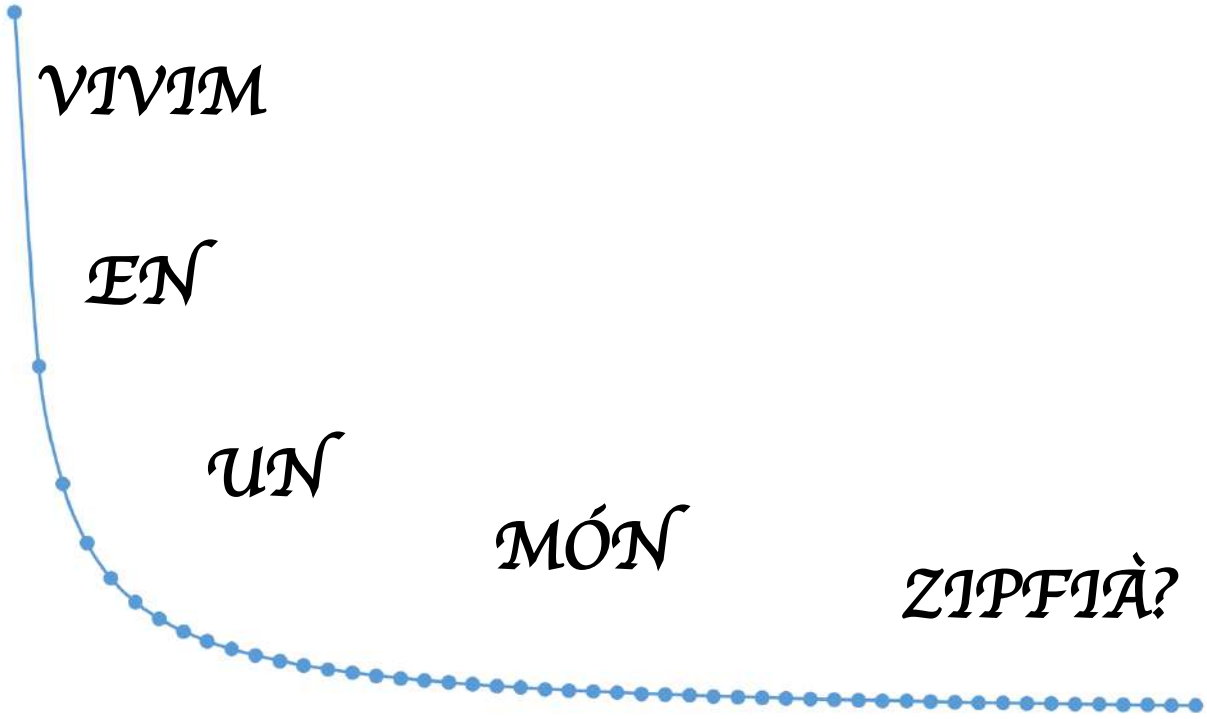
Treball de Recerca



Clement Hamilton
Turoria: Mª Àngels Picart
Ins Pere Alsius i Torrent
2016-2017

Entendre és percebre patrons - Isaiah Berlin

M’agradaria començar donant les gràcies al meu pare, a la meva mare i al meu germà pel suport i l’ajuda que m’han aportat i la paciència en tot moment.

Sens dubte, donar les gràcies a l’Àngels Picart per la direcció i tutoria en aquest treball. Sense les seves guies no hauria sabut ni per on començar.

Finalment, necessito donar les gràcies a Michael Stevens, el qual va ser la inspiració d’aquest projecte, per tot el coneixement que m’ha aportat i les noves maneres de veure els fenòmens que s’esdevenen a la vida quotidiana.

Índex

Índex de Figures

Índex de Gràfiques

Índex d'equacions

Índex de Taules

Introducció

Ja fa temps que part del meu temps lliure me'l passo a Internet, precisament a YouTube. És una plataforma on qualsevol individu pot penjar el seu vídeo sense haver de pagar. Aquest factor ha fet que una gran varietat de persones utilitzin aquesta plataforma, des de tutorials de maquillatge, fins a crítiques de productes tecnològics i jugadors professionals d'ordinador.

(https://www.youtube.com/yt/brand/media/image/YouTube-logo-full_color.png) *Fig. 1*

Uns anys més tard de l'arribada d'aquesta plataforma, va sorgir la possibilitat de monetitzar els vídeos que un penjava. Aquesta possibilitat va tenir tant d'èxit que avui en dia milers d'individus es guanyen la vida fent vídeos i compartint-los a YouTube. Un individu que freqüentment comparteix vídeos s'anomena *youtuber.*

Un *youtuber* que jo segueixo fanàticament s'anomena Michael Stevens, director del canal *Vsauce.* Stevens publica vídeos sobre una gran varietat de temes, aquests vídeos podrien ser anomenats minidocumentals.

(https://flic.kr/p/dUaEws) *Fig. 2*

Un d'aquests documentals que va publicar s'anomena *"El misteri de Zipf"*, on Stevens relacionava la llei de Zipf amb molts altres àmbits i intentava explicar el perquè d'aquest patró.

Seguidament després de la publicació va arribar el període d'elecció del meu treball de recerca. Des que era conscient que l'elecció del treball era meva, sabia que havia de triar un tema que m'agradés i em fos interessant.

Stevens havia despertat una espurna dins meu, i m'arribaren ganes d'explorar més sobre la llei de Zipf. Vet aquí el perquè d'aquest treball.

Objectius

Objectius principals:

- Entendre el concepte de la llei de Zipf
- Aprendre a buscar la llei de Zipf en un text
- Comprovar la llei de Zipf en una varietat de textos en català
- Comparar hipòtesis sobre si vivim en un món Zipfià

Objectius secundaris:

- Aprendre a llegir equacions i entendre el seu significat
- Aprendre a utilitzar el programa informàtic Excel
- Aprendre a mostrar gràficament un conjunt de dades
- Aprendre a relacionar dades utilitzant la correlació lineal i potencial

George Kingsley Zipf

George Kingsley Zipf va néixer a Freeport, Ilinois, als Estats Units el 7 de Gener, l'any 1902.

(http://www.robertbike.com/polaris/images/1910/1917_zipf.jpg) *Fig. 4*

Es va graduar a l'Universitat de Harvard l'any 1924 amb l'honor *summa cum laude*, un honor adjudicat per ser un dels millors (5%) del grau.

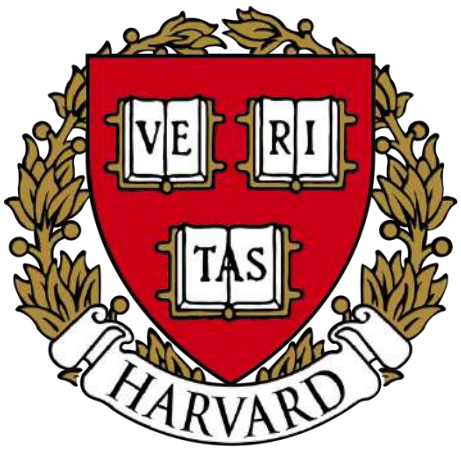


Fig. 3 *(Harvard University)*

A continuació va estudiar un any a Alemanya, a les universitats de Bonn i Berlín, per després retornar a Harvard on va rebre el seu doctorat en filologia comparada l'any 1930.

Va esdevenir instructor d'Alemany fins a l'any 1936, professor auxiliar fins al 1939 i professor universitari fins a la seva mort l'any 1950.

Durant la seva vida va fer cinc publicacions:

- "Estudis selectius i el Principi de la freqüència relativa en l'Idioma" (1932)
- "La Psico-Biologia de les Llengües" (Hughton-Mifflin 1935, MIT Press 1965)
- "Unitat Nacional y desunió: la nació com un Organisme Bio-Social" (1941)
- "La hipòtesi del (P1*P2)/D: En el moviment interurbà de les persones" (1946)
- "El comportament humà i el principi del menor esforç" (1949)

Escrits sobre la llei de Zipf (G. K. Z.)

“Estudis selectius i el Principi de la freqüència relativa en l’Idioma” (1932)

G.K. Zipf ensenya la seva recerca dintre la freqüència relativa en diferents àmbits: fonologia xinesa, l’abreviatura, el canvi semàntic i obres com *Plaute 'Aulularia, Mostellaria, Pseudolus i Trinummus*. L’autor transcriu 20.000 síl·labes del dialecte de Pequín i n’estudia la freqüència d’aquestes paraules i les seves síl·labes.

“La Psico-Biologia de les Llengües” (1935/1965)

G.K. Zipf mostra diferents patrons en la llengua humana. Com per exemple, que la longitud d’una paraula està pròximament relacionat amb la freqüència del seu ús; com més gran és la freqüència, més curta és la paraula. És més, es pot demostrar que com més complexa és la paraula, la freqüència és menor. Tota l’evidència de l’autor ens porta a la conclusió que existeix una condició de l’equilibri entre la forma i els hàbits o patrons de la parla en qualsevol idioma.

La llei empírica

Una llei empírica és aquella que està basada en la pràctica, l’experiència i l’observació.
El nom empíric deriva del pensament filosòfic anomenat Empirisme. Aquest va ser fundat per John Locke. L’empirisme va sorgir de les illes Britàniques al segle XVIII. Però, d’altra banda el nom prové del grec *empíria* que significa experiència.

Les lleis empíriques ens informen de fenòmens que existeixen i ens expliquen les seves característiques. A partir de l’observació de fenòmens que segueixen una llei empírica un mateix guanya, el que es coneix com, coneixement empíric.

La Llei de Zipf

La llei de Zipf és una llei empírica formulada utilitzant l'estadística matemàtica que fa referència al fet que dades de molts camps estudiats en les ciències físiques i socials es poden aproximar a una distribució Zipfiana.

Donat un conjunt de dades, ordenades per valor ($F_1 \geq F_2 \geq F_3 \ldots$), r és el rang de F_r. Es podria pensar que F_r és la mida del valor de r dintre el conjunt ordenat.

La llei de Zipf diu que la freqüència(F) d'una paraula és inversament proporcional al rang(r).

$$F \cdot r = k$$

Equació 1

Multiplicant la freqüència pel rang ens dóna una k, una constant. Per tant, si el rang disminueix, la freqüència augmenta i si el rang augmenta, la freqüència disminueix.
Aquesta equació la podem girar quan el nostre interès es basa en la freqüència i no pas en la seva relació.

Aquesta fórmula ens permet buscar la freqüència (F) si se sap la constant (k) i el rang (r).

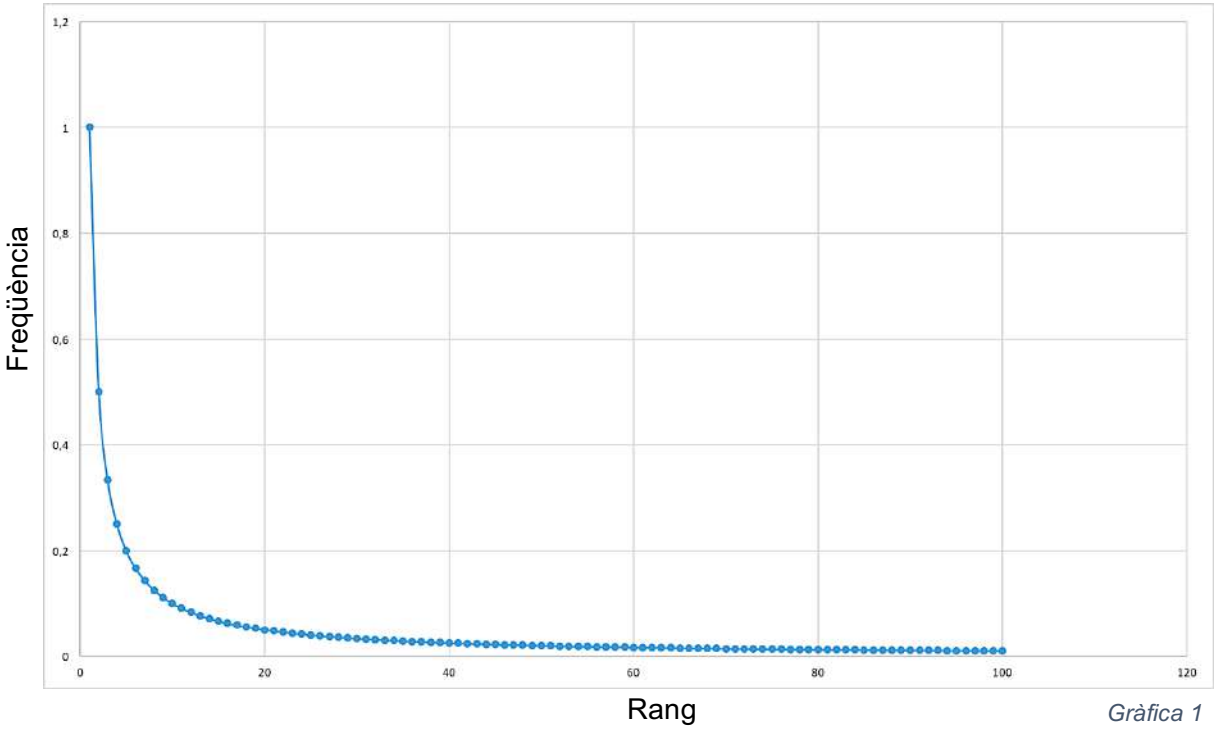
$$F = \frac{k}{r}$$

Equació 2

R	F
1	1
2	1/2
3	1/3

-Donat que k=1 *Taula 1*

Segons la llei, la paraula més freqüent apareixerà el doble de vegades que la segona més freqüent, tres vegades més que la tercera més freqüent, etc.



Gràfica 1

La relació entre rang i freqüència mostra una hipèrbola rectangular.

Trenta anys després de la popularització de la llei de Zipf, el Sr. Booth i Federowicz van ajustar la llei per millorar la relació rang-freqüència quan s'aplica a textos.

Van trobar que la llei s'ajustava més si es modificava el rang d'una mateixa manera en tot un text. La modificació que van afegir va ser elevar el rang a una constant (*b*, aquesta constant és propera a 1).

$$F = \frac{k}{r^b}$$

Equació 3

A mesura que augmentem el conjunt de dades, les paraules de rang menor augmenten en freqüència i s'observen més paraules. Aquests fenòmens causen que l'eix y contingui una gamma més gran de valors, fins a tal punt on no es pot apreciar les diferències entre dades properes.

Per a resoldre aquest problema ajustem la gràfica i posem en logaritmes els eixos. Per a fer aquest pas, recordem la propietat dels logaritmes i apliquem logaritmes a cada costat de l'equació 4.

