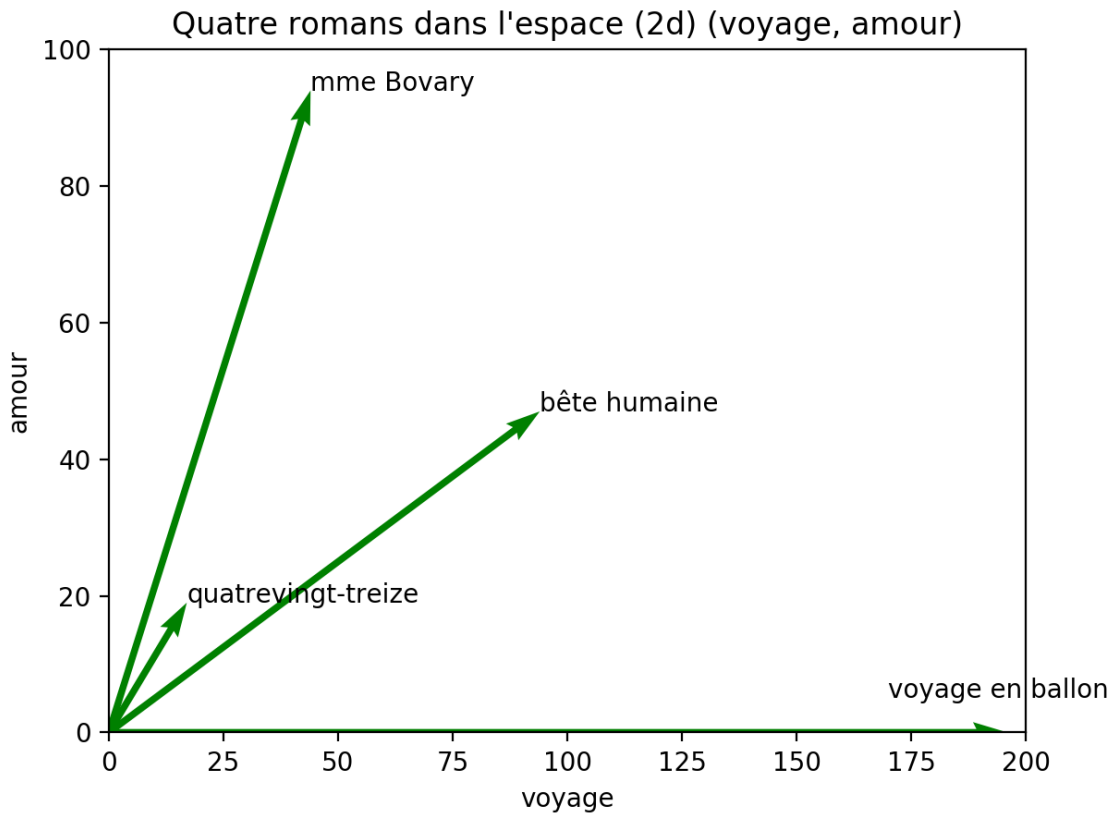
- *If A and B have almost identical environments we say that they are synonyms.* *(Harris, 1954)*
- *You shall know a word by the company it keeps.* *(Firth, 1957)*
- *The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear.* *(Lenci, 2008)*

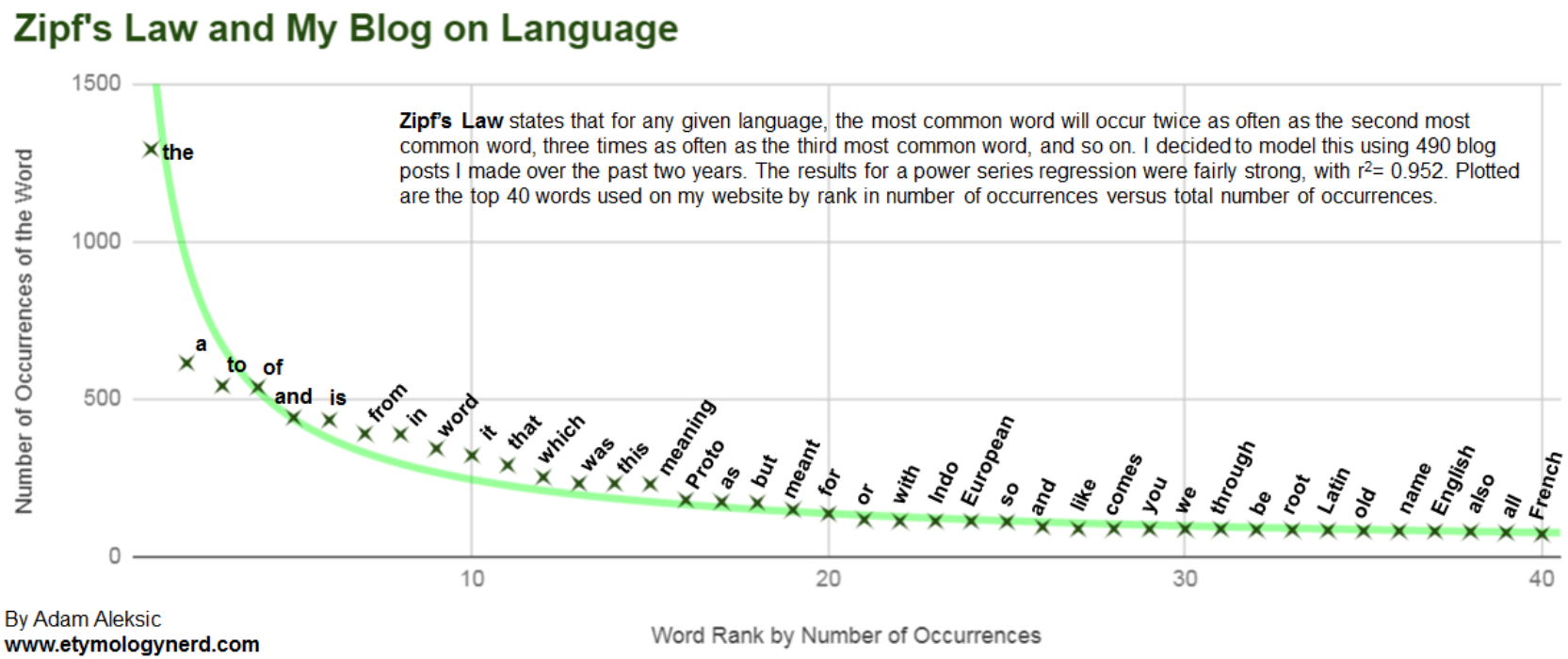
	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Mme Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

