universität wien

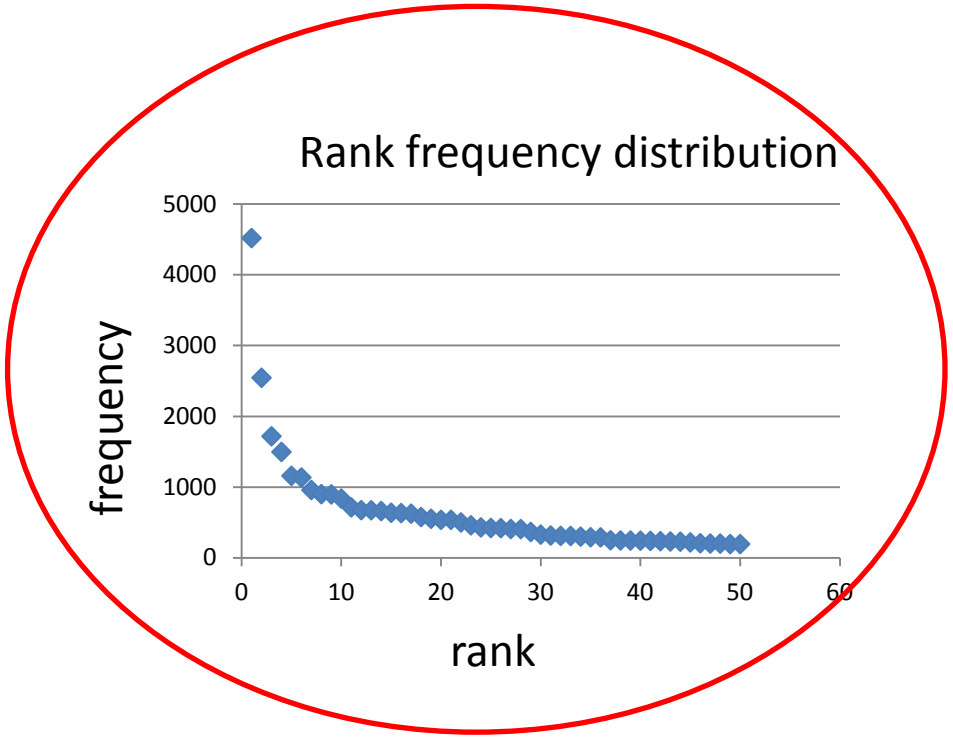
Dr. Emmerich Kelih

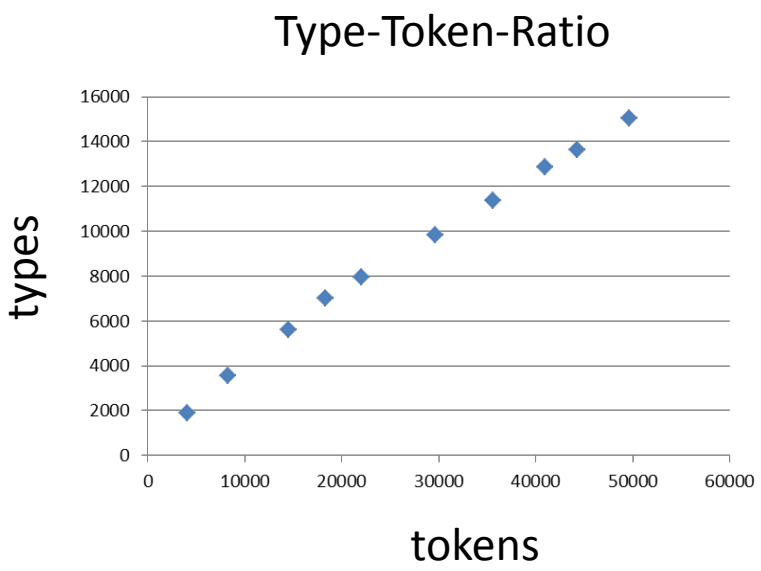
Institut für Slawistik

Lexical richness in text: new approaches to stylometrics

Todays talk:

1. Introduction to quantitative analysis of the **lexical structure of texts**





2. Relevance of the **Type-Token-Ratio** for crosslinguistic comparision and measurement of analytism/synthetism

Quantitative Analysis of the lexical structure

relevant for:

- Quantitative Linguistics: organisation of texts and behaviour of the lexical frequency
- Quantitative Text Analysis
- Natural language processing
- Computational linguistics
- Applied linguistics
- Psycholinguistics
- Lexicology (minimal lexicon, frequency dictionaries)
- **stylometrics**
- **authorship attribution**

→ frequency of word forms = vocabulary richness?