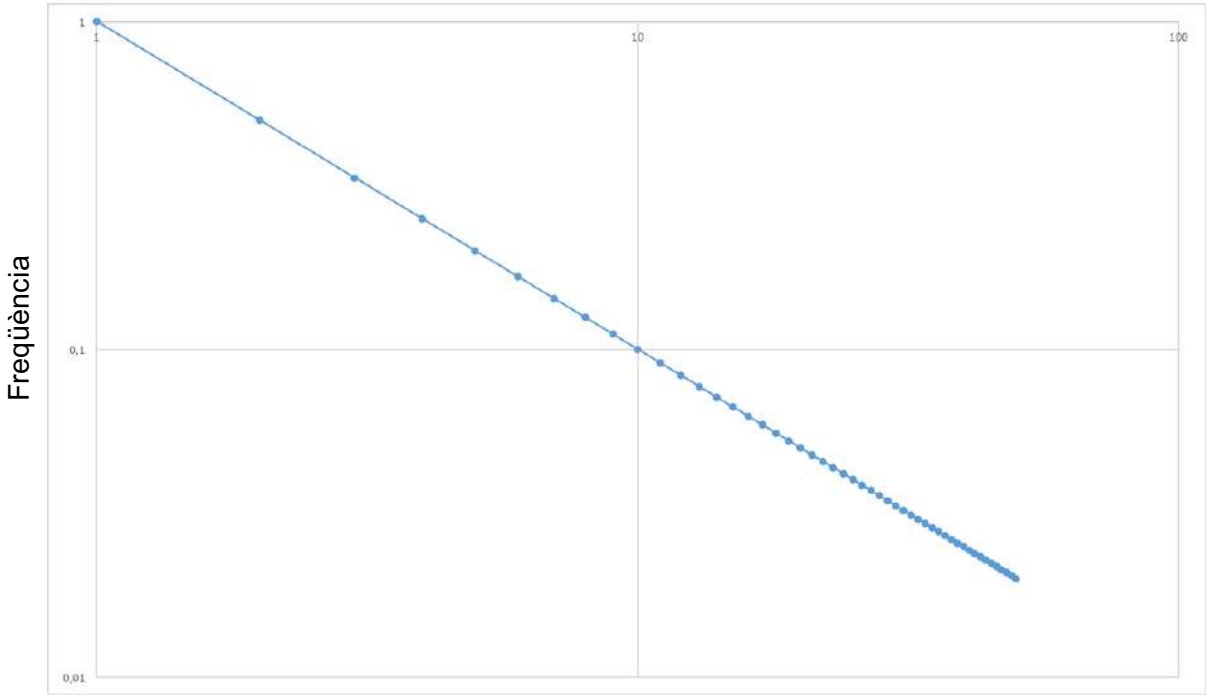
$$x = y \rightarrow lnx = lny$$

Equació 4

$$F = \frac{k}{r^b} \rightarrow lnF = lnk - b \cdot lnr$$

Equació 5

És una línia recta del tipus *y=mx+n*
m= -b n=lnk

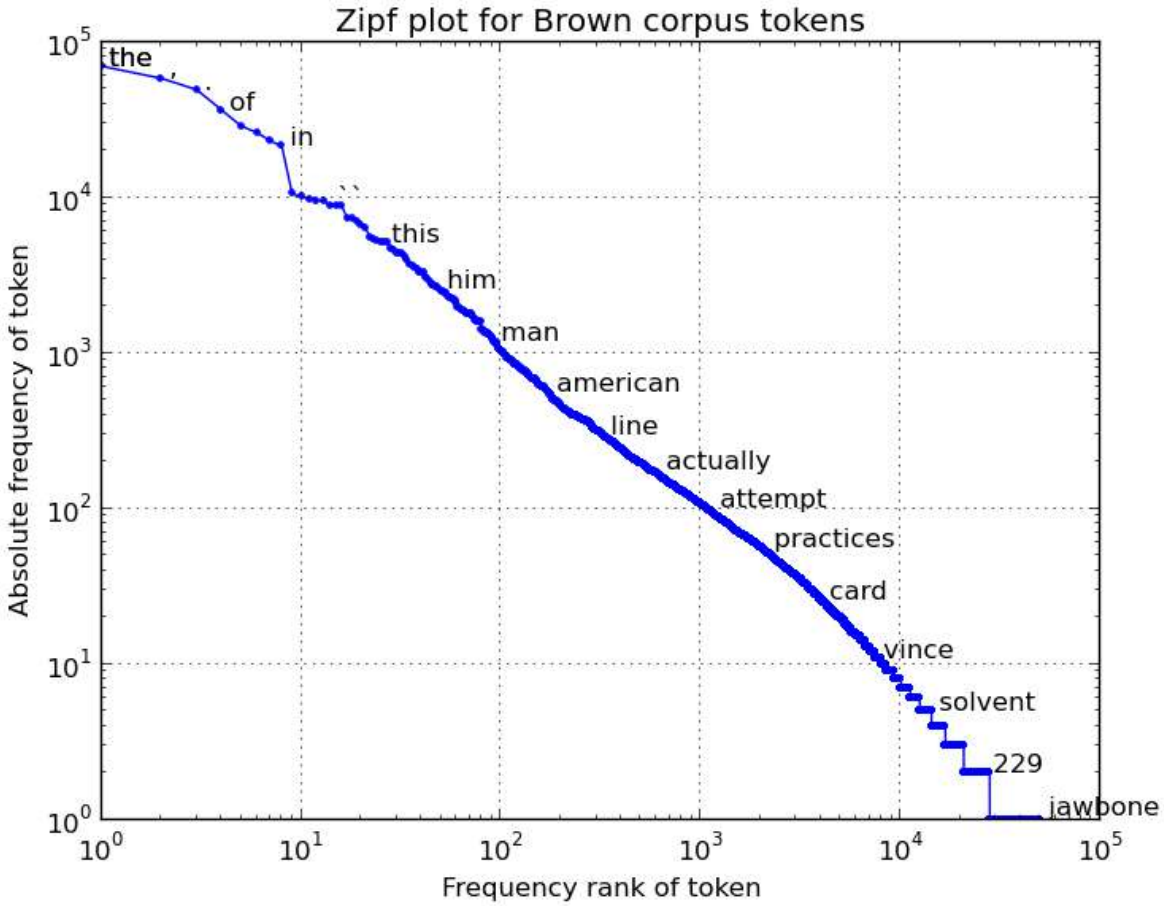


Gràfica 2

La relació entre rang i freqüència amb eixos logarítmics mostra una línia recta

En aquesta gràfica es mostra la relació rang-freqüència representada en eixos logarítmics. Quan passem a l'escala logarítmica deixem de veure una hipèrbola rectangular i es mostra una línia recta de pendent negativa.

En el Brown Corpus, conjunt de cinc-cents textos americans compilat l'any 1961, la paraula *"the"* és la paraula més freqüent, apareix 69.971 vegades i representa el 7% de totes les paraules. La segona paraula és *"of"*, la qual apareix 36.411 vegades i representa el 3,5% del corpus. Seguidament tenim la paraula *"and"* amb 28,852 aparicions. Només es necessiten 135 paraules per a formar 50% del tot el Corpus



(https://finnaarupnielsen.files.wordpress.com/2013/10/brownzipf.png) *Fig. 5*

En la gràfica superior (Fig.5) es veu clarament com es forma una línia recta a partir de les dades del corpus. També s'hi mostren exemples de quin tipus de paraules es mostren en diferents freqüències.

La Recerca

La recerca consta de quatre parts:

1. La selecció de textos
2. El comptador de freqüència
3. Les taules i gràfiques
4. Els resultats

La selecció de textos

En tota recerca el grup estudiat influeix gran part en el seguiment i el resultat de la recerca, per tant, de bon principi volíem ser curosos a l'hora de triar els nostres textos. Però ràpidament ens vàrem adonar que la llei es refereix al funcionament d'una llengua per tant el registre o el context del text hauria de ser irrellevant a la recerca.
Hem decidit emprendre la nostra recerca amb sis textos: dos llibres, un del 1490 i l'altre del 1974, un article de diari del 2011, dues cançons de principis dels anys noranta i l'Himne dels Segadors escrit l'any 1897.

Llibres

- "El mecanoscrit del segon origen", escrit per Manuel de Pedrolo i Molina
  Aquesta obra va tenir un gran èxit de públic, sobretot en el sector juvenil, resulta ser un dels llibres més venuts de la literatura catalana.

- "Els Episodis Amorosos" de "Tirant Lo Blanc", escrit per Joanot Martorell
  És considerada un dels màxims exponents de la novel·la cavalleresca en llengua catalana i del segle d'or valencià.

Articles

- La Vanguardia, 3 maig del 2011
  La Vanguardia porta des del 1881 publicant en castellà, però el 3 de maig del 2011 va publicar la primera edició en català.

Cançons

- *Boig per tu*, SAU
  La lletra d'aquesta cançó expressa una declaració d'amor cap a la Lluna, però també pot ser dedicada a una persona.

- *L'Empordà*, Sopa de Cabra
  El disc que inclou aquesta cançó ha venut 40.000 còpies, aquesta cançó s'ha convertit en l'emblema del grup .

Himnes

- *Els Segadors*
  És l'himne nacional oficial de Catalunya, el qual fa una crida per defensar la llibertat de la terra.

El comptador de freqüències

Per a continuar amb la nostra recerca necessitàvem un mètode per a comptar la freqüència que fos útil i eficient.

Els primers dies vam començar comptant la freqüència a mà. Amb una llibreta i un bolígraf apuntàvem les paraules per columnes i marcàvem amb un asterisc les paraules repetides. No vam tardar gaire a adonar-nos que no era un mètode gaire eficient. En cerca d'un mètode per comptar la freqüència vam decidir entrar al món d'Internet per a trobar una solució. Internet és un lloc magnífic ple de tot el que un pugui necessitar, des de les més noves invencions fins a exemplars dels primers textos asiàtics. Però tots aquests avantatges també comporten algun desavantatge, com per exemple, programaris amb imperfeccions que afecten el seu funcionament.

Ràpidament vàrem trobar un comptador de freqüències. Però, aquest, estava programat en anglès i no entenia els accents, aquest conflicte causava que les paraules que portaven accent eren separades i les síl·labes eren perdudes entre la multitud de paraules.

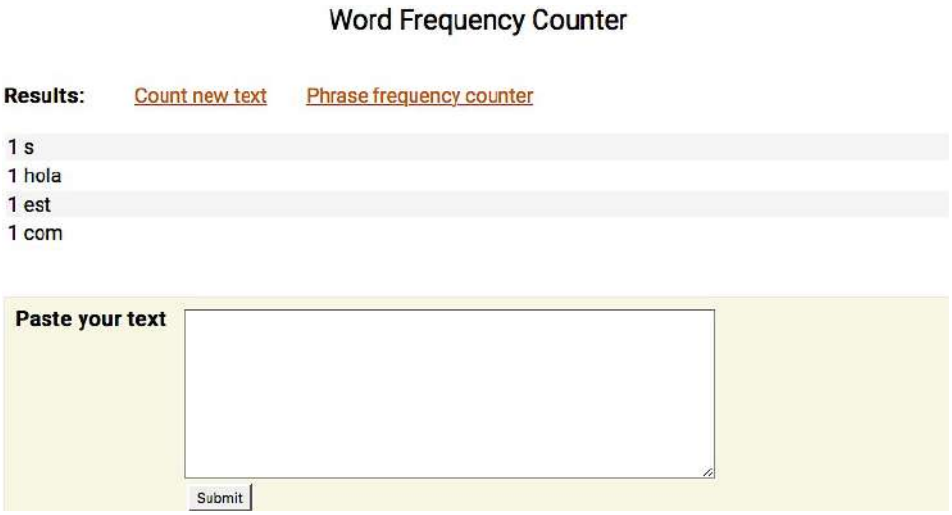


Fig. 6 *(http://www.writewords.org.uk/word_count.asp)*

A continuació d'aquest incident vam decidir buscar-ne un altre, i, efectivament, en vam trobar un altre en qüestió de minuts. N'havíem trobat un segon, i aquest, era capaç d'entendre els accents sense cap mena de problema. Però, no era perfecte, era una versió beta, i només et mostrava les dues-centes paraules més freqüents. Útil en textos curts, però no el podíem utilitzar en textos més llargs, perdíem massa informació.

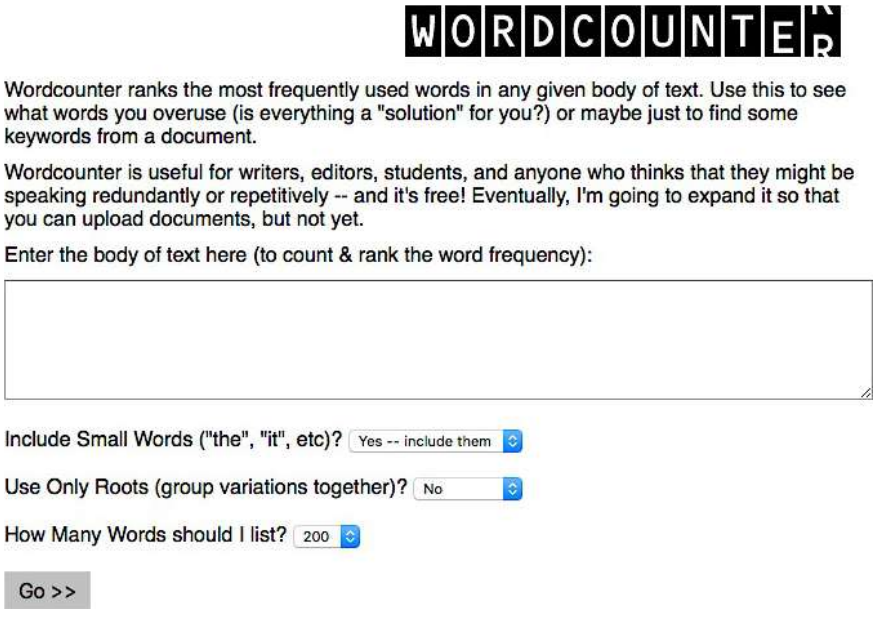


Fig. 7 *(http://www.wordcounter.com)*

Decididament vam continuar la nostra cerca per a un comptador útil i eficient, i finalment el vam trobar.