Table – Matrice terme-documents pour quelques mots et 4 romans (Quatrevingt-treize (Hugo) ; Le voyage en ballon (Verne) ; La bête humaine (Zola) ; Mme Bovary (Flaubert)).

	QuatreVT	Voyage Bal	Bête Hum.	Mme Bovary
voyage	17	196	94	44
amour	19	0	47	94



Loi de Zipf


Zipf's Law and My Blog on Language
Zipf's Law states that for any given language, the most common word will occur twice as often as the second most common word, three times as often as the third most common word, and so on. I decided to model this using 490 blog posts I made over the past two years. The results for a power series regression were fairly strong, with r²= 0.952. Plotted are the top 40 words used on my website by rank in number of occurrences versus total number of occurrences.
Number of Occurrences of the Word
1500
1000
500
0
10
20
30
40
the
a
to
of
and
is
from
in
word
it
that
which
was
this
meaning
Proto
as
but
meant
for
or
with
Indo
European
so
and
like
comes
you
we
through
be
root
Latin
old
name
English
also
all
French
By Adam Aleksic
www.etymologynerd.com
Word Rank by Number of Occurrences

tf-idf : formules

$$tf(t,d) = \begin{cases} \log(1 + f_{t,d}) & \text{si } f_{t,d} > 0 \\ 0 & \text{sinon} \end{cases}$$

$$idf(t,D) = \log\left(\frac{|D|}{df(t)}\right)$$

$$td\text{-}idf(t,d) = tf(t,d,D) \times idf(t,D)$$

bataille	(35,2)
politique	(11,5)
amour	(19,94)
voyage	(17,44)

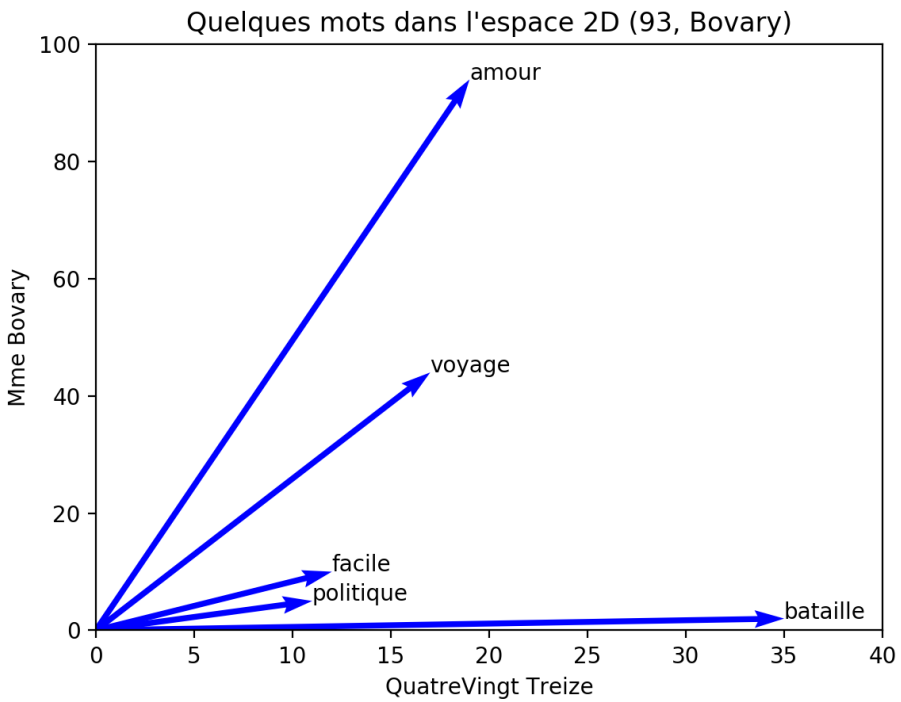


Figure – représentation graphique des mots dans le plan (93, Bovary)

