Word Distributions and Zipf's Law

J. Savoy
Université de Neuchâtel

C. D. Manning & H. Schütze : *Foundations of statistical natural language processing*. The MIT Press. Cambridge (MA)

P. M. Nugues: *An introduction to language processing with Perl and Prolog*. Springer. Berlin

R. H. Baayen : *Word Frequency Distributions*. Kluwer. Drodrecht

What is a word?

- Select the word as unit of measurement

我不是中国人

我　不　是　中国人

I　not　be　Chinese

- Other possibilities
letters, lemmas, grammatical categories, syntactic structures, themes
- But … What is a word? Sequence of letters?

What is a word?

- But… What is a word? Sequence of letters?
- Examples

  Richard Brown is painting in New York (or in NY)

  I'll send you Luca's book

  l'école, d'aujourd'hui

  le chemin de fer

  C|net

  Micro$oft

  IBM360, IBM-360, ibm 360, …
- Sequence of letters and digits?
- And the uppercase / lowercase

What is a word?

- The same word?
  - Richard *Brown*
    brown paint
    Brown is the …
  - Database system
    data base system
    data-base system (hyphen ?)
  - I *saw* a man with a *saw* (homograph)

What is a word?

- Particular problem with the "-"
  - the aluminium-export ban
  - a text-based medium
  - a final "take-it-or-leave-it" offer
  - the 45-year old
  - the New York-New Haven railroad

What is a word?

- Sometimes tricky:
  - Dates: 28/02/96 (French & British),
    2002/11/20/ (US, Swedish)
  - Numbers: 9,812,345 (English),
    9 812,345 (French and German) or
    9,812.345 (Old fashioned French)
  - Abbreviations: km/h. m.p.h.
  - Acronyms: S.N.C.F., UN, EU, US (but not the pronoun)

Frequency

- Select a sample (document/corpus) of size *n* of word tokens
- Example

  "The world considered the United States as a young country. Today, we are the world's oldest constitutional democracy."
- Count

  19 word *tokens (forme)*

  16 word *types (vocable)* {a, as, are, considered, constitutional, country, democracy, oldest, s, States, the, today, United, we, world, young}

  E.g.. the word type "the" appears three times

Frequency

- Counting the word *types (vocable)* means counting the vocabulary size

  Denote by V the vocabulary
  E.g., V = {country, democracy, States, the, United}
  and its size is |V| = 5 (cardinality of a set)
- Counting the number of tokens (*forme*) means counting the sample / document / corpus size
  Use *n* to indicate this size
- Usually *n* > |V| because some word types appear more than once in a sample / document / corpus.
- Use f(ω) to indicate the frequency (number of occurrences) of a given word ω in a sample (e.g., f("the") = 3)