Different approaches to vocabulary richness:

1. Capturing the vocabulary richness by **means of a measure, of an index**
   = first step toward an quantification

2. Capturing the **unfolding of the vocabulary by a curve**, whose representants are e.g. Herdan,Tuldava, Köhler, Galle and several Russian scholars.

3. Starting with the **empirical distribution of words (types)** occurring x-times (tokens) and deriving the theoretical distribution based upon combinatorial considerations

4. Based **upon stochastic processes resulting in distributions.** (pure mathematical approach by Brainerd, Gani, Haight, McNeil, Simon, Baayen and others)

Wimmer, Gejza; Altmann, Gabriel: On vocabulary richness. In: *Journal of Quantitative Linguistics* 6 (1), S. 1–9.

What are the ingredients of vocabulary richness (VR) ?

1. values of V (vocabulary) and N (text length)
2. the course of the increase of the vocabulary: (sequential analysis)
3. the whole distribution of word forms

Methodological aspects:

1. descriptive goals: VR of one text, one author etc.
2. comparison of VR in different texts, different authors
3. application of statistical methods

Problems of the analysis of VR

1. The choice of texts which is, as a matter of fact, irrelevant if we are interested in theory: the methods, statements, etc. must be applicable to and hold for all texts.

2. The **definition of the word** is associated with the problem of **lemmatization**. What should be done with conversions, portmanteau morphemes, liaison, sandhi reflexive verbs, with detachable affixes etc. ?

3. What are the criteria for determing the word/word form? orthographical, phonological, morphological morphosyntactical identification of words?

4. Considerable problems arise with the **sampling of the text**: should we take the whole text, only a part of the text, should we mix several smaller parts or take a random sample of words from the whole text?
5. Are we interested only in the classification/comparision of texts only?

6. Should we examine textual and extratextual properties of vocabulary richness too?

7. What does it mean that a text has a **rich vocabulary**?

Starting point: Zipf‘s law

- named after G.K. Zipf (1902-1950)
- author of *Psychobiology of Language* (1935) and
Human Behavior and the Principle of Least Effort (1949)

- Zipf‘s law is a umbrella term for several forms of “power laws”

- different mathematical formulations (continuous functions and discrete probability distributions)

- applied in linguistics and many other scientific branches (biology, sociology, economics etc.

- fundamental behaviour of nonlinear dynamic systems

- distribution curve of lexical frequencies can be captured by mathematical models

What is the linguistic background of Zipf‘s law?