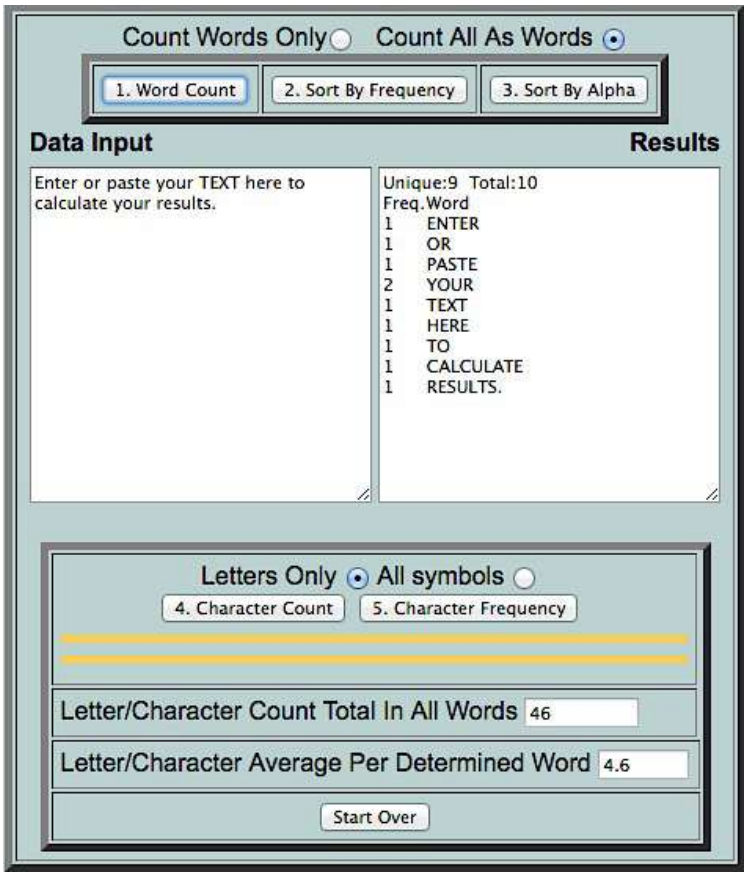
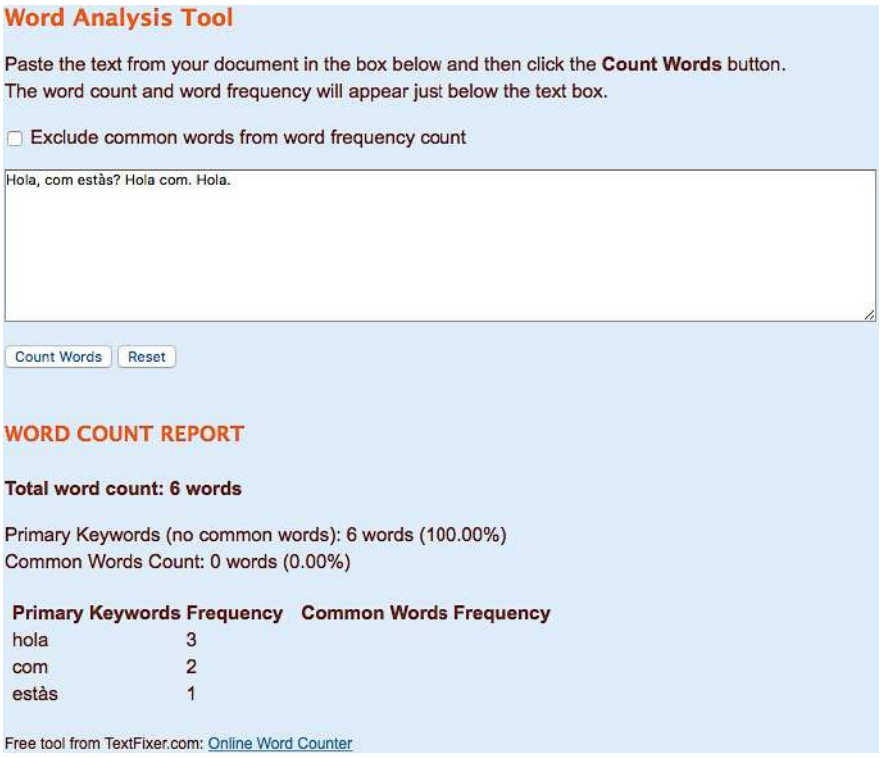


Fig. 8 *(http://www.csgnetwork.com/documentanalystcalc.html)*

Aquest programari ens permetia comptar les paraules, ordenar-les per aparició, freqüència i alfabèticament. El programari et mostrava el nombre total de paraules i el nombre de diferents paraules. A més, el programari també et podria comptar la freqüència dels caràcters i ordenar-los tant per ordre d'aparició com per freqüència.
Un altre avantatge d'aquest programari és que està formatejat de tal manera que pots traslladar els resultats en un full de càlcul sense alterar-los.

Per descomptat, no ens podem deixar l'avantatge que ha fet possible aquest estudi. És gratuït i compatible amb els sistemes operatius de Mac, Windows i Linux.

A part d'aquest últim comptador també en vam trobar dos més que ens podrien haver servit en cas que aquest fallés o fos eliminat del lloc web.



Word Analysis Tool

Paste the text from your document in the box below and then click the **Count Words** button.
The word count and word frequency will appear just below the text box.

☐ Exclude common words from word frequency count

Hola, com estàs? Hola com. Hola.

Count Words | Reset

WORD COUNT REPORT

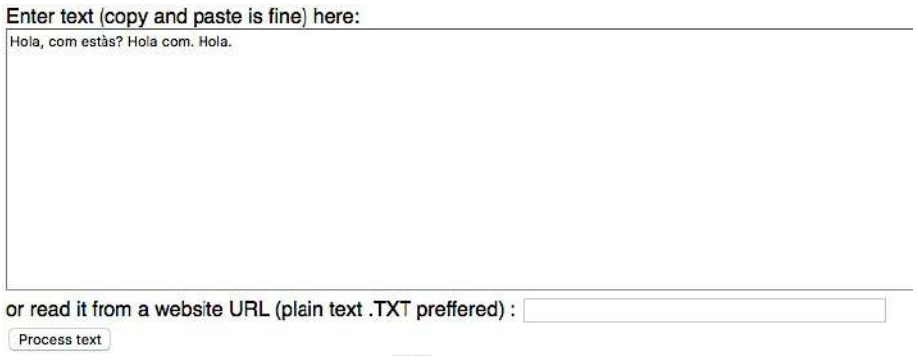
Total word count: 6 words

Primary Keywords (no common words): 6 words (100.00%)
Common Words Count: 0 words (0.00%)

Primary Keywords	Frequency	Common Words	Frequency
hola	3		
com	2		
estàs	1		

Free tool from TextFixer.com: Online Word Counter

Fig. 9 *Font: <http://www.textfixer.com/tools/online-word-counter.php>*



Enter text (copy and paste is fine) here:

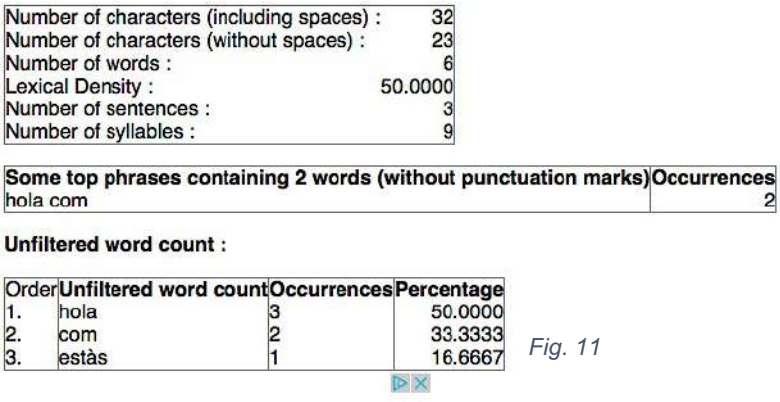
Hola, com estàs? Hola com. Hola.

or read it from a website URL (plain text .TXT preffered) :

Process text

Fig. 10

Font: <http://www.online-utility.org/text/analyzer.jsp>



Number of characters (including spaces) :	32
Number of characters (without spaces) :	23
Number of words :	6
Lexical Density :	50.0000
Number of sentences :	3
Number of syllables :	9

Some top phrases containing 2 words (without punctuation marks)	Occurrences
hola com	2

Unfiltered word count :

Order	Unfiltered word count	Occurrences	Percentage
1.	hola	3	50.0000
2.	com	2	33.3333
3.	estàs	1	16.6667

Fig. 11

Font: <http://www.online-utility.org/text/analyzer.jsp>

Un cop teníem els resultats traslladats en un full de càlcul vam ensopegar amb alguns problemes. En primer lloc, trobàvem que, a diferència de l'anglès, tots els substantius tenen gènere i per tant tots els articles i pronoms relacionats als substantius poden variar entre masculí i femení. En altres paraules, l'article anglès "*the"* és utilitzat tant per a "taula" com "llapis", "*the table and the pencil"*, mentre que en català utilitzem el i la, "*la taula i el llapis".* Vàrem decidir agrupar les paraules que tenien variacions entre masculí i femení.

Seguidament també ens qüestionem com manegem la diferència entre paraules singulars i plurals. Si acceptem *el* i *la* dins un mateix conjunt també tenim el mateix dubte amb el nombre de la paraula. Traduint de l'anglès: "*the table and the tables", "la taula i les taules".*

El funcionament de la llengua catalana és utilitzar articles abans de mencionar substantius, aquest article varia segons gènere i nombre, per tant s'ha de comptar la freqüència dels articles, i totes les altres paraules, en grups: el/la/els/les (639, Episodis Amorosos) i emperador/a (42, Episodis Amorosos).

Però els dubtes no es van acabar aquí. En anglès: "*I sing, you sing...",* però, *"jo canto, tu cantes i ell canta..."*. Dintre la mateixa conjugació verbal surten sis formes diferents, les quals no són reconegudes com a una mateixa pel programari. Així que un cop traslladàvem els resultats corregíem totes les formes verbals reduïdes a només el verb, deixant enrere la persona i el temps verbal del verb.

Les taules

Ja hem finalitzat la primera tapa, ara tenim totes les dades traslladades en un full de càlcul.

Com és mencionat en l'apartat anterior és necessari modificar la taula perquè les diferències de gènere, nombre, conjugació i pronominalització podrien marcar la diferència a l'hora d'aplicar la llei.

	Freq.	Word
1	2013	DE
2	1608	I
3	1603	QUE
4	1279	LA
5	1161	A
6	888	EN
7	786	EL
8	723	NO
9	663	VA
10	626	ELS

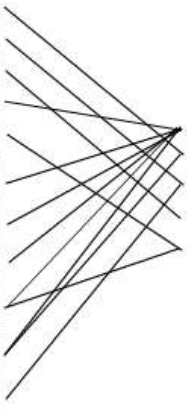
Taula 2

Modifiquem la taula seguint quatre directrius:

1. Les paraules amb variacions, de nombre o gènere, s'ajunten. (*el+la+els+les*)
2. Totes les conjugacions d'un verb es reuneixen sota l'infinitiu. (*volen+volia+vull→voler*)
3. Els pronoms es despronominalitzen i s'agrupen sota la paraula que substituïen (*-los → els, -ne → en*)
4. Les paraules compostes, com per exemple: *del, pel, al, etc,* seran separades i comptades com a paraules diferents. (*del→de+el, pel→per+el, al→a+el).*

Un cop hem modificat la taula seguint les directrius, ordenem de nou les paraules per ordre de freqüència.

Rank	Frequència	Paraula
1	2013	DE
2	1608	I
3	1603	QUE
4	1279	LA
5	1161	A
7	786	EL
10	626	ELS
14	527	LES
18	339	AL
22	287	DEL
26	254	I,



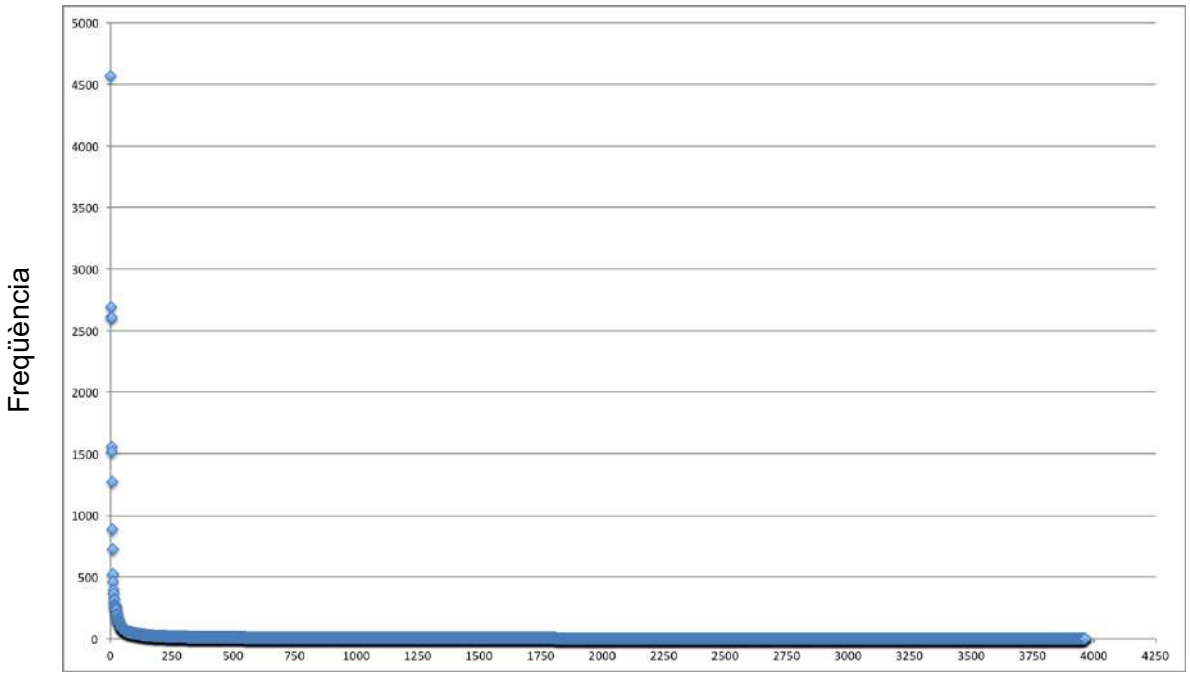
Rank	Frequència	Paraula
1	4568	LA
2	2692	DE
3	2608	I
4	2603	QUE
5	1555	A

Taula 3

(Totes les modificacions de les taules es troben a l'Annex 1)

Les gràfiques

Quan un individu mostra un conjunt de dades sobre una gràfica, hi ha moltes variables per modificar el significat de les gràfiques afegint subjectivitat. Per tant, hem de trobar una manera de mostrar el nostre conjunt de dades perquè sigui el més favorable possible. Una vegada disposem sobre un full de càlcul les dades obtingudes i estan ordenades mitjançant les directrius anteriors és hora de començar a representar-les. La gràfica més adient als nostres objectius és la representació mitjançant els eixos X i Y. Un cop creat el gràfic obtenim un resultat similar a aquest:



Gràfica 3

Un cop hem representat les nostres dades veiem que la hipèrbola rectangular mencionada a la gràfica 1 té relació amb la llei de Zipf. Precisament, utilitzant la línia que formen les dades podrem trobar un número per establir l'exactitud de la llei de Zipf. Aquest procés el durem a terme a partir de les anomenades línies de tendència.

Una línia de tendència és una corba que intenta encaixar amb el conjunt de dades. Les línies de tendència poden ser configurades en sis modes diferents: linear, logarítmica, polinòmica, potencial, exponencial i la mitjana mòbil. Cadascun d'aquests modes ens ajuda a determinar l'aparença d'un conjunt de dades a una equació.