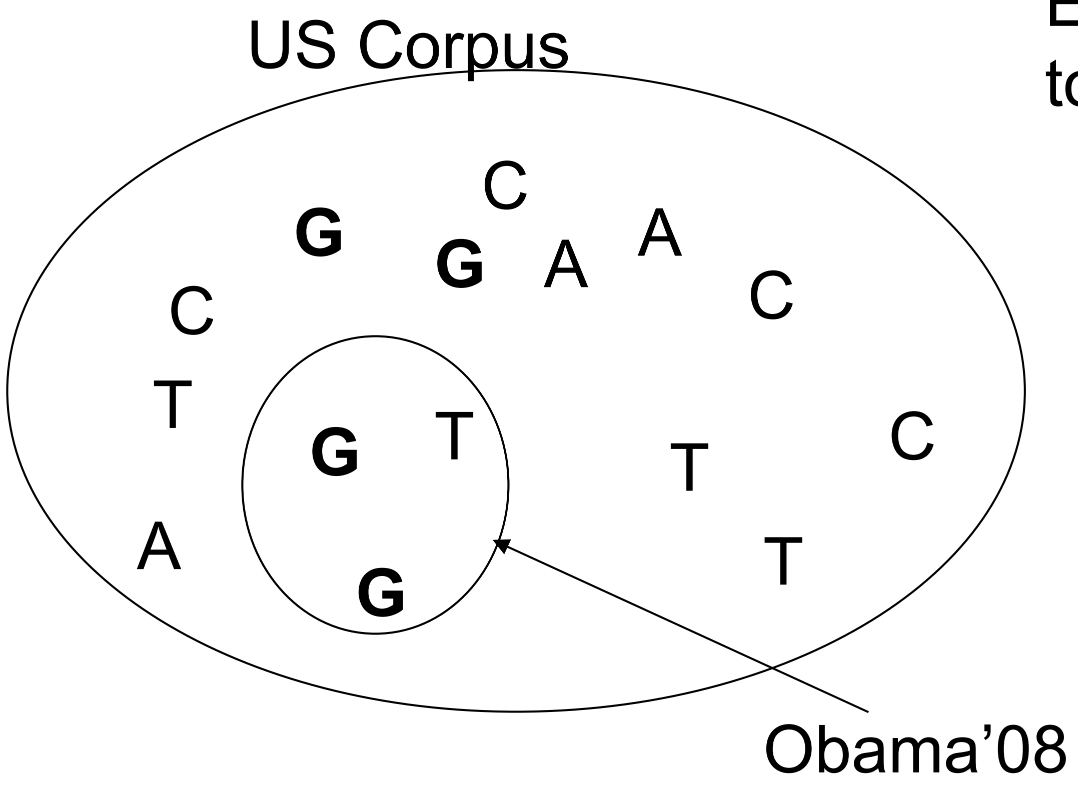
Frequency

- Given a corpus. can we model the word distribution?
- Can we find general law(s) governing the word distribution?
- Are words used randomly?
- Does the word distribution differ from one author to the other?
- Can we find constant(s) when analyzing the word distribution of a given author within a given genre? A set of authors in a given genre? An author in general?
- Can we use such information to describe an author’s style?

Our US Corpus

US: all speeches given by B. Obama & J. McCain during the years 2007 & 2008

Example with 15 tokens and 4 types



Our US Corpus

- Speeches given by Senator Barack Obama

  150 speeches from Feb., 10th 2007

  420,410 tokens, 9,014 types

  For 2008 only: 113 speeches

  294,553 tokens, 7,663 types

  http://www.barackobama.com/

- Speeches given by Senator John McCain

  94 speeches. from Apr., 25th 2007

  206,899 tokens, 9,401 types

  For 2008 only: 71 speeches

  154,365 tokens, 7,792 types

  http://www.johnmccain.com/

Frequency

The most frequent word types f(ω)

With
|V| = 7,792
for J. McCain and
|V| = 7,663
for B. Obama
the number of distinct types (or vocabulary size)

	McCain'08		Obama'08	
Rank	**Word**	**f(ω)**	**Word**	**f(ω)**
1	the	7759	the	13027
2	and	6157	and	10950
3	to	5413	to	9072
4	of	4773	that	7446
5	in	3137	of	6985
6	a	2940	we	6203
7	I	2345	a	5562
8	that	2243	in	5340
9	we	2160	is	4986
10	for	1762	I	4216

Frequency (Brown Corpus)

Collected in 1961

A real sample

1,014,312 tokens

Given by lemmas (e.g., “be” = “is”, “was”, “be”, “were”, etc.)

Rank	Word	Freq.	%
1	the	69975	6.90%
2	be	39175	3.86%
3	of	36432	3.59%
4	and	28872	2.85%
5	to	26190	2.58%
6	a	23073	2.28%
7	in	20870	2.06%
8	he	19427	1.92%
9	have	12458	1.23%
10	it	10942	1.08%