Frequency of word forms

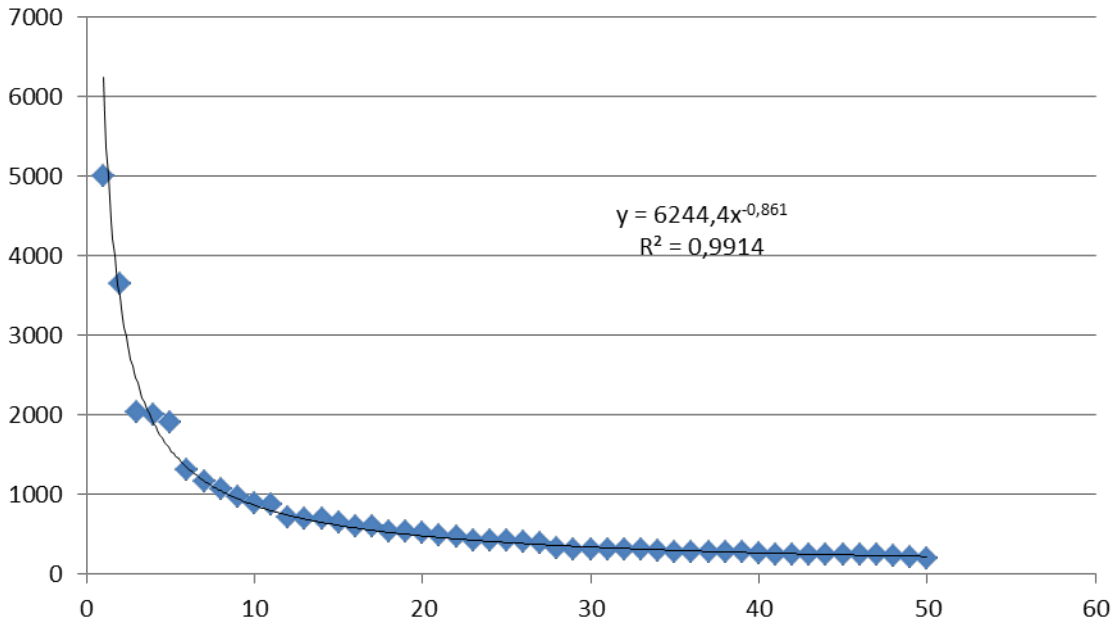
1. Text (running text)
2. Lemmatization (yes/**no** ?)
3. Determination of the number of word forms (tokens)
4. Frequency of word forms (types)
5. Compilation of a rank frequency distribution

Однажды весною, **в** час небывало жаркого заката, **в** Москве, **на** Патриарших прудах, появились два гражданина. Первый из них, одетый **в** летнюю серенькую пару, **был** маленького роста, упитан, лыс, свою приличную шляпу пирожком нес **в** руке, а **на** хорошо выбритом лице его помещались сверхъестественных размеров очки в черной роговой оправе. Второй — плечистый, рыжеватый, вихрастый молодой человек **в** заломленной **на** затылок клетчатой кепке — **был** в ковбойке, жеваных белых брюках и **в** черных тапочках.

Russian text: Master i Margarita

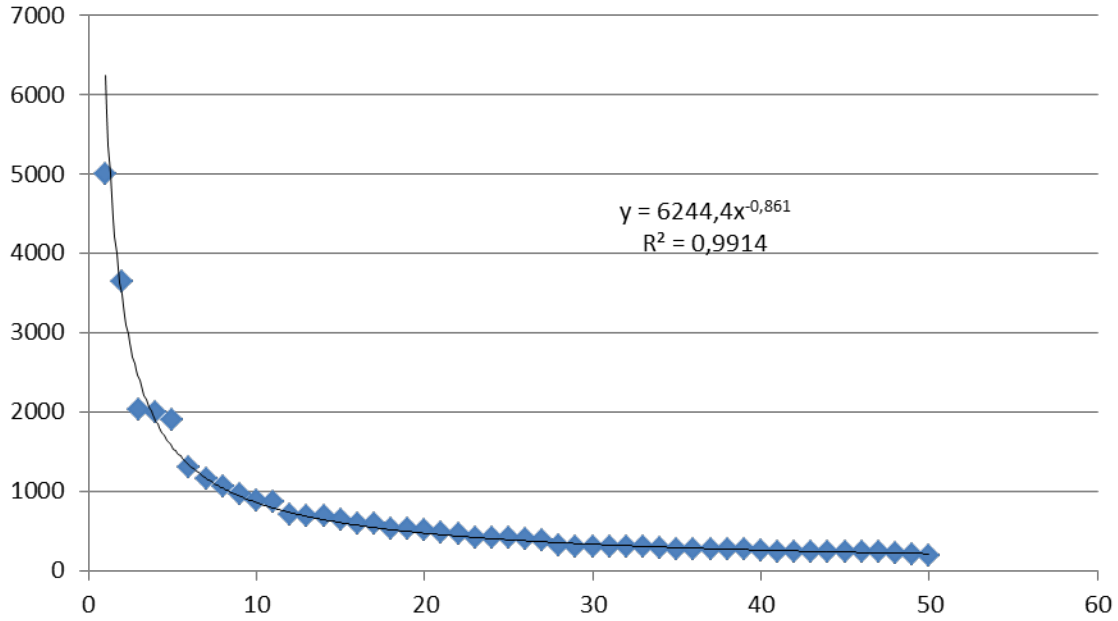
Rang	Word	Freq.	Rang	Word	Freq.
1	И	5006	16	ИЗ	602
2	В	3640	17	ЖЕ	594
3	НЕ	2025	18	ПО	534
4	НА	2003	19	У	527
5	ЧТО	1905	20	ЗА	515
6	С	1307	21	ВСЕ	477
7	ОН	1153	22	БЫЛО	465
8	ТО	1071	23	МАРГАРИТА	419
9	А	974	24	ТАК	414
10	КАК	883	25	ВЫ	411
11	Я	867	26	ОНА	396
12	НО	713	27	ОТ	382
13	К	699	28	О	326
14	ЕГО	688	29	ТУТ	308
15	ЭТО	647	30	ДА	306