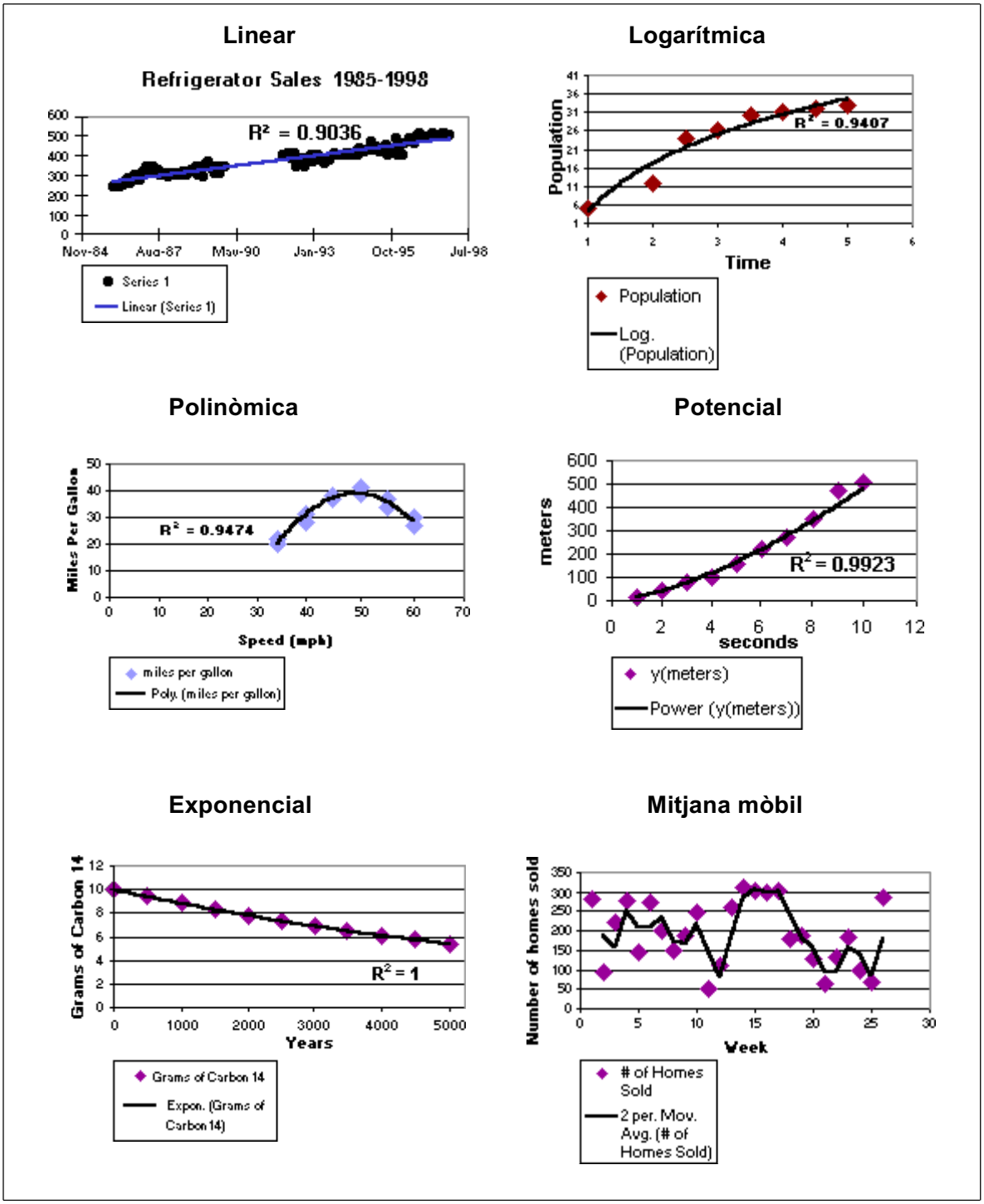
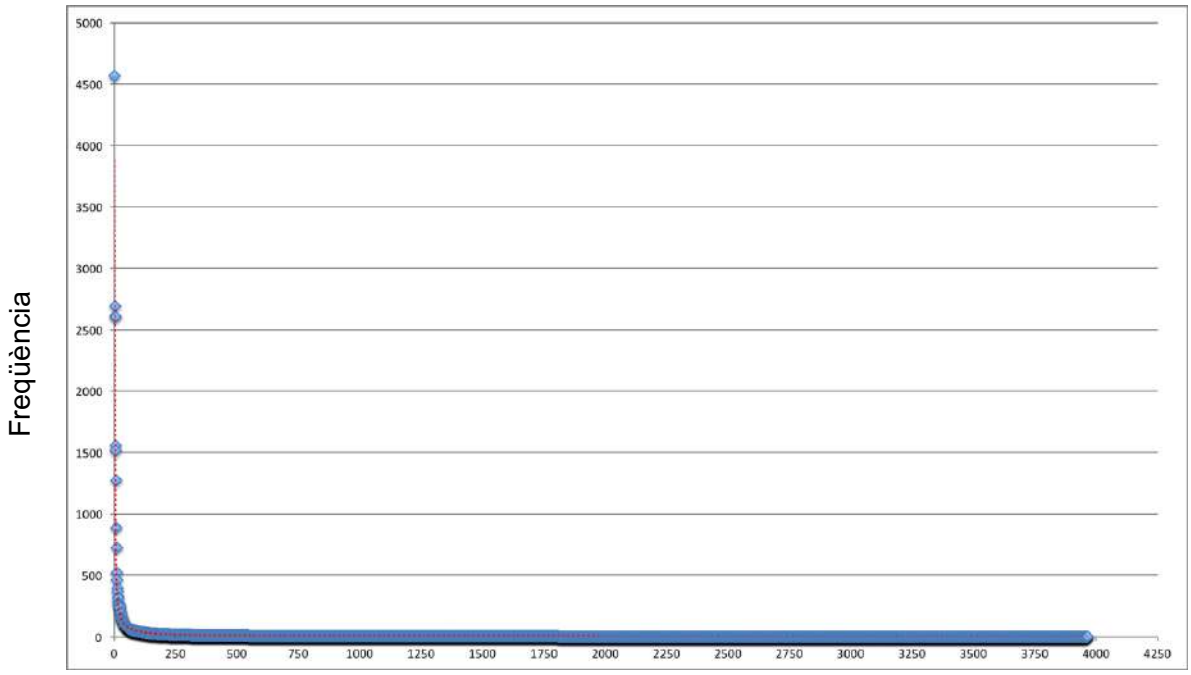


Fig. 12 *Font:https://support.office.com/en-us/article/Choosing-the-best-trendline-for-your-data-1bb3c9e7-0280-45b5-9ab0-d0c93161daa8*

En el nostre cas, la línia de tendència que necessitem és la potencial. La línia de tendència té la funció de comparar el conjunt de dades amb l'equació més similar a aquella. Aquesta comparació es mesura mitjançant el coeficient de determinació (R^2).
El coeficient de determinació és un nombre que oscil·la entre -1 i 1 en relació a la correspondència que té l'equació generada per l'ordinador amb el conjunt de dades.



Gràfica 4

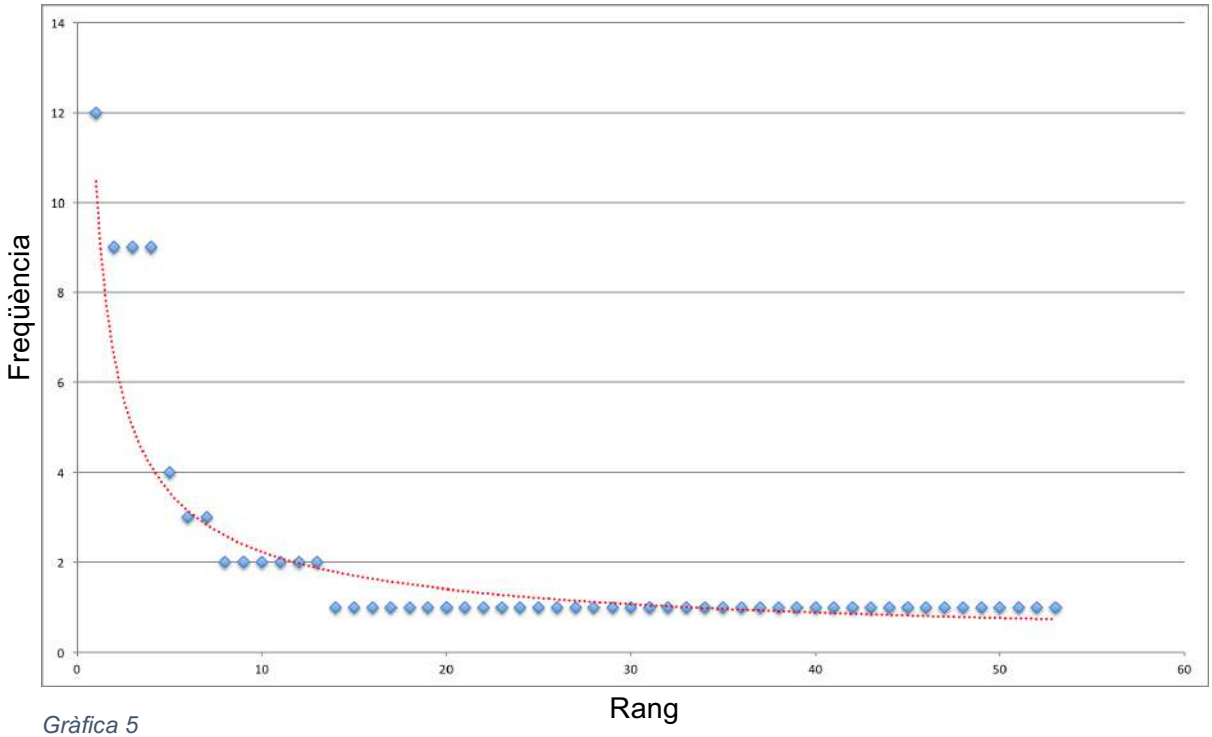
En la Gràfica 4 veiem l'equació $F = 3875 \cdot r^{-0,985}$ representada en negre, aquesta equació és capaç de preveure el conjunt de dades amb un coeficient de correlació (R^2) de 0,993. El coeficient de correlació serà el que ens servirà de base per confirmar o no la llei de Zipf. Si R^2 s'apropa a 1, vol dir que la llei es confirma.

Els resultats

Després d'aquest llarg procés de seleccionar els textos, comptar les paraules, ordenar i reordenar les taules (veure annex), representar les gràfiques i mostrar l'equació juntament amb el coeficient de correlació és hora de mostrar els resultats.

Els Segadors

Els segadors no és un text llarg, conté 101 paraules, de les quals 53 en són diferents. Només les paraules del rang 1-10 són necessàries per formar 51% del text.



Gràfica 5

Visualment la línia de tendència i les dades del text no s'ajusten molt, això ve donat perquè hi ha tres paraules amb F=9, cinc amb F=2 i quaranta amb F=1. Com veiem en l'equació de la línia de tendència, b (el modificador del rang) no és gaire proper a 1. Però, no obstant això, l'himne dels segadors té un coeficient de relació elevat.

$$F = 10,465 \cdot r^{-0,671}$$

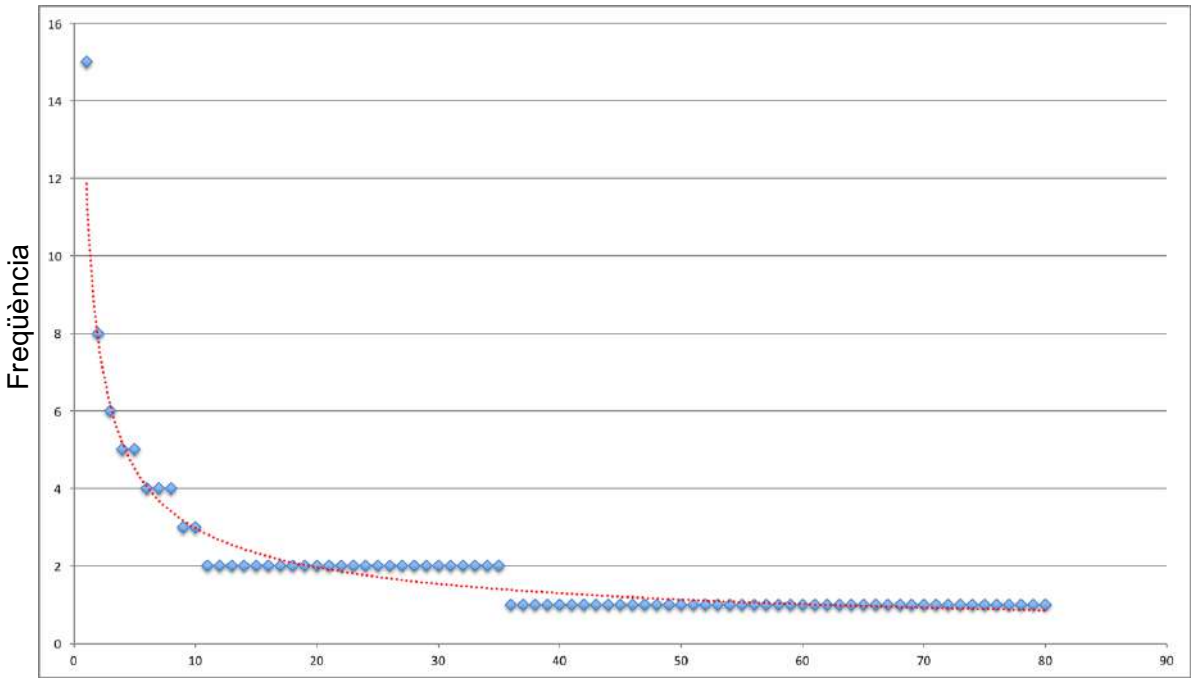
Equació 7

$$R^2 = 0,82905$$

Equació 7

L'Empordà – Sopa de Cabra

La cançó conté 80 paraules diferents, les quals formen un text de 152 paraules. D'aquestes 152 paraules, 50% són escrites utilitzades els rangs 1 a 20.