Zipf's Law

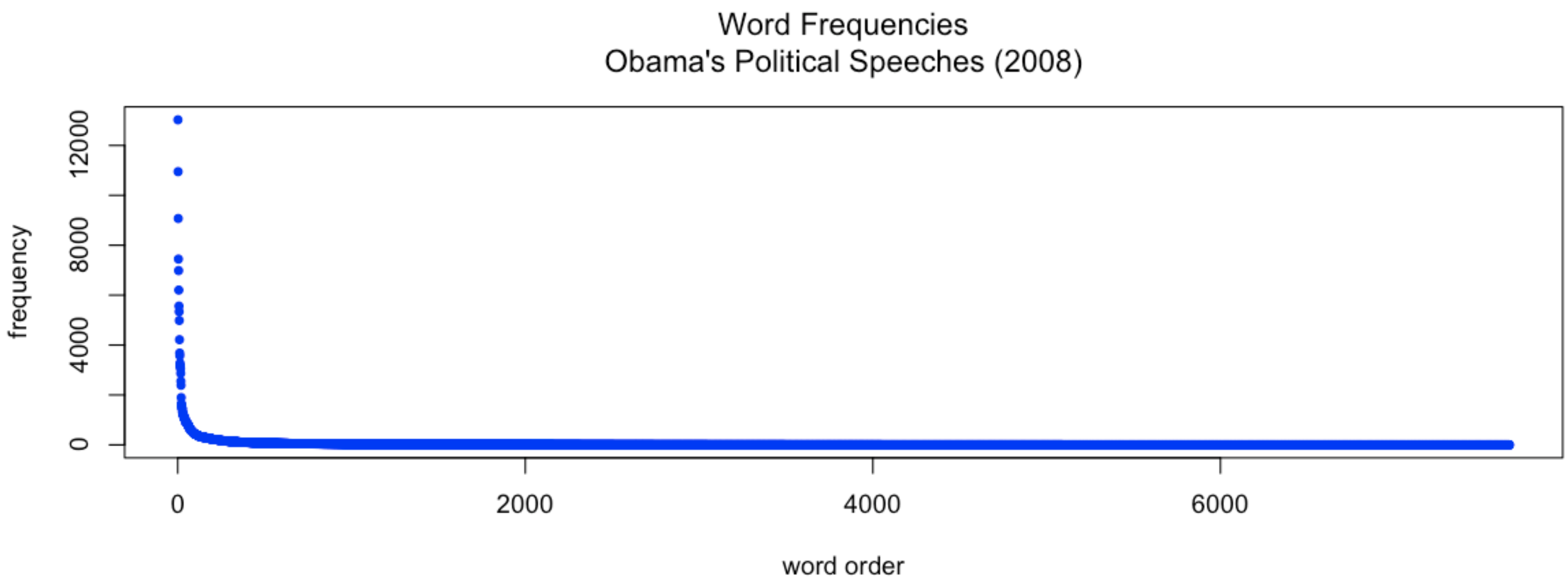
- More a regularity than a strict law
- The frequency (of a word type) ($f(\omega)$) is related to the inverse of its rank (z) (with $\alpha = 1$ for Zipf)
- We could use the absolute frequency ($f(\omega)$) of the relative frequency ($f(\omega)/n$)

$$f(\omega) = \frac{c}{z^{\alpha}} = c \cdot z^{-\alpha}$$

- Based on Obama's Speeches (2008)
  max frequency: 13027 ("the")
  number of types: 7663
- Graph: from the most frequent ("the") to the less frequent