Graphical representation: rank frequency distribution



- systematic interrelation of frequent and less frequent word forms
- few word forms have a high frequency
- form of the (nonlinear) curve and it's interrelation to entropy and repeat rate of words forms

Graphical representation: rank frequency distribution



Form of the rank frequency distribution depends on

- author
- text type
- functional style
- language

+ rank frequency distribution are relevant for any linguistic entity
+ analysis of rank frequency distributions is a basic task of QL
+ QL can be understood as a part of general system analysis (complex, nonlinear systems)

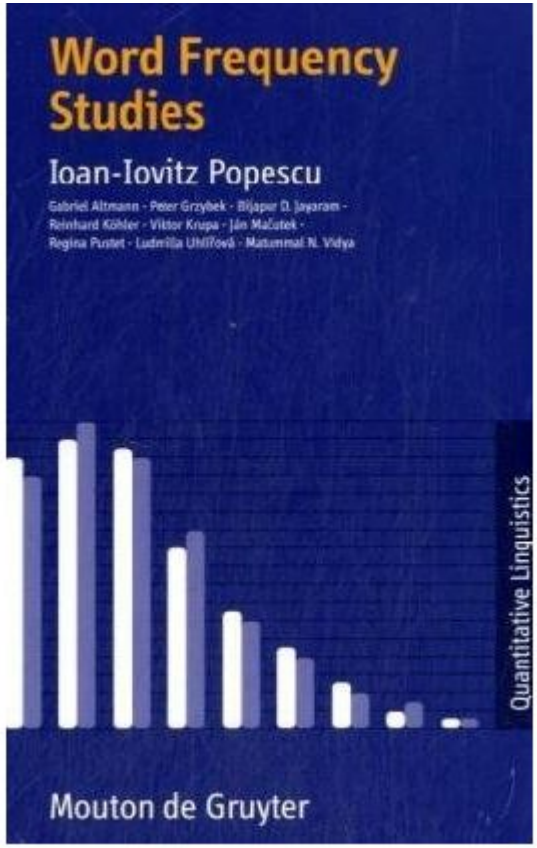
Zipf‘s law and the lexical organization of texts: The State of the Art

- many mathematical modifications of Zipf‘s law are available - not all of the them are motivated by linguistic considerations
- no systematic empirical analysis of the behaviour of Zipf‘s parameter in different texts and languages
- linguistics claims about Zipf‘s law are in fact very general
- same law is valid for organization of the population structure of cities, citations of internet pages, scientometric and bibliometrics issues

<u>But,</u>

- Zipf’ law is determining the lexical structure of texts
- it has an impact for lexicological and lexicographical objectives and task
- is relevant for the study of collocations (corpus linguistics)
  e.g. impact on the specific organization of the combination of frequent and less frequent word forms
- shape of the rank frequencies contains information about the “lexical richness” of texts
- lexical frequency interrelates with many other linguistic entities and properties

Word frequency studies: State of the art



Ioan-Iovitz Popescu (2009)

in cooperation with
Gabriel Altmann
Peter Grzybek
Bijapur D. Jayaram
Reinhard Köhler
Viktor Krupa
Ján Mačutek
Regina Pustet
Ludmila Uhlířová
Matummal N. Vidya

- analysis of many different languages
- many new methodological approaches
- emphasis on the linguistic information of the rank frequency distribution

Contents

- **Problems and Presentations**
- **The h-point and its Relatives**
- The geometry of word frequencies
- The dynamics of word classes
- Thematic concentration of the text
- Crowding, pace filling and compactness
- Autosemantic text structure
- Distribution models
- The relation of frequency to other word properties
- Word frequency and position in sentence
- **The type-token relation**

Word **frequency** as a very simple property laying on the surface of a text?