Gràfica 6

Visualment les gràfiques dels textos curts, com cançons, himnes i poesia, no és molt clar la relació entre les dades i la línia de tendència. Comparant aquest text amb l'himne anterior veiem que aquest té més variació de freqüències. Aquest també té el coeficient de correlació més elevat, de 0,83 a 0,91.

$$F = 11,863 \cdot r^{-0,601}$$

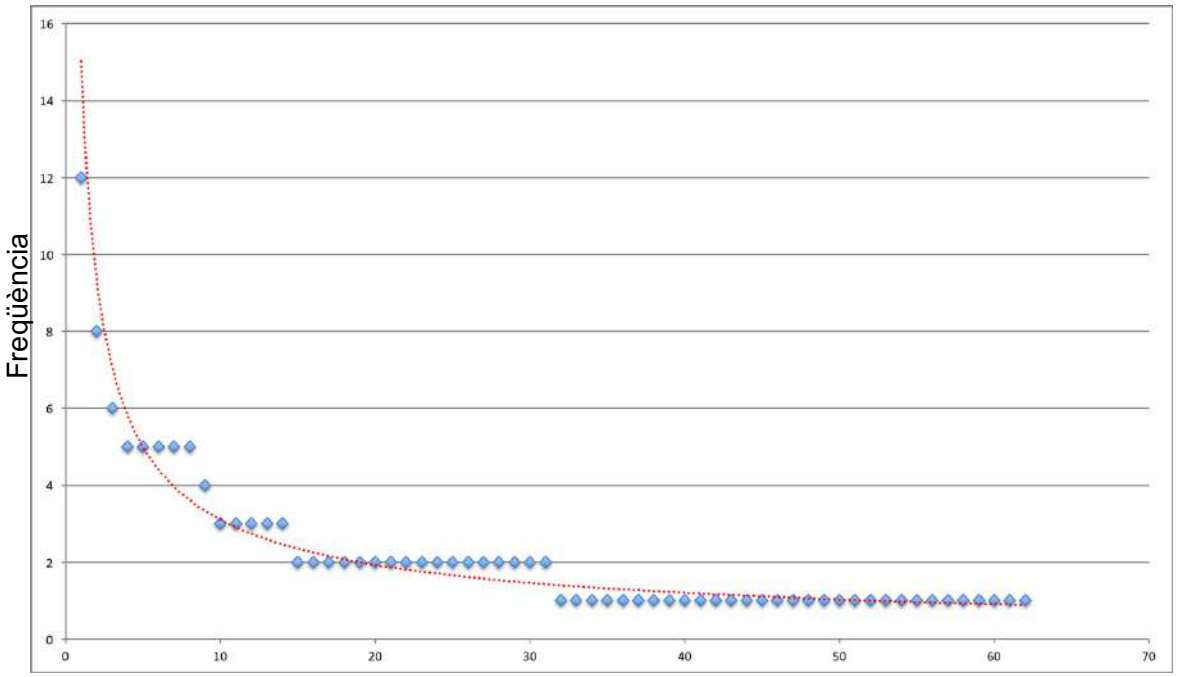
Equació 9

$$R^2 = 0,91408$$

Equació 9

Boig Per Tu – SAU

Seguint l'estudi amb cançons no es veu extremadament clar, però tot i així es segueix mostrant. En aquest text de 135 paraules només són necessàries les paraules del rang 1-13 per a formar el 50% de la cançó.



Gràfica 7

"Boig per Tu" té un coeficient de correlació més elevat que els dos textos anteriors. Per la quantitat de paraules que conté, és extraordinari que la llei es confirmi.

$$F = 15,024 \cdot r^{-0,685}$$

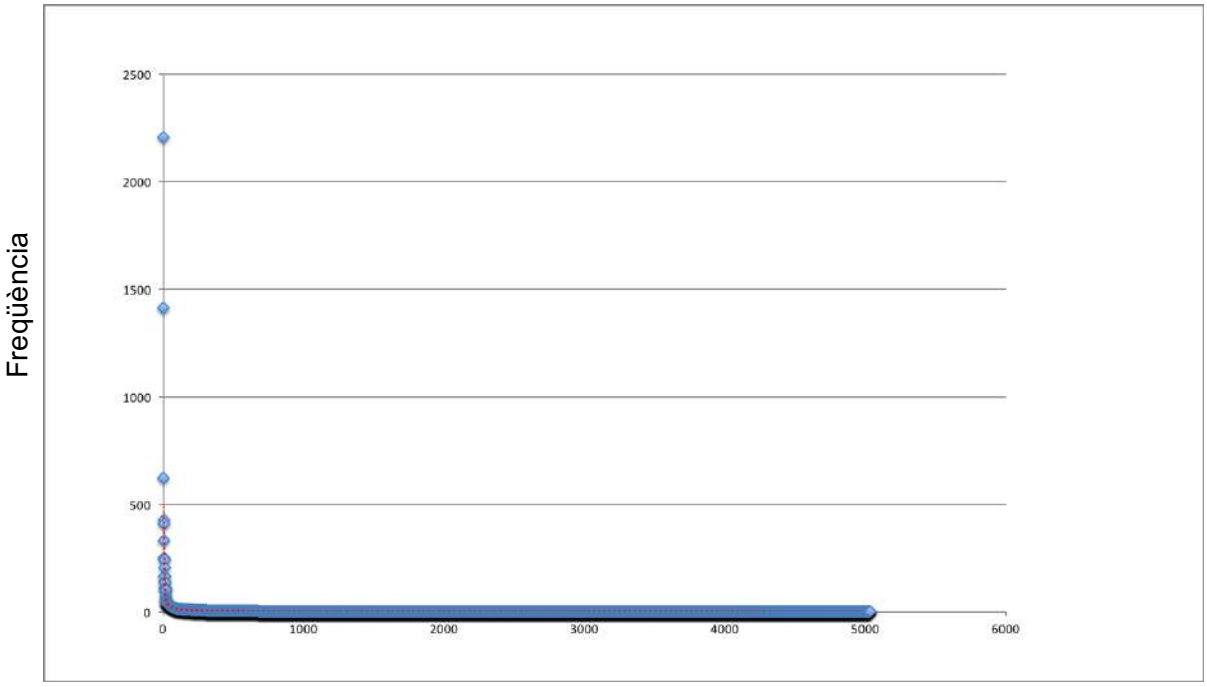
Equació 11

$$R^2 = 0,929861$$

Equació 11

La Vanguardia 3/05/2011

En l'edició de la Vanguardia comencem a treballar amb conjunts de dades més grans. En la publicació (3/03/2011) estem treballant amb 19.120 paraules, cent vegades més que en els textos anteriors.



Gràfica 8

$$F = 488,6 \cdot r^{-0,762}$$

Equació 12

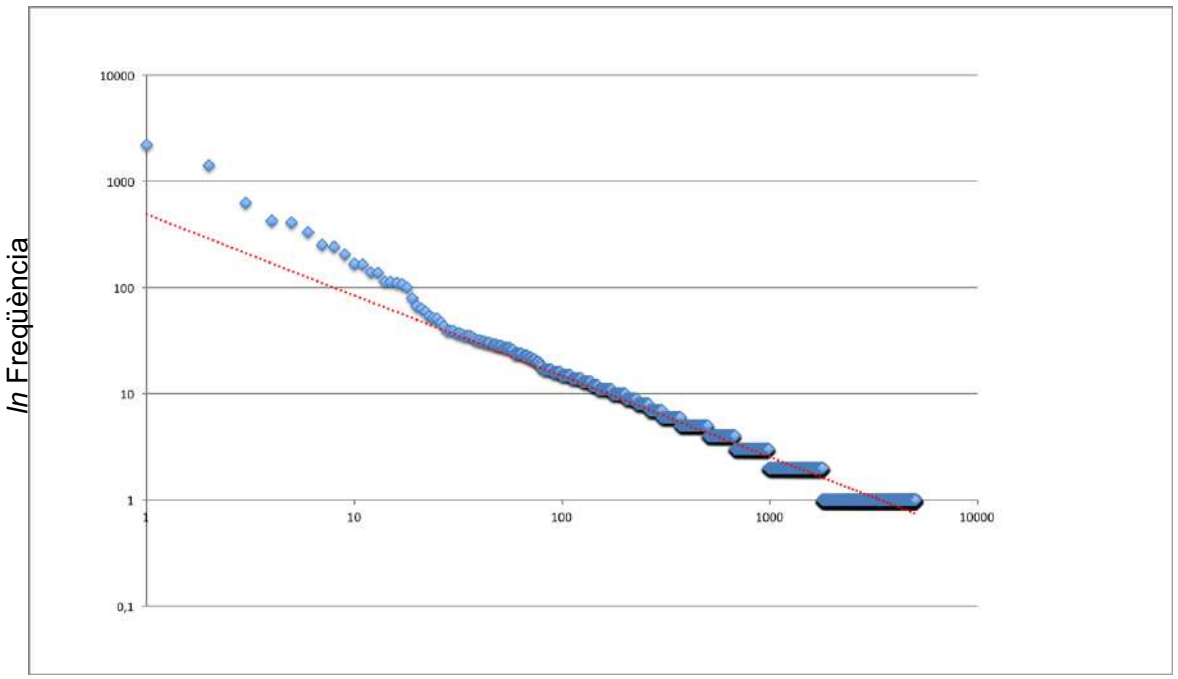
La quantitat de text és tan gran que significa que hi ha moltíssimes més paraules amb freqüències menors (3.233 paraules amb F=1, 811 paraules amb F=2 i 208 amb F=3). Aquestes tres freqüències donen lloc a 84% de les paraules.

A la gràfica no s'aprecia la diferència de valors ni la hipèrbola rectangular. És en aquest moment on hem de canviar la configuració de la gràfica i posar els eixos en escala logarítmica.

Cap a lineal: utilitzant la propietat dels logaritmes

En el programa d'edició de fulls de càlculs que nosaltres utilitzem (Excel) és molt senzill fer aquest canvi.

1. Es fa clic amb el botó dret sobre l'eix que vols editar.
2. Dintre el menú que apareix se selecciona *Format de l'eix...*
3. Surt una finestra amb diferents opcions, es busca la casella que digui *Escala Logarítmica*, i es prem.



Gràfica 9

A la gràfica 9 veiem que deixen de formar una hipèrbola rectangular i formen una línia recta. L'equació de la línia de tendència és la mateixa que utilitzant els eixos linears. Si es vol buscar l'equació lineal de la línia de tendència, només s'ha de passar l'equació de la freqüència a logaritmes:

$$F = \frac{k}{r^b} \rightarrow lnF = lnk - b \cdot lnr$$

Equació 16

Equació 16

$$F = 488,6 \cdot r^{-0,762}$$

$$lny = 6,19 - 0,762 \cdot lnr$$

Equació 16

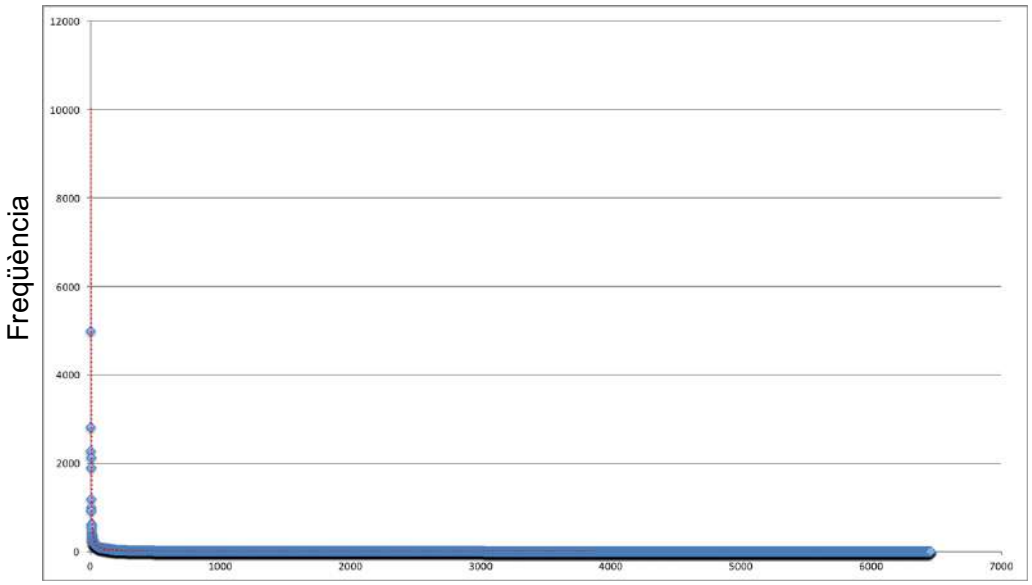
El coeficient de correlació calcula la diferència entre les dades i la línia de tendència, per tant el resultat no és afectat pel canvi d'escala lineal a logarítmica ni el canvi de la línia de tendència.

$$R^2 = 0,937$$

Equació 16

Els Episodis Amorosos de Tirant Lo Blanc – Joanot Martorell

En el moment que comencem a analitzar llibres, veiem que la llei es mostra fàcilment. Els Episodis Amorosos de Tirant Lo Blanc contenen 56.439 paraules, formades per 6.462 paraules diferents.