Zipf's Law

From Obama's
speeches in 2008



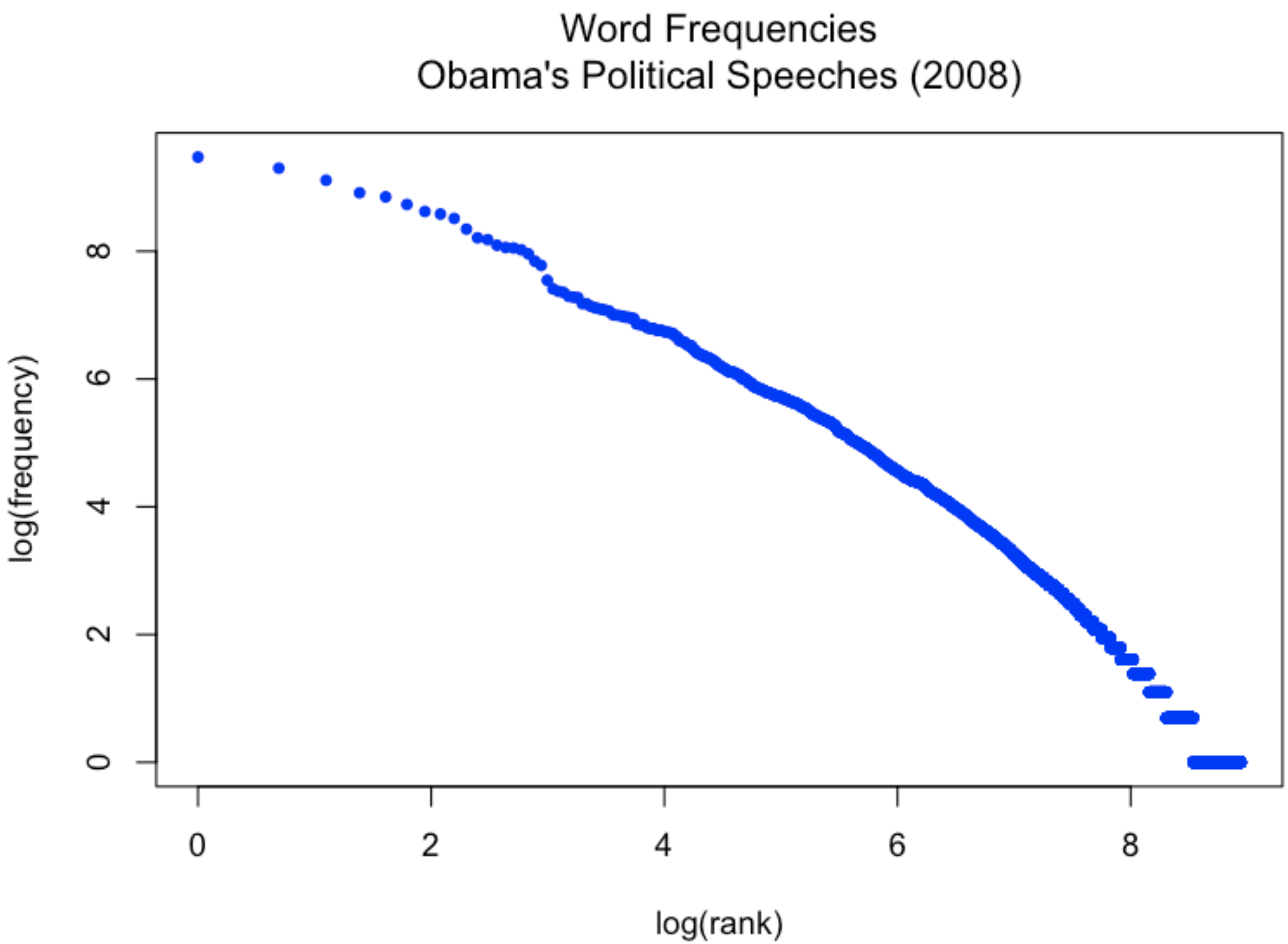
Zipf's Law

- The Zipf's law could be more useful when considering the log-log relationship between the absolute frequency (f(ω)) and the rank (z)

$$
\begin{aligned}
f(\omega) &= \frac{c}{z^{\alpha}} = c \cdot z^{-\alpha} \\
\text{we may obtain} & \\
log(f(\omega)) &= log\left(\frac{c}{z^{\alpha}}\right) \\
&= log(c) - \alpha \cdot log(z) = \beta - \alpha \cdot log(z)
\end{aligned}
$$

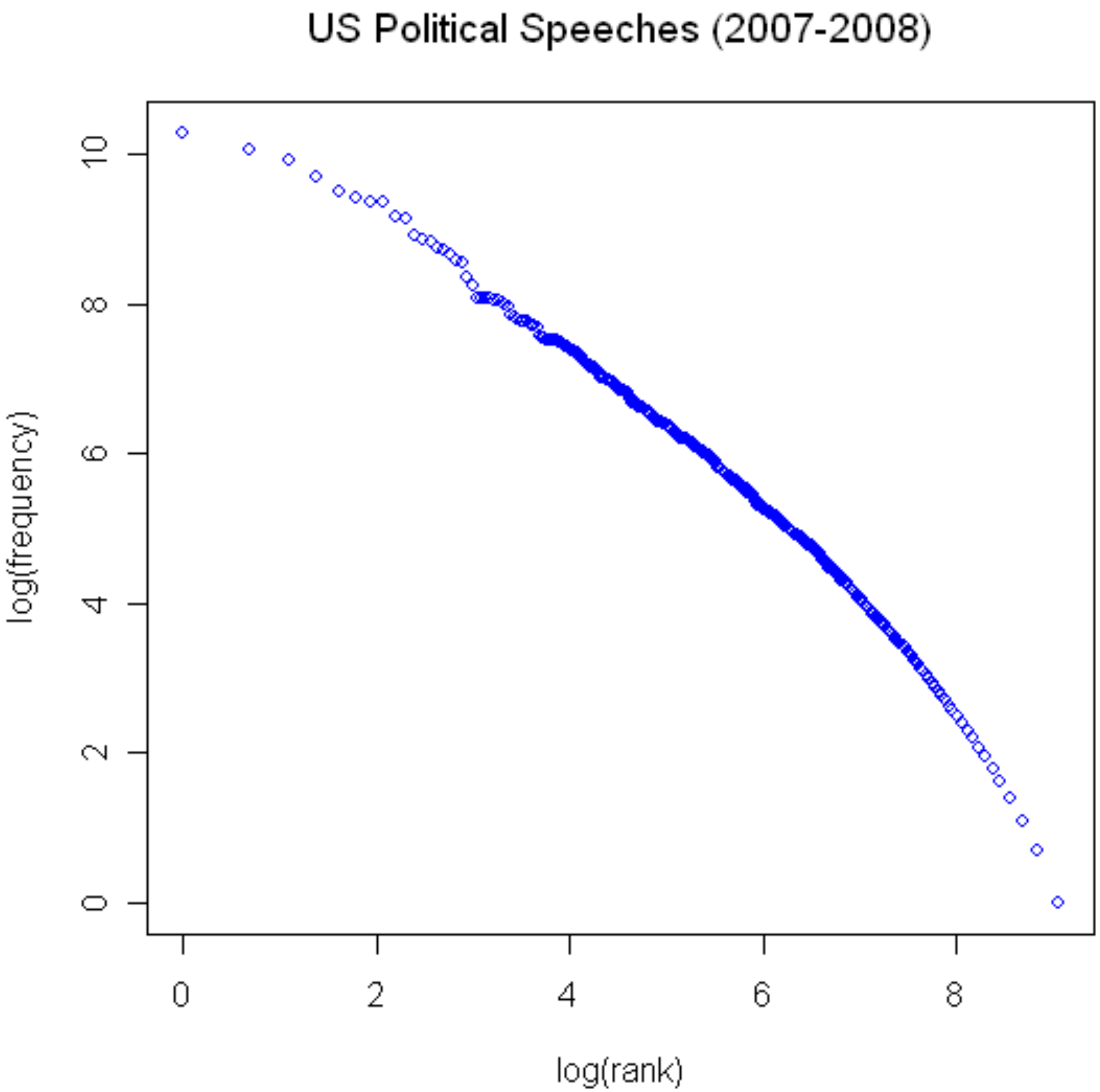
- Zipf's law is an example of power law
another example is the 80-20 rule
- Property: scale invariant

