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

What is a word?

- Select the word as unit of measurement
- What is a word?
  Trivial…
- Sequence of letters?
  "This painter is known in Paris"
- Sequences of letters and digits starting with a letter?
  "The first computer of the third generation was the IBM360 built in 1964"

What is a word?

- 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
- Uppercase vs. lowercase
  "The big clock" vs. "the big clock"
  "John with me" vs. "Me with John"
- All in uppercase
  "Stay with us" vs. "Stay with US"

5

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)

6

What is a word?

- Other possibilities
  lemma (entry in the dictionary, dogs -> dog),
  with grammatical categories (record/NN vs. record/VB)
- Other languages, other problems

我不是中国人

我	不	是	中国人
I	not	be	Chinese

7

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

8

Frequency

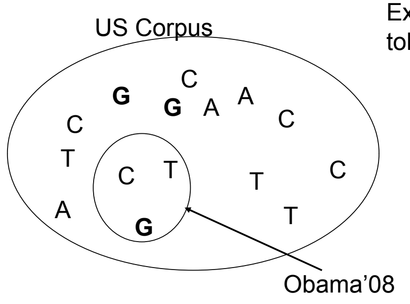
- Counting the word *types (vocable)* means counting the vocabulary size
  Denote by V the vocabulary
  E.g., V = {country, democracy, States, the, United}
  and its size is |V| = 5 (cardinality of a set)
- Counting the number of tokens (*forme*) means counting the sample / document / corpus size
  Use *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 infer pertinent information from word distribution?
- 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 electroal 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
  n = 420,410 tokens, |V| = 9,014 types
  For 2008 only: 113 speeches
  n = 294,553 tokens, |V| = 7,663 types
  http://www.barackobama.com/
- Speeches given by Senator John McCain
  94 speeches. from Apr., 25th 2007
  n = 206,899 tokens, |V| = 9,401 types
  For 2008 only: 71 speeches
  n = 154,365 tokens, |V| = 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

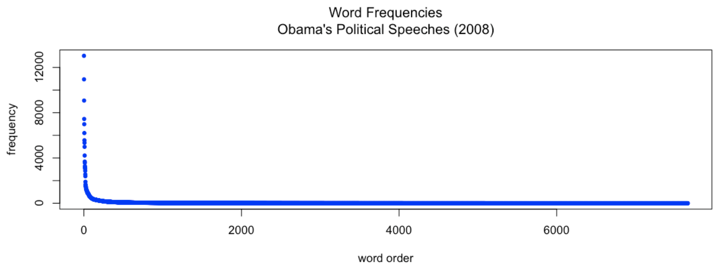
- More a regularity than a strict law
- The frequency (of a word type) (f(ω)) is related to the inverse of its rank (z) (with α = 1 for Zipf)
- We could use the absolute frequency (f(ω)) of the relative frequency (f(ω)/*n*)

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

- 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*)

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

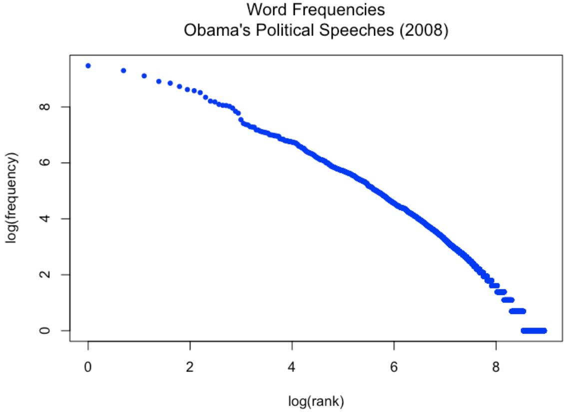
we may obtain

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

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

17

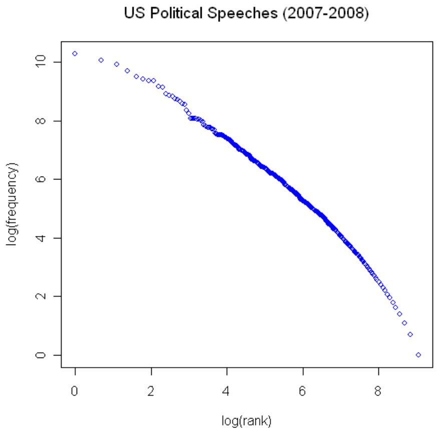
Zipf's Law



18

Zipf's Law

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



19

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)

20

Rank	Word	Freq. f(ω)	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

25

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%

26

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

27

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

28

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)?
- Can we model the growing of the vocabulary?