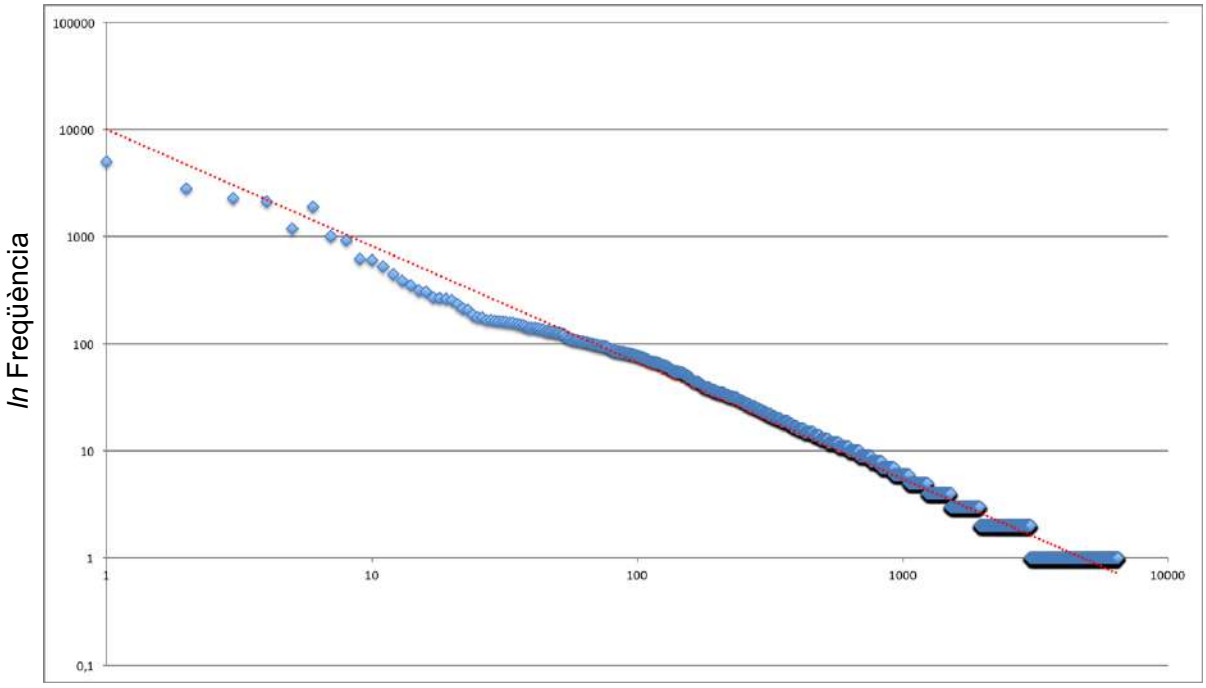


Gràfica 10

Rang

$$F = 10.015 \cdot r^{-1,088}$$

Equació 17



ln Rang

Gràfica 11

$$lnF = 9,21 - 1,088 \cdot lnr$$

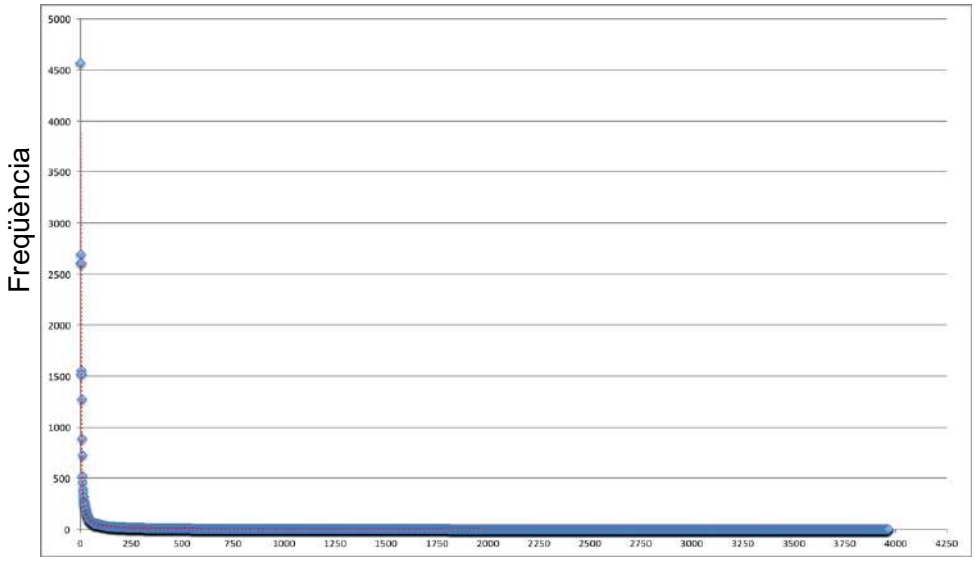
Equació 18

$$R^2 = 0,969$$

Equació 19

El Mecanoscrit del Segon Origen – Manuel de Pedrolo i Molina

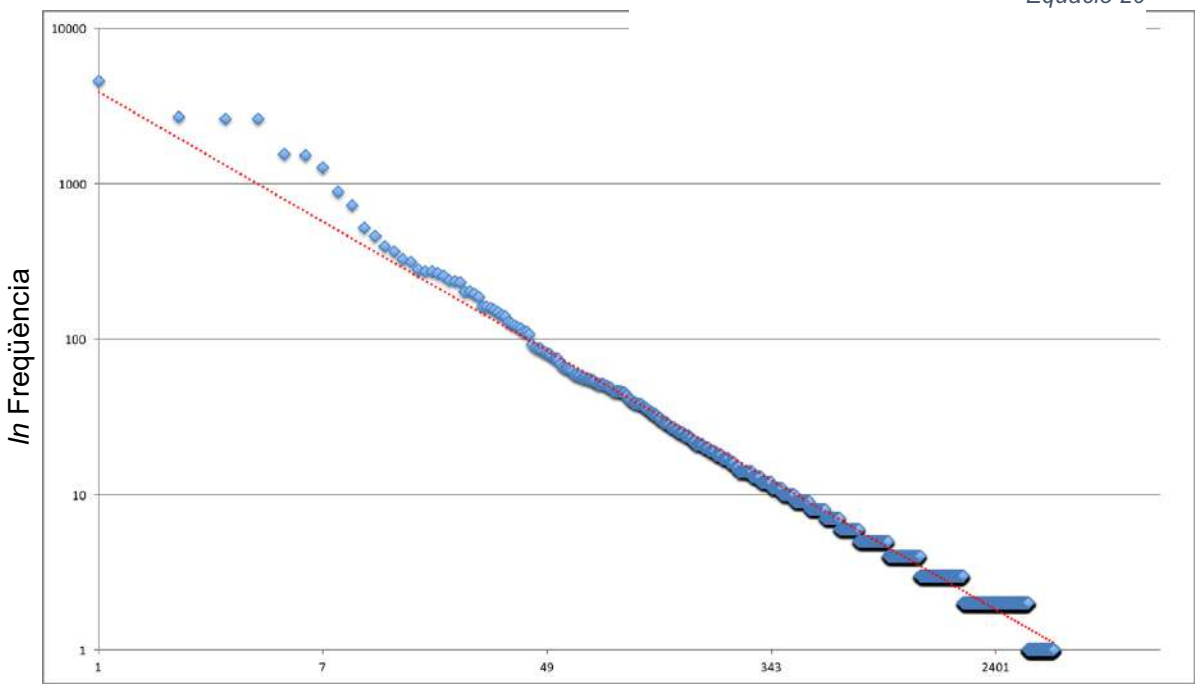
El mecanoscrit del Segon Origen va ser el primer llibre que vàrem començar a estudiar, també és el que aconseguírem més fàcilment en versió PDF per a fer la compta de paraules. Aquest text és el que hem mostrat durant l'explicació del procés, ja que és un conjunt de dades molt fiables les quals corresponen idealment amb la llei de Zipf.



Gràfica 12

$$F = 3875 \cdot r^{-0,985}$$

Equació 20



Gràfica 13

$$lnF = 8,26 - 0,985 \cdot lnr$$

Equació 21

Equació 18

$$R^2 = 0,99297$$

Corpus

En la majoria d'idiomes s'hi associa un institut per regular les normes gramaticals i altres matèries semblants. Una d'aquestes matèries és crear un recull de tota la literatura escrita en l'idioma que representen, aquest recull s'anomena un Corpus. En la llengua catalana existeix un corpus proveït per l'Institut d'Estudis Catalans. Quan vam trobar el corpus català no enteníem molt bé el funcionament. Jo buscava un fitxer que contingués tots els escrits mentre que el que trobàvem era la pàgina de consultes, http://ctilc.iec.cat

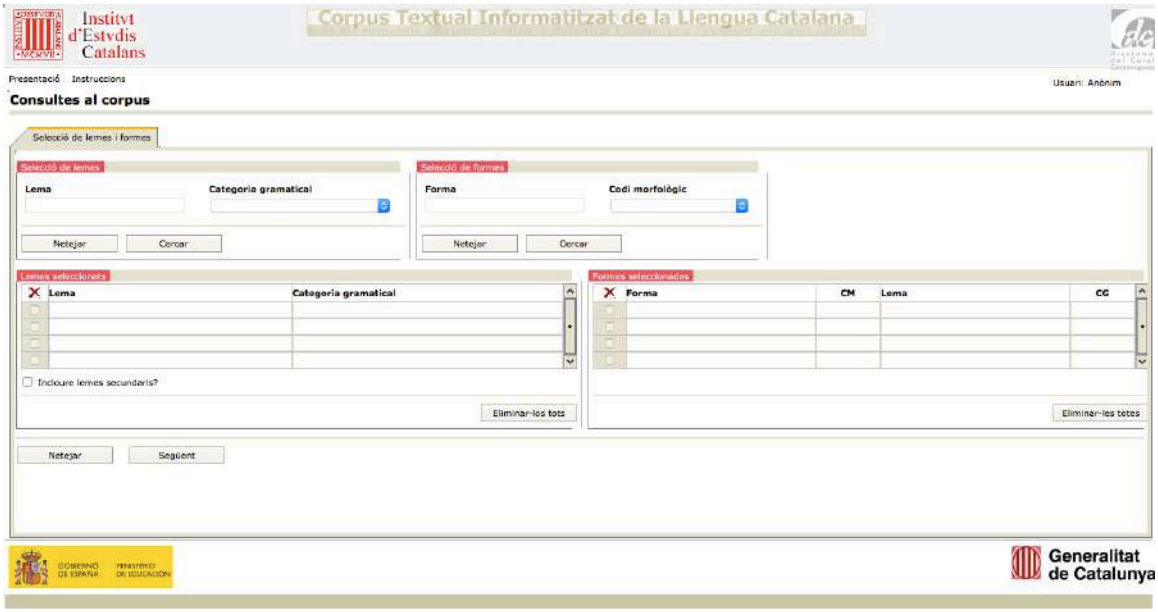
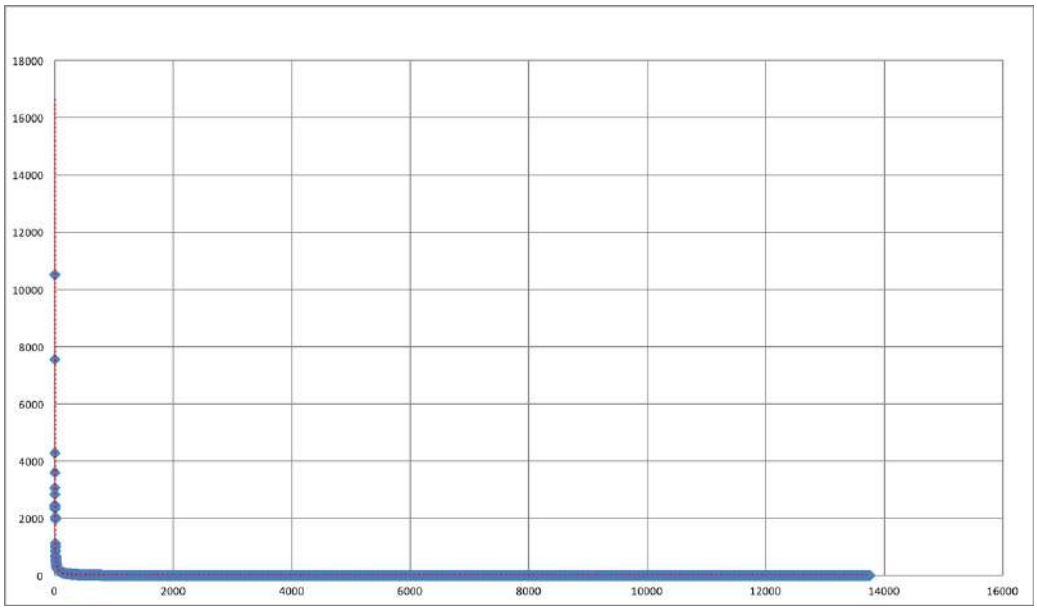


Fig. 13

Davant d'aquest fet, vam decidir abandonar aquesta cerca, però uns dies després vam reobrir la cerca per trobar corpus catalans proveïts per altres entitats. Efectivament, però no era del tot el volíem. Era l'anomenat WikiCorpus, un recull de totes les entrades a la Viquipèdia de l'any 2006 en català, castellà i anglès, proveït per la Universitat Politècnica de Catalunya. Els fitxers estaven separats en cada un dels tres idiomes. Malauradament, els fitxers en català estaven corruptes, el format del fitxer havia desplaçat accents i canviat caràcters.

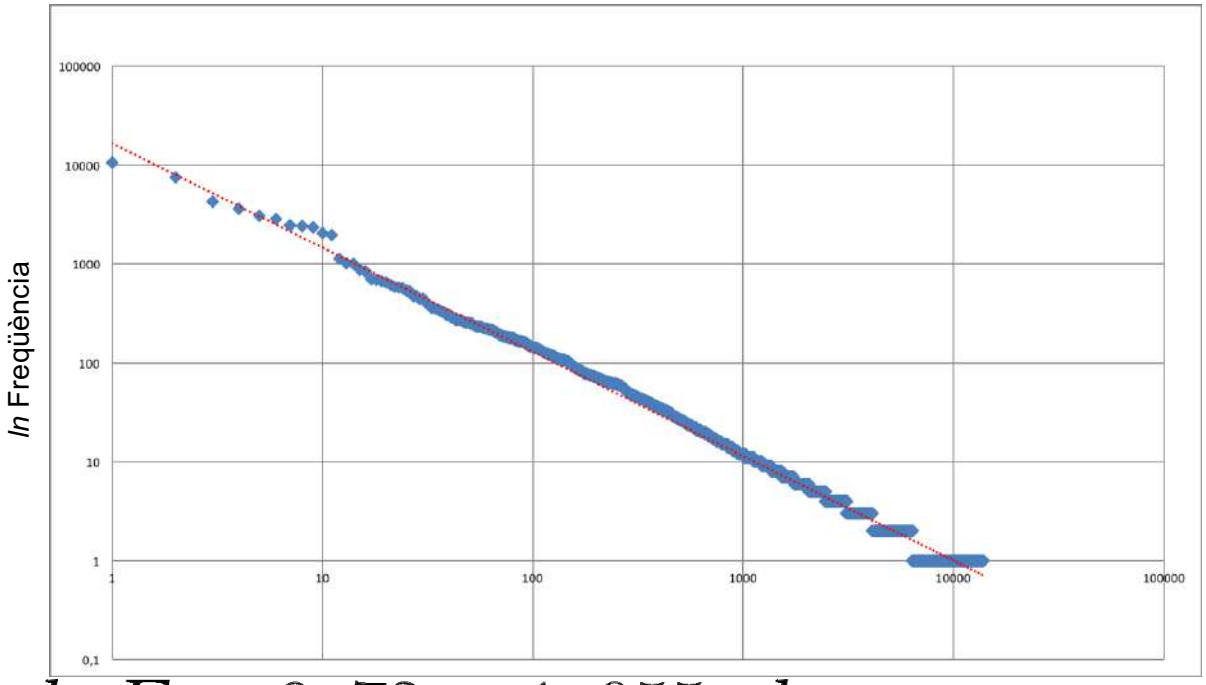
En resposta al fet de no haver trobat un corpus funcional hem decidit fer un corpus basat en tots els textos utilitzats en aquest treball per veure si el que s'acostava a una verificació individual també es pot observar quan es forma un conjunt amb les dades. El corpus conté 131.431 paraules, d'aquestes, 13.753 són úniques, en comparació al corpus Brown (mostrat al principi) aquest només té un desè de les paraules.



Gràfica 14

Rang $F = 16.622 \cdot r^{-1,055}$

Equació 23



$lnF = 9,72 - 1,055 \cdot lnr$

Equació 24

ln Rang

Gràfica 15

$R^2 = 0,97$

Equació 25

En el Corpus, resultat del recompte, té un coeficient de correlació molt elevat però no tant com els altres conjunts grans de text. Imaginem que aquest fet ve donat per la diferència d'autors. Les paraules utilitzades l'any 1490, el 1970 i l'any 2011 varien molt, tot i això la correspondència és bastant estricte.

Zipf fora dels textos

Durant la recerca d'aquest treball, molts autors i pàgines web han fet referència a l'aparició de la llei de Zipf en diversos entorns.

L'exemple que sempre es mostra és la llei de Zipf aplicada a les ciutats més grans.