Zipf’s Law


Word Frequencies
Obama's Political Speeches (2008)
log(frequency)
0
2
4
6
8
log(rank)
0
2
4
6
8

Zipf's Law

Using the US corpus with |V| = 12,573



Zipf's Law (French Language)

- From the French language
- Based on the newspaper *Le Monde* and ATS
- 34,508,866 tokens and 251,017 types (*vocables*)
- With the first 16 most frequent types, we cover around 30% of all French documents (news articles)

Rank	Word	Freq. $f(\omega)$	Rel. Freq.	Cumul.	r x freq.
1	de	1,891,468	0.0548	0.0548	0.0548
2	la	1,062,987	0.0308	0.0856	0.0616
3	l	811,217	0.0235	0.1091	0.0705
4	le	807,145	0.0234	0.1325	0.0936
5	à	682,670	0.0198	0.1523	0.0989
6	les	657,241	0.0190	0.1713	0.1143
7	et	592,668	0.0172	0.1885	0.1202
8	des	584,412	0.0169	0.2054	0.1355
9	d	548,764	0.0159	0.2214	0.1431
10	en	477,379	0.0138	0.2352	0.1383
11	du	439,227	0.0127	0.2479	0.1400
12	a	409,561	0.0119	0.2598	0.1424
13	un	394,582	0.0114	0.2712	0.1486
14	une	335,561	0.0097	0.2809	0.1361
15	est	279,495	0.0081	0.2890	0.1215
16	dans	265,387	0.0077	0.2967	0.1231

Zipf's Law (German Language)

- Based on the newspaper *NZZ, Der Speigel,* and SDA
- 70,000,000 tokens and 1,081,681 types (*vocables*)
- With the first 16 most frequent types, we cover more than 20% of all German documents (news articles)

Rank	Word	Freq.	Rel. Freq.	Cumul.	r x freq.
1	der	2,420,534	0.0346	0.0346	0.0346
2	die	2,407,558	0.0344	0.0690	0.0688
3	und	1,489,787	0.0213	0.0902	0.0639
4	in	1,243,042	0.0178	0.1080	0.0710
5	den	790,054	0.0129	0.1193	0.0564
6	von	668,300	0.0095	0.1288	0.0573
7	das	668,163	0.0095	0.1384	0.0668
8	mit	586,284	0.0084	0.1468	0.0670
9	im	568,533	0.0081	0.1549	0.0731
10	zu	556,061	0.0079	0.1628	0.0794
11	für	534,454	0.0076	0.1705	0.0840
12	des	489,420	0.0070	0.1775	0.0839
13	auf	481,672	0.0069	0.1843	0.0895
14	sich	456,291	0.0065	0.1909	0.0913
15	dem	429,675	0.0062	0.1970	0.0921
16	ein	421,569	0.0060	0.2030	0.0964