- results of mechanical word counting can be used for practical purposes
  (typography, psychology, language teaching, cryptography, software)

- problems of frequency in linguistics are manifold (quantitative linguistics, usage-based approach, corpus linguistics, language learning)

- basic problems of the research unit

- basic difference of word forms and lemmas

- fuzziness of the term word form and lemma

- which of these units one should take? Word form or Lemmas?

→it depends on the hypothesis, one examines!

→ criteria for determining words forms and lemmas must be discussed!

→ usually some common definitions of word forms in written texts are used
  (orthographical criteria)

I. Counting word forms or lemmas?

Both can be in turn shaped using different criteria, e.g. one can consider clitics as parts of word-forms or not, i.e. do we consider grammatical or phonetic word-forms? In Slovak, a syllabic preposition takes the accent of the word and can be considered as a component of the phonetic word. In Indonesian, the interrogative particle kah can be added to almost any word, in Japanese writing one cannot always see whether the same particle ka is part of the word or not. In some Slavic languages one writes the reflexive pronoun together with the word, in other ones separately.

Lemmas can be set up in narrow conception e.g. you, they and your, their are four different lemmas but in a broad conception there are only two, or even only one since person is a grammatical category.

→ Set up the criteria of your counts!

→ for the analysis of vocabulary richness, one should favour lemmas

II. Different views on word frequency:

The second distinction is that between a **rank frequency distribution, cumulative frequencies** and corresponding **frequency spectrum**.

Words are ordered according to their decreasing frequency, i.e. the variable x is the rank, and f (x) is the frequency.

1. Plain ranking of word frequencies

(II)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	rank x
	15	13	8	7	6	5	5	4	3	3	2	2	2	2	1	1	1	1	1	1	$f(x)$

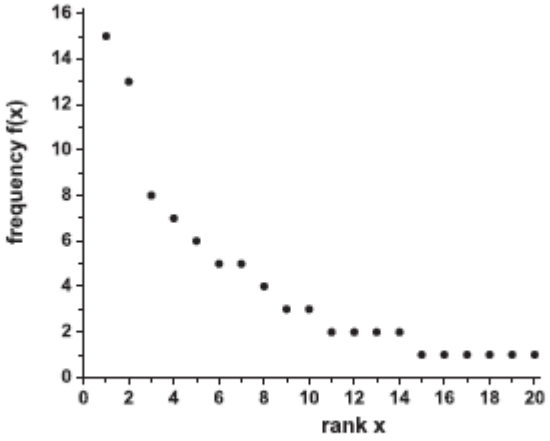


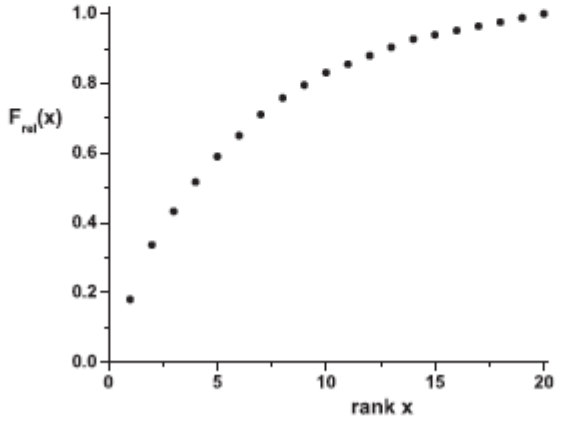
Figure 2.2: Plain ranking of frequencies

2. Cumulative ranking of word frequencies

(III)

x	$F(x)$	$F_{rel}(x)$	x	$F(x)$	$F_{rel}(x)$
1	15	0.1807	11	71	0.8554
2	28	0.3373	12	73	0.8795
3	36	0.4337	13	75	0.9036
4	43	0.5181	14	77	0.9277
5	49	0.5904	15	78	0.9398
6	54	0.6506	16	79	0.9518
7	59	0.7108	17	80	0.9639
8	63	0.7590	18	81	0.9759
9	66	0.7952	19	82	0.9880
10	69	0.8313	20	83	1.0