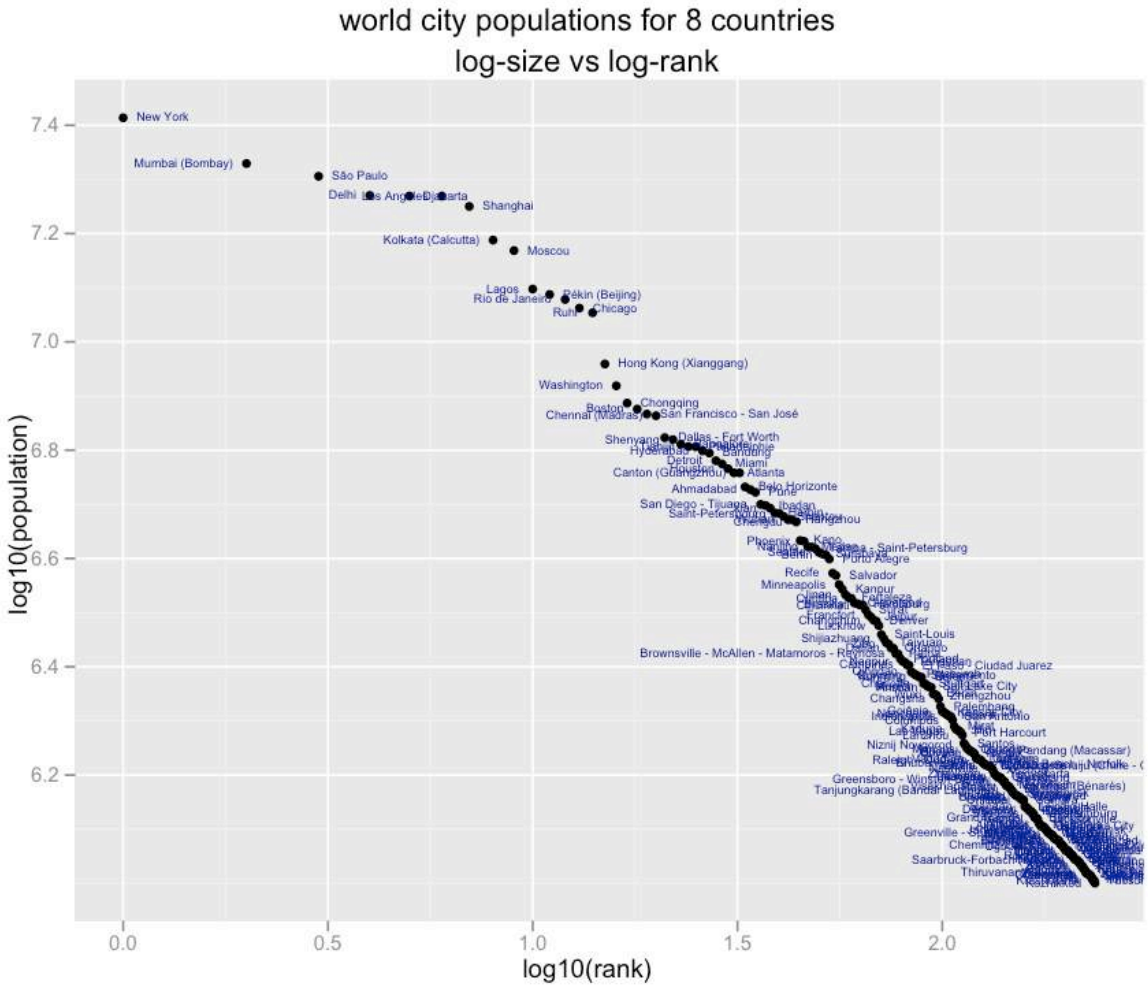


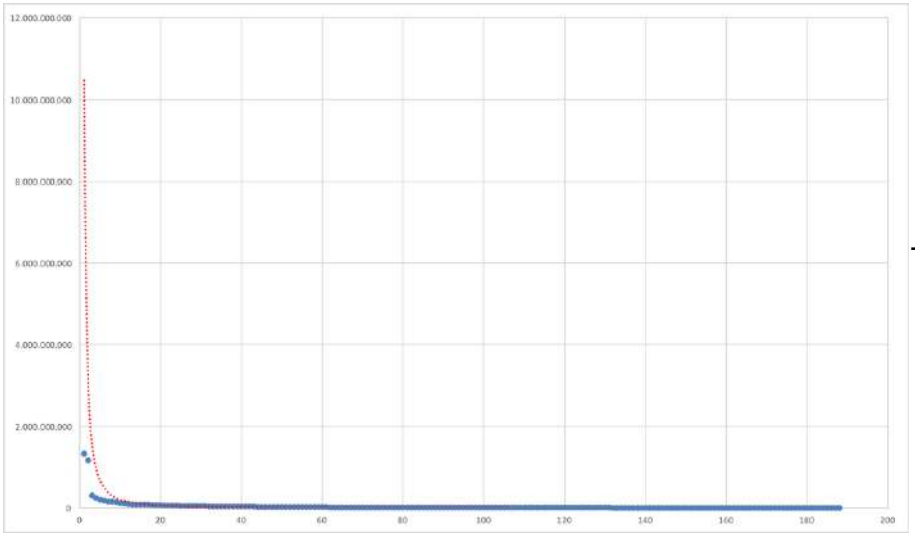
Fig. 14

La fig. 14 mostra una gràfica en una escala logarítmica les ciutats més grans del món ordenades per població. Tot i que les ciutats grans varien una mica, a mesura que la població va disminuint es veu clarament el que seria la línia de tendència.

Zipf al voltant del món

Ens atreia la idea d'estudiar la població, però no volíem fer una gràfica com la de la pàgina anterior, per això vam decidir, en comptes d'estudiar la població per ciutats, fer-ho per països. L'ordre i el nombre d'habitants que utilitzem són proveïts per Viquipèdia, l'any 2016.

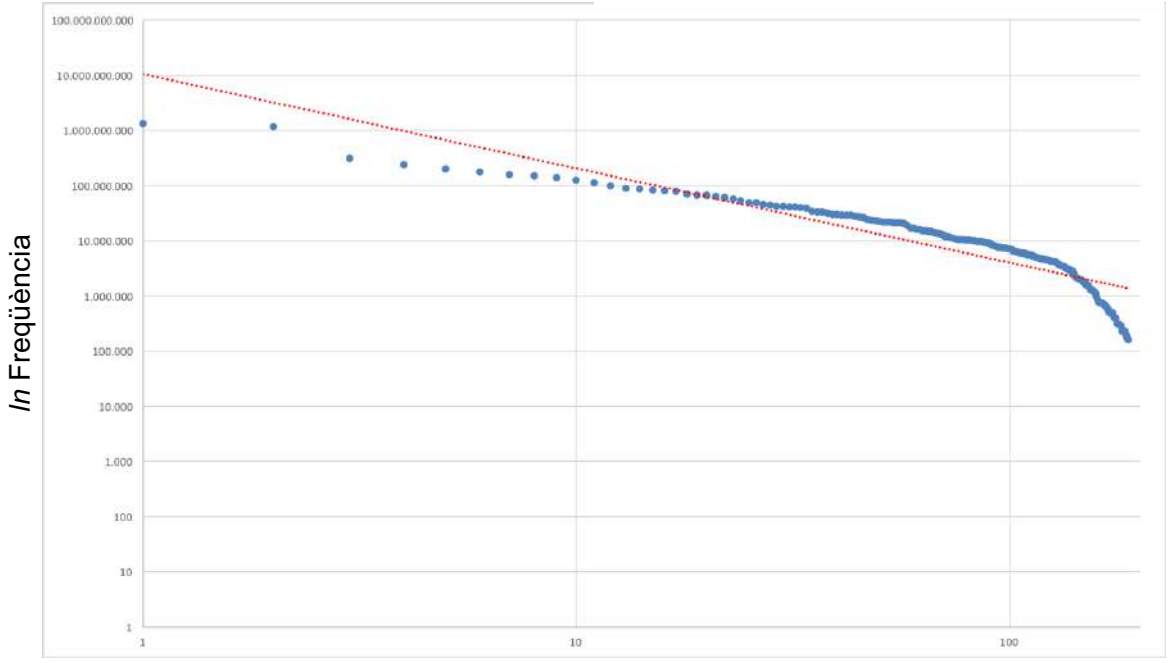
En l'estudi de població per països, no veiem una correlació tan evident com en els textos, però tot i així mostra una correlació bastant elevada.



Gràfica 16

$$F = 1 \cdot 10^{10} \cdot r^{-1,707}$$

Equació 26



Gràfica 17

$$lnF = 23,03 - 1,707 \cdot lnr$$

Equació 19

Equació 28

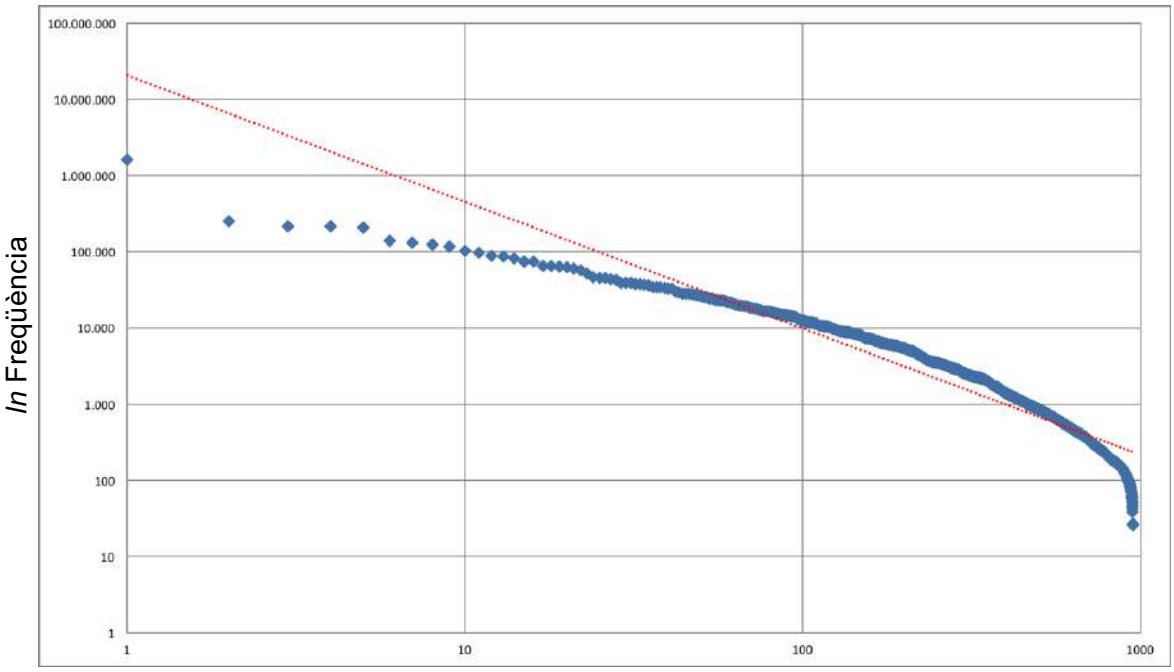
$$R^2 = 0,8269$$

La llei de Zipf a Catalunya

Com que el treball es basa a comprovar la llei de Zipf en català, especificar l'estudi a Catalunya i la seva població.

Catalunya és un pèl escassa si es parla de ciutats, però de municipis i comarques no n'hi falta. Gràcies a l'Institut d'Estadística de Catalunya vam poder trobar la població distribuïda en municipis i comarques de l'any 2015.

Municipis



Gràfica 18

A diferència dels textos, existeixen pocs rangs amb dades iguals, en els textos hi ha moltes paraules que comparteixen freqüència, mentre que els conjunts de població no comparteixen un mateix nombre d'habitants, tan freqüentment.

$$lny = 6,19 - 0,762 \cdot lnr$$

Equació 29

$$F = 2 \cdot 10^{7} \cdot r^{-1,66}$$

Equació 30

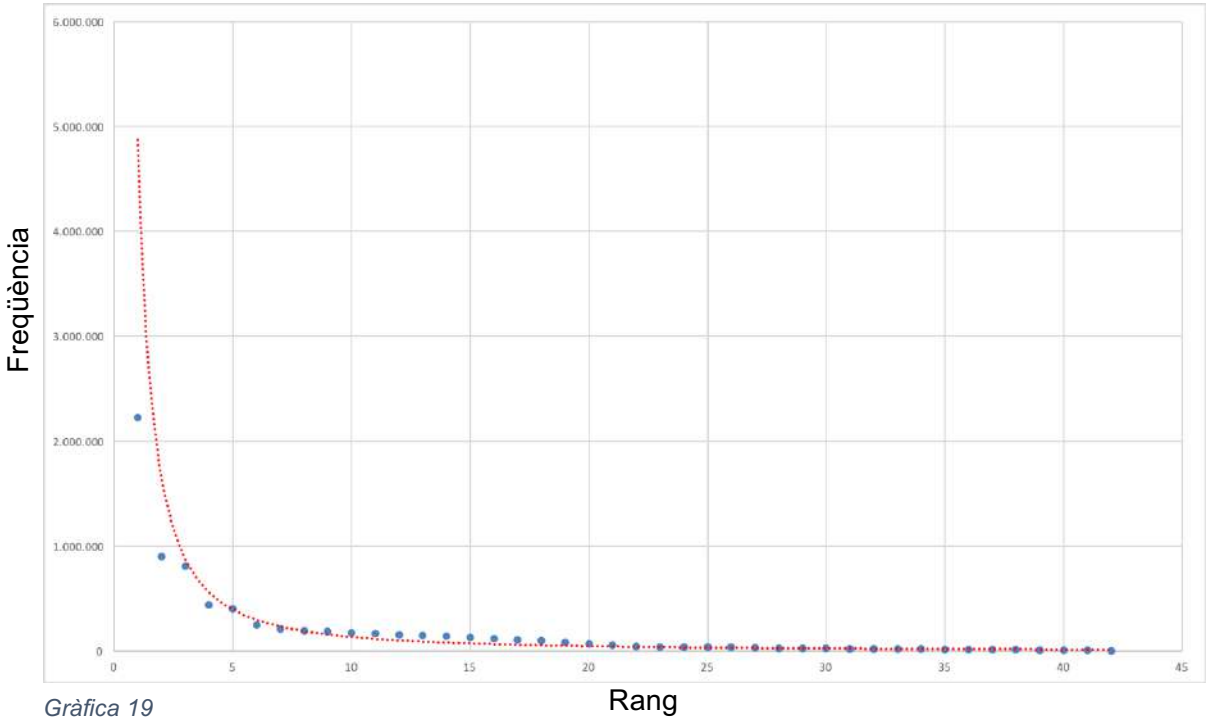
$$R^2 = 0,8269$$

Equació 20

Annex 1.9 per veure la gràfica amb eixos lineals

Comarques

No sabíem del tot si les poblacions de Catalunya seguirien un ordre zipfià. Però ha resultat tenir una relació més apreciable que l'esperada.