Zipf’s Law (Spanish Language)

- Based on the news agency *EFE*
- 71,987,982 tokens and 377,945 types (*vocables*)
- With the first 12 most frequent types, we cover more than 30% of all Spanish documents (news articles)

Zipf's Law (Spanish Language)

Rank	Word	Freq.	Rel. Freq.	Cumul.	r x freq.
1	de	5,004,275	0.0695	0.0695	0.0695
2	la	2,876,708	0.0400	0.1095	0.0799
3	el	2,452,367	0.0341	0.1435	0.1022
4	que	2,171,101	0.0302	0.1737	0.1206
5	en	2,046,482	0.0284	0.2021	0.1421
6	y	1,613,223	0.0224	0.2245	0.1345
7	a	1,376,522	0.0191	0.2437	0.1338
8	los	1,228,087	0.0171	0.2607	0.1365
9	del	1,094,641	0.0152	0.2759	0.1368
10	por	809,824	0.0112	0.2872	0.1125

Zipf's Law

- On the other tail (the less frequent word types)
- Lot of word types with frequency = 1 (*hapax legomena*) and many with frequency = 2
- Number of word types: 7663 (Obama'08), 7792 (McCain'08)

Frequency	Obama'08		McCain'08	
1	2573	33.6%	2958	38.0%
2	1042	13.6%	1112	14.3%
3	556	7.3%	641	8.2%
4	446	5.8%	435	5.6%
5	308	4.0%	313	4.0%

Zipf’s Law

- The Zipf’s law predict 50% *hapax legomena*
- Why?
  - Spelling errors (performance & diacritics)
  - Many proper names
  - but this is a general pattern
    few word types cover a large number of tokens
    large number of word types cover a few number of tokens

Zipf's Law

- Example of *hapax legomena*

in McCain 2008	in Obama 2008
MI	AK
BMW	zionist
denial	WTO
bird	odd
richer	petrodollar
motel	Dupont
NALEO	Dehli

Vocabulary Growth

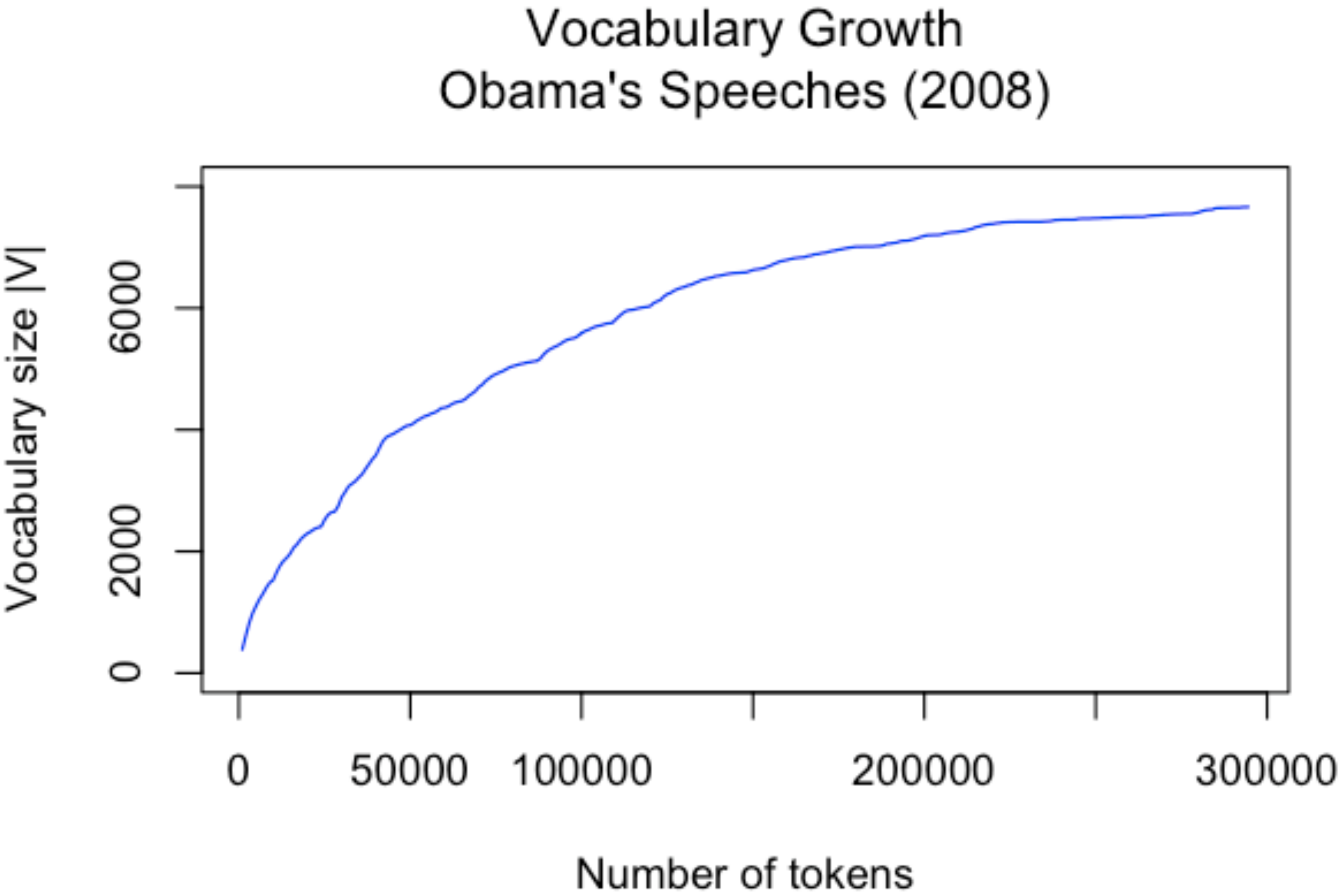
- Can we characterize the growth of an author’s vocabulary?
- After a progression phase (introducing new words), do we reach a plateau?
- Can we model the evolution of the number of *hapax*?
- Can we model the evolution of the vocabulary increase (by step of 1000 tokens)?