Here one usually uses the relative cumulative frequencies, i.e. all numbers divided by their sum, e.g. $F(1) = 15/83 = 0.1807$; $F(2) = 28/83 = 0.3373$ etc. The highest rank r_{max} (here 20) is the vocabulary V of the text, the greatest $F(r_{max})$ (here 83) is text length N. In Figure 2.3 one can see the cumulative ranking of relative frequencies (II) which is slightly smoother than the plain ranking.



3. Reversed cumulative ranking of word frequencies

(IV)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	rank
	1	1	1	1	1	1	2	2	2	2	3	3	4	5	5	6	7	8	13	15	reversed $f(x)$

then the frequencies will be cumulated

(V)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	reversed rank
	1	2	3	4	5	6	8	10	12	14	17	20	24	29	34	40	47	55	68	83	cumulative $f(x)$

In a last step all ranks will be divided by the highest rank (here 20) and the cumulative frequencies will be divided by the sum of frequencies (practically by the last frequency, here $F(20) = 83$). Thus we obtain finally

(VI)	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
	0.011	0.024	0.036	0.048	0.060	0.072	0.096	0.120	0.145	0.169
	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
	0.205	0.241	0.289	0.349	0.401	0.482	0.566	0.663	0.819	1.000

The graphical presentation of (VI) can be seen in Figure 2.4.

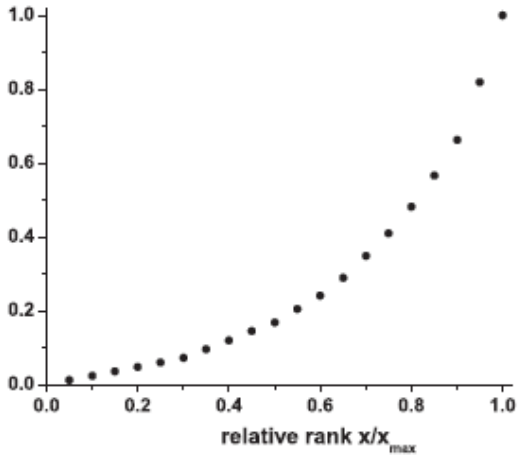
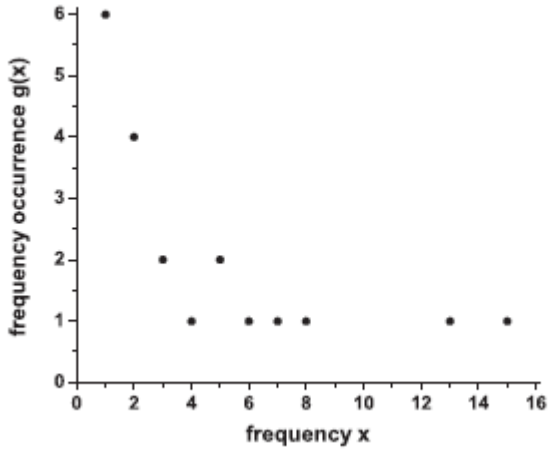


Figure 2.4: Reversed cumulative relative frequency with relative ranks

4. Frequency spectrum

In the latter, g(x) is the number of words occurring exactly x-times. We shall differentiate f (x) and g(x) on practical grounds.

(VII)	1	2	3	4	5	6	7	8	13	15	x = frequency
	6	4	2	1	2	1	1	1	1	1	$g(x)$ = number of words with $f(x)$



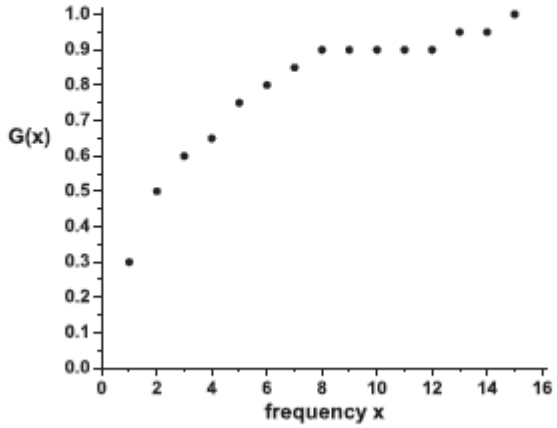
5. cumulative frequency spectrum

The cumulative frequency spectrum $G(x)$ is obtained by simply adding up the frequencies in (VII) without skipping the missing values of the variable, yielding