Gràfica 19

En les comarques, en comparació amb els municipis, s'aprecia més la semblança entre les dades i la previsió de la llei. Visualment és una de les gràfiques en què la línia de tendència passa per més punts.

$$F = 5 \cdot 10^6 \cdot r^{-1,55}$$

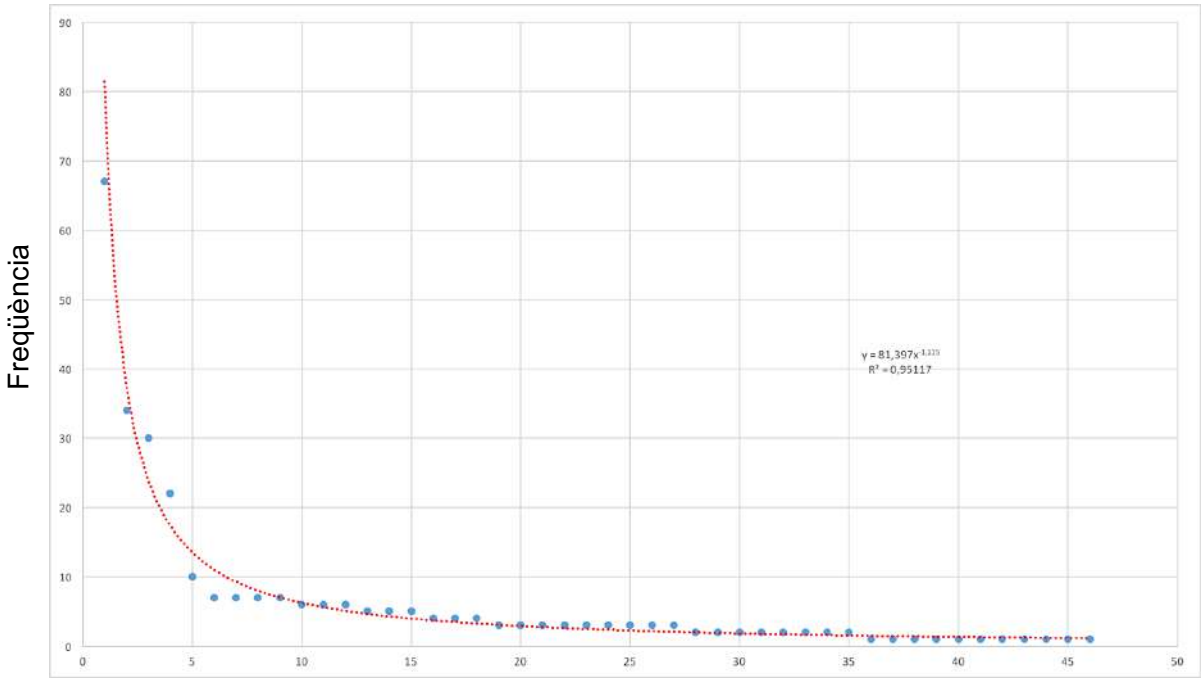
Equació 32

$$R^2 = 0,912$$

Equació 33

Missatges de text

Finalment, donat que avui en dia tothom està enganxat al mòbil, vam decidir estudiar la distribució de trucades en el mòbil de la mare per veure si la llei de Zipf s'hi mostrava. Malauradament, no vàrem aconseguir les factures telefòniques i només vàrem rebre la informació dels missatges de text enviats. Ho vam comprovar amb els missatges que teníem. Un recull de 290 missatges enviats a 49 destinataris diferents, en dotze mesos.



Gràfica 20

Tot i tenir un conjunt de dades inferior al desitjat, els missatges varen donar bons resultats quan ajustats a la llei. És impressionant veure que la llei de Zipf s'aplica a tants camps diferents.

$$F = 81,397 \cdot r^{-1,115}$$

Equació 34

$$R^2 = 0,951$$

Equació 35

Conclusions

Principalment, l'objectiu del treball era desenvolupar i demostra la llei de Zipf en la llengua catalana.

Els resultats mostren coeficients de correlació entre les dades dels textos i la llei de Zipf molt elevats. Recordant que aquest coeficient s'expressa entre -1 i 1, tots els textos estableixen una correlació més gran que 0,8 i la majoria més gran que 0,9.

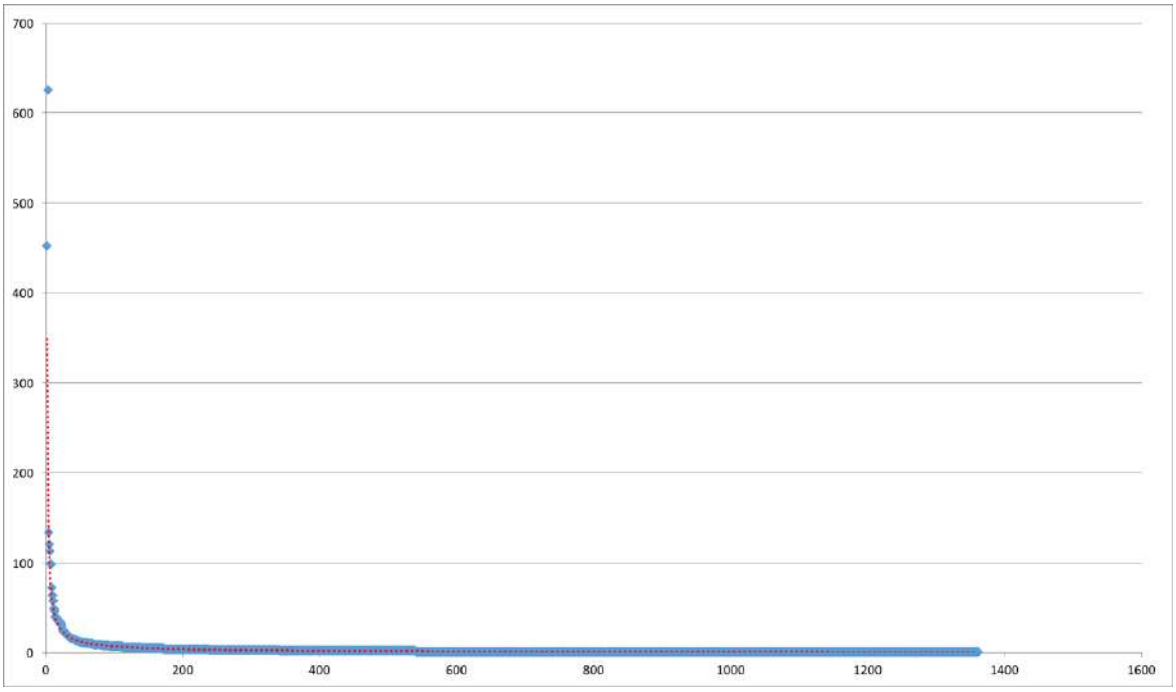
Es pot afirmar, a partir de l'estudi, que la llei de Zipf es confirma en la llengua catalana. El català és una llengua Zipfiana.

Personalment, els resultats han sigut superiors a les meves expectatives, sobretot en conjunts de dades petits com l'himne i les cançons, i també en la distribució de població a Catalunya.

Si el projecte continués en marxa, voldria aconseguir un conjunt de dades més gran, com per exemple el WikiCorpus, format per tots els articles escrits en català a la Viquipèdia o el Corpus català de l'Institut dels Estudis Catalans.

Per acabar, volíem mostrar la llei de Zipf aplicada a aquest treball de recerca. El qual, conté 1361 paraules diferents que formen un text de 5650 paraules.

Gràfica 21



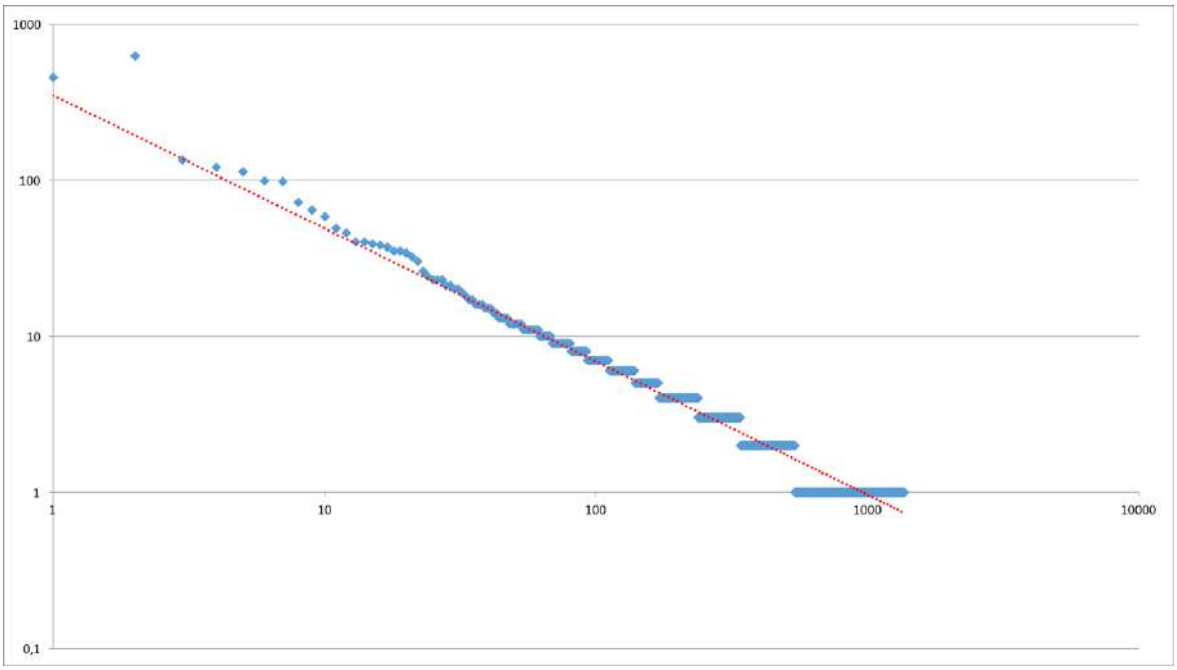
$$F = 348,73 \cdot r^{-0,853}$$

Equació 36

$$R^2 = 0,96$$

Equació 37

Efectivament, veiem que el treball dedicat a la llei de Zipf, confirma la llei de Zipf.



Gràfica 22

$$lnF = 5,85 - 0,853 \cdot lnr$$

Equació 21

Annex de les taules

1. L'himne dels Segadors
    a. Original
    b. Corregit

2. Empordà – Sopa de Cabra
    a. Original
    b. Corregit

3. Boig per tú – Sau
    a. Original
    b. Corregit

4. Diari La Vanguardia
    a. Original
    b. Corregit

5. Episodis amorosos de Tirant lo Blanc
    a. Original
    b. Corregit

6. Mecanoscrit del segon origen
    a. Original
    b. Modificant
    c. Corregit

7. Corpus
    a. Original
    b. Corregit

8. Llista de països ordenats per poblacions
9. Llista de municipis de Catalunya
10. Comarques catalanes
11. Missatges de text durant un l'any 2012

Bibliografia

Articles en suport digital:

ADAMIC, Lada A. “Zipf, Power-laws, and Pareto –a ranking tutorial” (en línia) Publicat per Information Dynamics Lab. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html> [Consulta: 5 agost 2016].

BAILÓN-MORENO, R., JURADO-ALAMEDA, E., RUIZ-BAÑOS, R., COURTIAL,J.P., “The Unified Scientometric Model. Fractality and transfractality” Publicat a Scientometrics, Vol. 63 Nº2, 2005 per la Univerdidad de Granada y Université de Nantes <http://www.bordalierinstitute.com/scientometricModel.pdf> [Consulta:30 setembre 2016]

FROST, Jim, “Regression Analysis: How do I interpre R-squared and asses the goodness-of-it?”, 2013 <http://blog.minitab.com/blog/adventures-in-Consulta:of-fit> [Consulta: 27 setembre a 2 octubre].

LI, W. “Random texts exhibit Zipf’s-law-like word frequency” (en línia) Publicat per IEET Transactions on Information Theory (Volume 38, Issue:6, Novembre 1992) <http://ieeexplore.ieee.org/document/165464/ > [Consulta: 5 agost 2016].

NORDQUIST, Richard “(principle of) least effort [Zipf’s Law]” Publicat per About Education, 2016 <http://grammar.about.com/od/il/g/Least-Effort.htm>[Consulta:18 setembre 2016].

O’CONNOR, Brendan, “Zipf’s law and world city populations” Publicat a AI and Social Science, 2009 <https://brenocon.com/blog/2009/05/zipfs-law-and-world-city-populations/ >[Consulta: 2 octubre 2016].

POWERS, David M.W. “Applications and Explanations of Zipf’s Law” (en línia) Publicat per Department of Computer Science, The Flinders University of South Australia, 1998 <http://www.aclweb.org/anthology/W98-1218> [Consulta: 5 agost 2016].

WEST, Marc “The mystery of Zipf” (en línia) Publicat per +plus Magazine. <https://plus.maths.org/content/mystery-zipf > [Consulta: 7 agost 2016].

WIKIPEDIA, “Zipf’s Law”, 2016 <https://en.wikipedia.org/wiki/Zipf%27s_law> [Consulta 5 agost a 2 octubre 2016].

WIKIPEDIA, “Power Law”, 2016 <https://en.wikipedia.org/wiki/Power_law> [Consulta:18 septembre a 2 octubre 2016].

WIKIPEDIA, “Pareto Principle”, 2016 <https://en.wikipedia.org/wiki/Pareto_principle>

WIKIPEDIA, “Zipf’s Law”, 2016 https://en.wikipedia.org/wiki/Zipf%27s_law

WIKIPEDIA, “Llista d’estats per població”, 2016 <https://ca.wikipedia.org/wiki/Llista_d%27estats_per_població>

Llibres en suport digital:

HARALD BAAYEN, R. “Word Frequency Distributions” Kluver Academic Publishers per Springer Science and Business Media Dordrecht. The Netherlands, 2001.

ZIPF, George K. “The Psycho-Biology of Language – An introduction to Dynamic Phililogy”, The MIT Press, Estats Units, 1965.

ZIPF, George K. “The P1 P2/D Hypothesis: On the Intercity Movement of Persons”, American Sociological Review, Vol. 11 No. 6 pp.677-686, 1946

ZIPF, George K. “Selected Studies of the Principle of Relative Frequency in Language”, Harvard University Press, Cambridge, Massachusets, Estats Units, 1932

Pàgines web com a eines de suport:

<www.textfixer.com> [19 setembre 2016]
<http://www.hermetic.ch/wfc/wfc.htm> [25 juliol a 2 octubre 2016]
<http://textmechanic.com/text-tools/basic-text-tools/count-characters-words-lines/> [1 agost a 2 octubre 2016]
<http://www.writewords.org.uk/word_count.asp> [25 juliol a 2 octubre 2016]

<http://www.csgnetwork.com/documentanalystcalc.html> [1 agost a 2 octubre 2016]
<https://support.office.com/en-us/article/Choosing-the-best-trendline-for-your-data-1bb3c9e7-0280-45b5-9ab0-d0c93161daa8> [26 setembre a 2 octubre]
<https://www.math.vt.edu/courses/math2015/Labs/Excel/TrendLine.html>[26 setembre a 2 octubre]
<https://www.codecogs.com/latex/eqneditor.php> [19 setembre a 2 octubre]

YouTube:

STEVENS, Michael, V-Sauce, 2015 <https://www.youtube.com/watch?v=fCn8zs912OE>