Vocabulary Growth

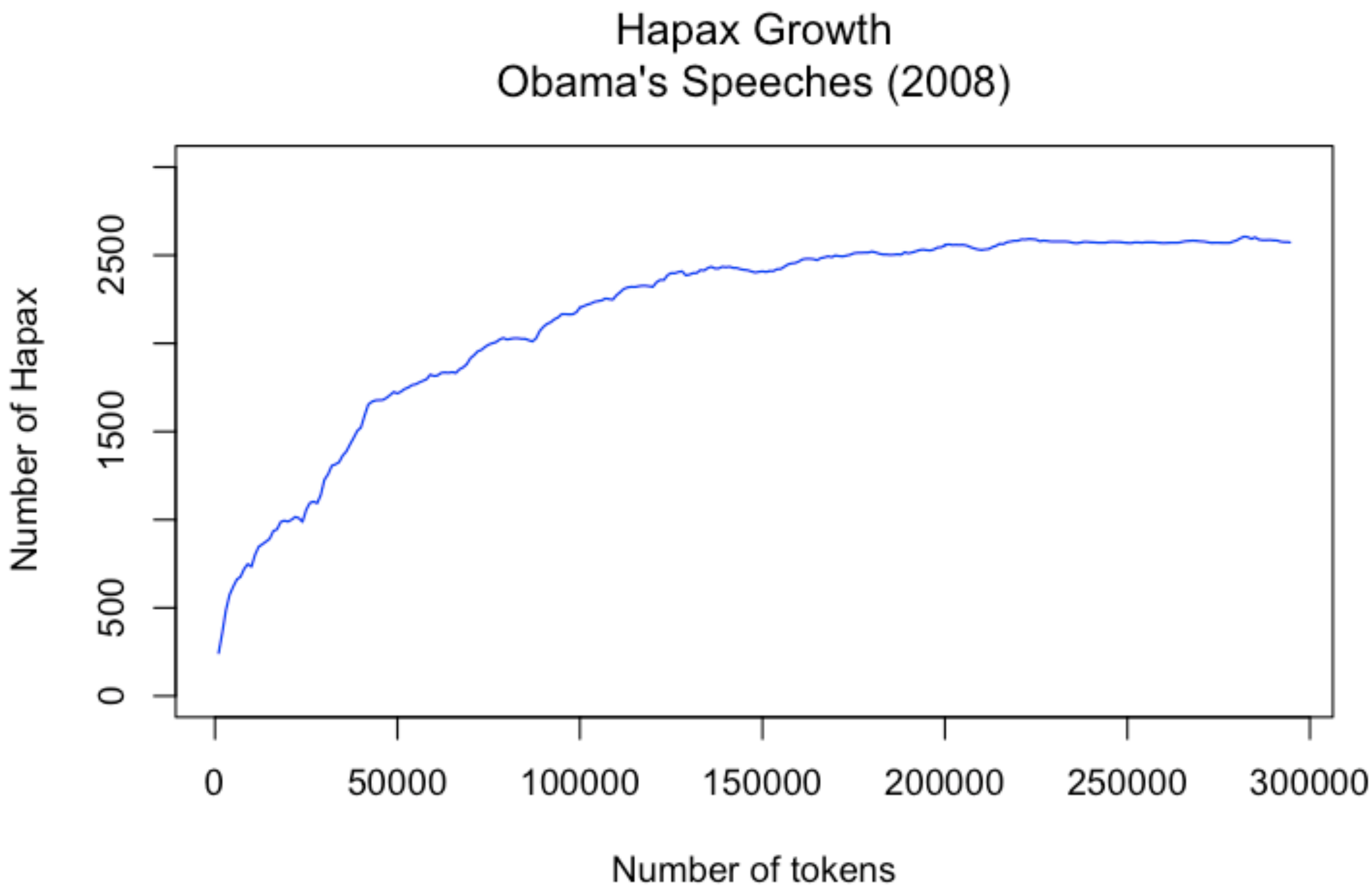
Obama's speeches (2008)