29

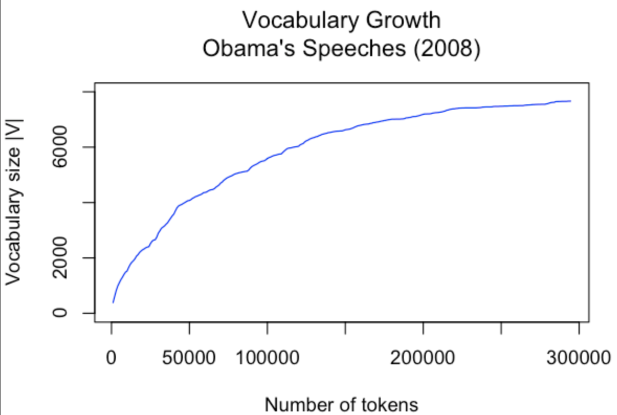
Vocabulary Growth

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

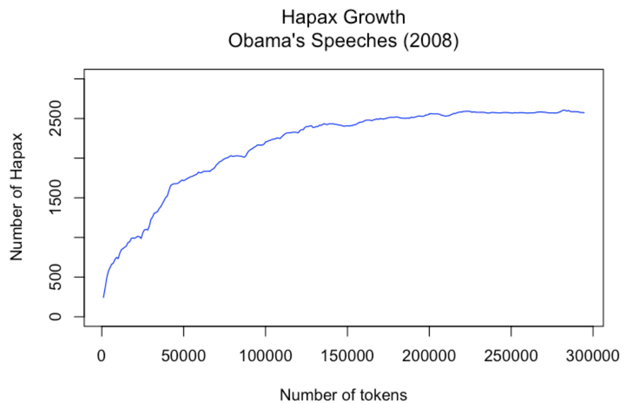
30

Vocabulary Growth



31

Hapax Evolution



32

Word Frequency

- Model of the growing of the vocabulary
  $|V| = k \cdot n^{\beta}$, with $10 \leq k \leq 20$, $0.5 \leq \beta \leq 0.6$
- 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?

33

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%

34

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)
- Basic elements for a language model
- Authorship attribution: Molière vs. Corneille

35

Language Model

- Objective (language model)
  Predicting the character / word sequence
- Probability of the sequence "h, o, r, s, e, s"
  And use a special symbol "Δ", beginning of a word
- Unigram: letter by letter, "h", "o", "r", …
- Bigram: "ho", "or", "rs", …
- Trigram: "hor", "ors", "rse", …
- Same for words Prob[*s* = *It was a bright cold day in April*]?
- Unigram
  Prob[*s*] = Prob[It| Δ] · Prob[was | It] · Prob[a | was] · … · Prob[April | in]

36

Language Model: Estimation

- Using the Maximum Likelihood Estimate (MLE) for bigrams ($C(w_k)$ = count / frequency of word w_k)

$$Prob_{MLE}[w_i|w_{i-1}] = \frac{C(w_{i-1},w_i)}{\sum_w C(w_{i-1},w)} = \frac{C(w_{i-1},w_i)}{C(w_{i-1})}$$

and for trigrams

$$Prob_{MLE}[w_i|w_{i-2},w_{i-1}] = \frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})}$$

37

Language Model: Estimation

- The model is trained on a part of the corpus: the training set.
- It is tested on a different part: the test set
- The vocabulary can be derived from the corpus, for instance the 20,000 most frequent words, or from a lexicon.
- It can be closed or open
- A closed vocabulary does not accept any new word
- An open vocabulary maps the new words, either in the training or test sets, to a specific symbol, <UKN>

38

Language Model: Example

- <s> *A good deal of the … way* </s>
- Unigram model with a corpus of 7,072 sentences.
- The word "A" occurs 2,482 times (and "good" 53 times, "deal" 5, "of" 3310 …).
- In this corpus, we found 115,212 tokens, 8,635 types (including 3,928 hapax legomena).
- Prob["A"] = 2,482/115,212 = 0.0215
  Prob["the"] = 6,248/115,212 = 0.0542
  and for the sentence
  Prob[*s*] = Prob["A"] . Prob["good"] . Prob["deal"] . … . Prob["way"] = $1.18 \cdot 10^{-48}$
- Which is the most probable sentence of three words?