(VIII)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	x
	6	10	12	13	15	16	17	18	18	18	18	18	19	19	20	$G(x)$

The cumulative frequencies can be transformed in relative values – dividing by the $G(x_{max})$, here 20 – yielding the cumulative relative frequency spectrum for which we do not reserve a special symbol, call it $G(x)$ but the presentation clearly shows whether the absolute or relative frequencies are meant. In our case the cumulative relative frequencies will be

The graphical presentation is shown in Figure 2.6. This representation is used for capturing the text coverage. As can be seen, this presentation is not so "smooth" as that in Figure 2.3.



6. Relativized forms of the rank frequency distributions:

In addition one can present both variables in relativized form, namely as $< x/x_{max}, f(x)/f(1) >$ for ranks or $< x/x_{max}, g(x)/g(1) >$ for the spectrum, yielding an easy optical comparison between texts. In Figure 2.7 the rank frequency distribution is presented in this form. Hence one could extend Table 2.1 (p. 16) into a third dimension in which all relativized presentations would be placed.

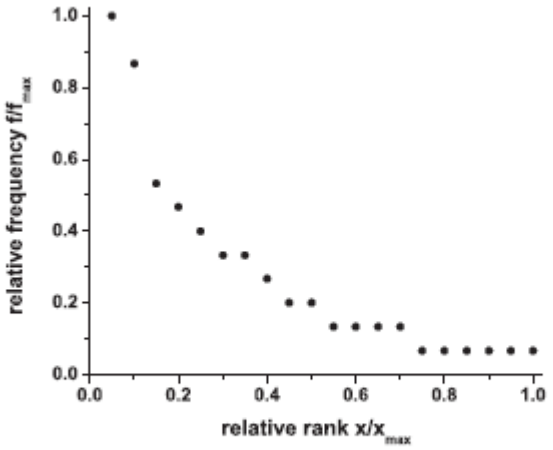


Figure 2.7: Both variables in relativized form (rank-frequency)

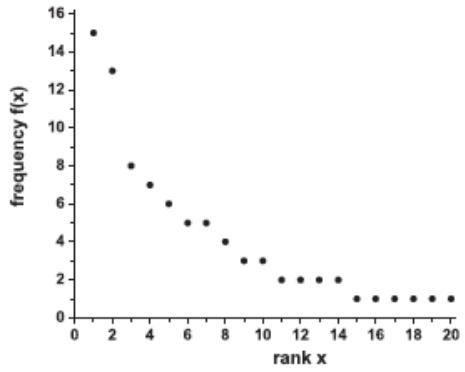
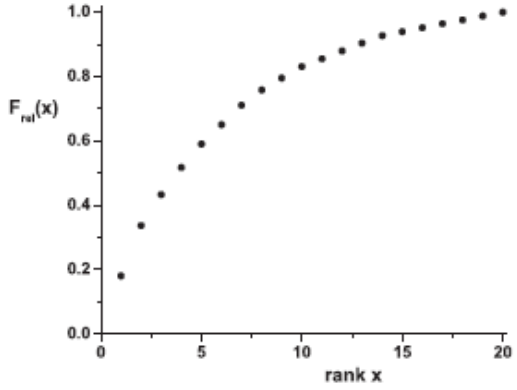
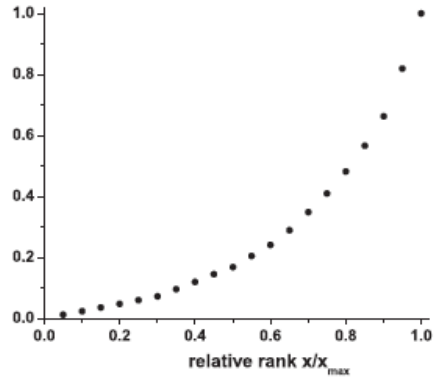
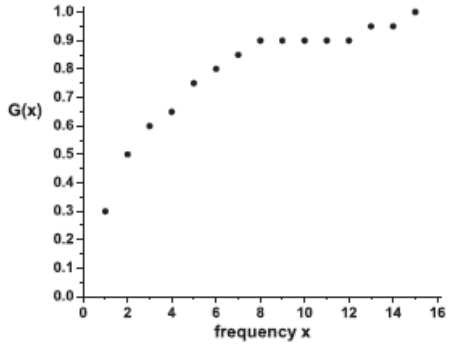