Tokens	| V |	Increase	Hapax
1,000	386	386	243
2,000	606	220	357
3,000	818	212	486
4,000	982	164	574
5,000	1,102	120	620
…	…	…	…
292,000	7,654	7	2,577
293,000	7,661	0	2,575
294,000	7,661	2	2,575

Vocabulary Growth



Hapax Evolution



Word Frequency

- Can we find useful features to help us finding the underlying characteristics of an author?
- We can find some differences between common American English (Brown corpus) and US electoral speeches by considering the top 10 / 20 most frequent word types
- Mainly on limited interest
- What are the differences between Obama's & McCain's speeches? Vocabulary? Topics? Style?

Rank	Brown		US	
1	the	6.90%	the	4.69%
2	be	3.86%	be	3.81%
3	of	3.59%	and	3.78%
4	and	2.85%	to	3.30%
5	to	2.58%	of	2.61%
6	a	2.28%	that	2.17%
7	in	2.06%	a	1.95%
8	**he**	1.92%	in	1.88%
9	have	1.23%	**we**	1.85%
10	it	1.08%	**I**	1.50%
11	**that**	1.05%	have	1.36%
12	for	0.89%	not	1.19%
13	not	0.87%	for	1.18%
14	I	0.83%	our	1.10%
15	they	0.82%	it	1.01%
16	with	0.72%	will	0.98%
17	on	0.61%	this	0.85%
18	**she**	0.60%	you	0.68%

Overall Lexical Measure

- We may consider forms used frequently by one author, less by the other
- Determinant “the” more frequent in ordinary language (6.9% vs. 4.7%)
- Used more frequently by politicians: “we”, “I”, “that”, “will”
- Used more often by common American English (Brown corpus): “he”, “she”
- Large variations when considering the same author but different periods, styles (e.g., tragedies, novels) and genres (prose vs. poetry)

Overall Lexical Measure

- In general, difficult to define an overall lexical measure and compare it with other authors/documents
- We can used:
  - |V| vocabulary size (number of word type)
  - ratio |V| / n
- not really satisfactory. Why?
  - depends on the sample size (not stable)
  - LNRE Large Number of Rare Events (many events do not occur in the sample!)

Conclusion

- Zipf's law (power law)
- Lexical distribution differs from the normal behavior (the Gaussian or Normal)
- LNRE distribution and phenomena more difficult to describe and analyze

Derivation from the Zipf's Law

- Starting with

$$f(\omega) = \frac{c}{z} \; or \; \frac{f(\omega)}{n} \cdot z = c'$$

where c is a constant, f(ω) the absolute frequency associated with word ω, n the total number of tokens, and z the rank

We may define by z_k the rank of word occurring k times in the corpus, we have:

$$z_k = \frac{c' \cdot n}{k}$$

Derivation from the Zipf's Law

- We can define I_k the difference between the rank z_k and the rank z_{k+1} with $z_{k+1} < z_k$

$$I_k = z_k - z_{k+1} = \frac{c' \cdot n}{k} - \frac{c' \cdot n}{k+1} = \frac{c' \cdot n}{k \cdot (k+1)}$$

$$I_1 = z_1 - z_2 = \frac{c' \cdot n}{2}$$

The rank difference between word occurring once and twice is 50% of all word types