KWIC voyagiste

je connaissais le	voyagiste	qui les amène
voyages : Différent du	voyagiste	ou tour-opérateur . En
compagnie à un	voyagiste	. Animaux domestiques : Lorsqu'
que disent certains	voyagistes	, légalement , aucune n'
Transport / Hébergement Le	voyagiste	FRAM propose les
plus vite votre	voyagiste	. Le Comité scientifique
le monde ... 6ème	voyagiste	européen , Kuoni est
autres , les grands	voyagistes	essayent de prendre
fév 2008 Ces	voyagistes	, qui opèrent principalement
parmis les meilleurs	voyagistes	en ligne . Service
les 15 principaux	voyagistes	en ligne en
ses amies des	voyagistes	locaux qui fêtent
foyers ...) et des	voyagistes	. La commercialisation Le
lancé vers des	voyagistes	susceptibles d' offrir
clients chez un	voyagiste	- Développement d' un
des plus grands	voyagistes	européens dont ils
voyage à cheval .	Voyagiste	spécialiste du voyage
clients chez un	voyagiste	- Développement d' un
sur cette déstination	Voyagiste	Catalogue déstination du
hôtel , à votre	voyagiste	, à votre ambassade

frWaC, consultation 2021-08-11

	bataille	voyage	homme	femme
arriver	246	470	1 819	890
tomber	100	83	1 205	384
habiller	2	4	339	384
mourir	180	116	339	1 088
	55 331	208 520	668 289	346 093

Table – Matrice Terme-Terme obtenue dans frWaC. Contextes en ligne.

	arriver	tomber	habiller	mourir	
bataille	246	100	2	180	55 331
voyage	470	83	4	116	208 520
homme	1 819	1 205	339	1 499	668 289
femme	890	660	384	1 088	346 093

Table – Matrice Terme-Terme obtenue dans frWaC. Contextes en colonnes.

f	aardvark	computer	data	pinch	result	sugar
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

f	computer	data	pinch	result	sugar	Σ
apricot	0	0	1	0	1	2
pineapple	0	0	1	0	1	2
digital	2	1	0	1	0	4
information	1	6	0	4	0	11
Σ	3	7	2	5	2	19

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0/19	0/19	1/19	0/19	1/19	2/19
pineapple	0/19	0/19	1/19	0/19	1/19	2/19
digital	2/19	1/19	0/19	1/19	0/19	4/19
information	1/19	6/19	0/19	4/19	0/19	11/19
p_{*b}	3/19	7/19	2/19	5/19	2/19	19/19

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0,00	0,00	0,05	0,00	0,05	0,11
pineapple	0,00	0,00	0,05	0,00	0,05	0,11
digital	0,11	0,05	0,00	0,05	0,00	0,21
information	0,05	0,32	0,00	0,21	0,00	0,58
p_{*b}	0,16	0,37	0,11	0,26	0,11	

ppmi	computer	data	pinch	result	sugar
apricot	-	-	2,25	-	2,25
pineapple	-	-	2,25	-	2,25
digital	1,66	0,00	-	0,00	-
information	0,00	0,57	-	0,47	-

f	computer	data	pinch	result	sugar
apricot	2	2	3	2	3
pineapple	2	2	3	2	3
digital	4	3	2	3	2
information	3	8	2	6	2

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0,03	0,03	0,05	0,03	0,05	0,20
pineapple	0,03	0,03	0,05	0,03	0,05	0,20
digital	0,07	0,05	0,03	0,05	0,03	0,24
information	0,05	0,14	0,03	0,10	0,03	0,36
p_{*b}	0,19	0,25	0,17	0,22	0,17	

ppmi	computer	data	pinch	result	sugar
apricot	-	-	2,25	-	2,25
pineapple	-	-	2,25	-	2,25
digital	1,66	0,00	-	0,00	-
information	0,00	0,57	-	0,47	-
ppmi [add2]	computer	data	pinch	result	sugar
apricot	0,00	0,00	0,56	0,00	0,56
pineapple	0,00	0,00	0,56	0,00	0,56
digital	0,62	0,00	0,00	0,00	0,00
information	0,00	0,58	0,00	0,37	0,00

Figure – Comparaison de mesures de *ppmi* sur l'exemple de (Jurafsky & Martin, 2019), avec ou sans lissage laplacien +2

Références

Firth, John Rupert. 1957. *Papers in Linguistics (1934-1951)*. Oxford : Oxford University Press.

Harris, Zellig S. 1954. Distributional structure. *Word*, **10**(2-3), 146–162.

Jurafsky, Daniel, & Martin, James H. 2019. *Speech and Language Processing : An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall. drafts of August 29, 2019. Chap. 6 Vector Semantics.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, **20**(1), 1–31.