39

Smoothing techniques

This is a black art in NLP.

40

Smoothing Technique

- Data sparseness is a serious and common problem in statistical NLP.
- The probability of a sequence is zero if it contains unseen elements (types, bigram)
- Problem 1: Low frequency *n*-grams
  if *n*-gram *x* occurs twice and *n*-gram *y* occurs once,
  is *x* really twice as likely as *y*?
- Problem 2: Zero counts
  If *n*-gram *y* does not occur in the training set, does that mean that it should have probability zero?

41

Laplace Smoothing

- Laplace smoothing

$$Prob[w_{i+1}|w_i] = \frac{C(w_i,w_{i+1}+1)}{\sum_w C(w_i,w)+1} = \frac{C(w_i,w_{i+1})+1}{C(w_{i-1})+|V|}$$

- Pro: Very simple technique
- Cons:
  - Too much probability mass is shifted towards unseen *n*-grams
  - Probability of frequent *n*-grams is underestimated
  - Probability of rare (or unseen) *n*-grams is overestimated
  - All unseen *n*-grams are smoothed in the same way
- Instead of adding 1 to all counts, add λ = 0.1 (Lidstone's rule)
- This gives much less probability to those extra events

42

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!)

43

View/Verify the Context

- Finding pertinent (significant) features is the first step
- Explaining such phenomena is the second step
- Usually it is important to see the context
  and again the computer science may help
- How?
  KWIC
  + Perl script to specify multiple constraints in selecting words / contexts / sentences

44

KWIC Keyword In Context

- Besides counting linguistic phenomena, computer science may provide other useful tools
- *KWIC* is such an example
- Provide the left and right context (number of words, number of characters) of a given word (exact spelling)
- Can be used to see the context around a term
- Example:
  Translation of “fort” (JJ) into the English language
  by “strong” or “powerful”
  “un fort orage”, “un café fort”, “un médicament fort”

Context around “Strong”

s pointed toward the December report as	strong	evidence of the long-awaited reversal in the nation's
5.8 billion Canadian dollars largely on	strong	foreign sales of forest products. *E* *S* However,
, and basically a black school that was	strong	in academics, "Dade said. *E* *S* "Before, we
finishing third in Iowa, maintened a	strong	lead in New Hampshire - but he no longer had the huge
etts Gov. Michael Dukakis maintened a	strong	lead in the Democratic race. *E* *S* ABC reported he
S* In both polls, Dukakis maintened a	strong	lead in the Democratic race. .End of Discourse *E* *
Er whose poll you're looking at - and a	strong	one, too, "said Jeff Alderman, chief of polling
Port on the seacost. *E* *S* Kemp, a	strong	proponent of states rights, has asked federal regu
rsuit of peace, NATO must soon offer a	strong	proposal on conventional and chemical weapons control
rsuit of peace, NATO must soon offer a	strong	proposal on conventional and chemical weapons control
ri Dubini Friday morning to "lodge a	strong	protest. *E* *S* "Defense Secretary Franl C. Carl
er Alexander Bessmertnykh read him a "	strong	protest. *E* *S* "The Soviet side cannot but view
the administration immediately lodged a	strong	protest with the Soviet ambassador here, saying the

Context around “Powerful”

ted. *E* *S* It also said two other "	powerful	bombs" were defused "in the last several days"
ederation of Economic Organizations, a	powerful	business alliance, is planning a leap into the 21s
itian army Col. Jean-Claude Paul, the	powerful	commander of the key batallion in Port-au-Prince,
. *E* *S* Despite the existence of two	powerful	drugs to treat the rare form of pneumonia, scienti
and simulated windsurfing in front of a	powerful	fan. *E* *S* Among the poeple wearing shorts were
nd West Germany, both with politically	powerful	farming lobbies, have sought an increase of $3.1 b
till was a land of barbarian tribes and	powerful	feudal warriors - one of Japan's last frontiers. *
out. *E* *S* "It's a vera silent but	powerful	force in Southern politics, "Rose said. *E* *S*
en. *E* *S* The reflex is particulary	powerful	in children, doctors say. *E* *S* Kendall was in
en. *E* *S* The reflex is particulary	powerful	in children, doctors say. *E* *S* Tecklenburg sai
ficient in the short-term, it provides	powerful	incentive for workers to sabotage innovative techno
eight straight term. *E* *S* With the	powerful	infrastructure of the governing Colorado Party at h
k was retained as head of South Korea's	powerful	intelligence agency, the Agency for National Secur
hn Moo-hyuk was retained as head of the	powerful	intelligence organization, the Agency for National

Strong vs. Powerful

- Are you drinking a “strong coffee” or a “powerful coffee”?
- Are you working with a “strong PC” or a “powerful PC”?
- Given the context, the translation could be “strong” or “powerful” (but the distinction is not always (for a computer at least) very clear, e.g., “strong/powerful drug”)

- Based on newspaper articles, we can find

Strong vs. Powerful

C(w)	C(strong w)	C(powerful w)	w
3418	4	13	force
933	0	10	computers
2337	0	8	computer
588	0	6	machines
2266	0	5	Germany
3745	0	5	nation
3685	50	0	support
3616	58	7	enough
3741	21	0	sales
1093	19	1	opposition
802	18	1	showing
2501	14	0	defense

t-test

- Compute the probability of having the word w^1 follows by the word w^2 in a given corpus
- We assume independance and estimate it as:
  Prob[w^1, w^2] = Prob[w^1] · Prob[w^2]
- As a second way, we simply count the number of observed bigrams in the corpus
- Example
  we have a corpus of 14,307,668 tokens
  we have 15,828 times the word type“new”
  we have 4,675 times the word type “companies”
  we have 8 times the bigram “new companies”
- Question: Does the bigram “new companies” form a collocation?

t-test

- Compute the probabilities
  Prob[new] = 15,828 / 14,307,668
  Prob[companies] = 4,675 / 14,307,668
  and assume independance (hypothesis H_0)
  Prob[w^1, w^2] = Prob[w^1] · Prob[w^2] = $3.615 \cdot 10^{-7}$
- Second model: the direct estimation
  Prob[new companies] = 8 / 14,307,668 = $5.591 \cdot 10^{-7}$
  and we can see this as a Bernoulli process with
  $\mu = p = 5.591 \cdot 10^{-7}$
  $\sigma^2 = p \cdot (1-p) \approx p = 5.591 \cdot 10^{-7}$ (because $(1-p) \approx 1$)
- Compare the models

t-test

- Example

$$t_{obs} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{5.591 \cdot 10^{-7} \cdot 3.615 \cdot 10^{-7}}{\sqrt{\frac{5.591 \cdot 10^{-7}}{14{,}307{,}668}}} = 0.999932$$

- In the table, with a significance level of α = 5% (dof = n-1 = ∞), we have t_{lim} = 2.576 (Normal table)
- The obsersed t_{obs} is lower that the t_{lim}
  Thus H_0 is not rejected.
  The words “new” and “companies” appear independently

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
- Language model use to predict word occurrence or bigram (trigram) of character / word
- Spelling error detection and correction
- Genre / authorship attribution

Derivation from the Zipf's Law

- Starting with

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

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:

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

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 $z_{k+1} < z_k$

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

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

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

Benford's Law

- Probability of occurrence of the most significant digit (1 to 9) given a sample of numbers
- Based on our prior knowledge (feeling), we may estimate that each digit owns the same chance to occur
  Uniform distribution for all digit = 1/9 = 0.111.
- This uniform distribution doesn not match real sample
- The distribution of the most significant digit follows the Benford's law
- The probability of occurrence of the digit "*d*" is
  Prob[*d*] = $\log_{10}$ [1 + (1/*d*)]

Benford's Law

- Estimations

d	prob	cumul. distribution
d = 1	0.30103	0.30103
d = 2	0.17609	0.47712
d = 3	0.12493	0.60206
d = 4	0.09691	0.69897
d = 5	0.07918	0.77815
d = 6	0.06694	0.84510
d = 7	0.05799	0.90309
d = 8	0.05115	0.95424
d = 9	0.04575	1.0