Figure 2.2: Plain ranking of frequencies



2.3: Cumulative ranking of relative frequencies ($F(x)$)





ure 2.6: Cumulative relative frequencies of the spectrum ($G(x)$)

- Each of the representations has its own tasks and characterizes an aspect of the word frequency distribution
- Not all of them are used in linguistics with the same intensity!
- Plain ranking and cumulative ranking are the most favourite ones
- all of the possibilities require different mathematical models

Plain ranking frequencies: The h-point and related points:

- Such distributions have always a monotonously decreasing hyperbolic form
- they are not always "smooth" in the sense that the frequencies are not positioned exactly on a theoretical curve.
- Here we can state that with increasing rank *r and decreasing frequency f (r) there is a* point at which *r = f (r).*
- This point is called *h-point and its distance to the* origin [0,0]

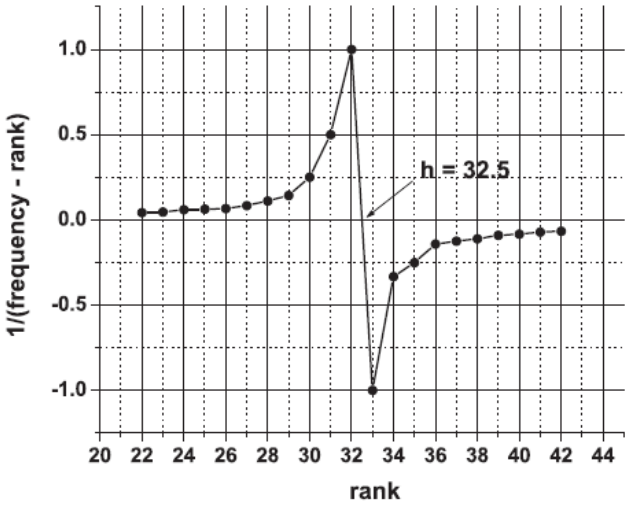
$$d = h\sqrt{2}.$$

$$C = \frac{1}{f(r) - r} = \frac{1}{frequency - rank}.$$

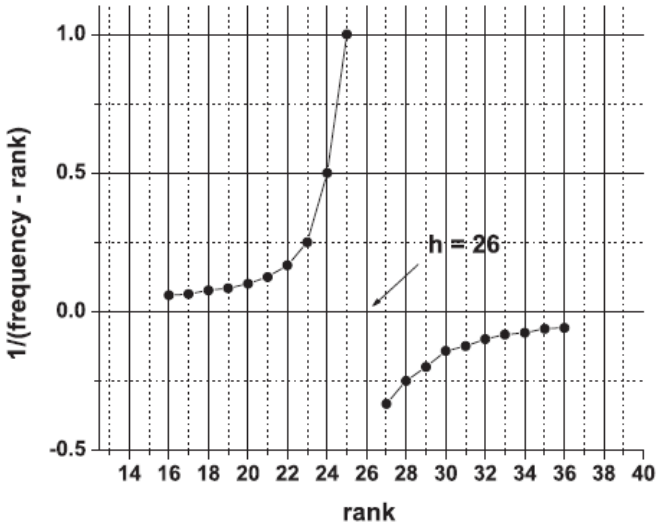
- Obviously, the *h-point of actual discrete* distributions is closely related to the mathematical attractive fixed point of continuous functions
- h-point introduced into scientometrics (Hirsch 2005)
- used and explored in linguistics mainly by G. Altmann and I.-I. Popescu

The h-point and related points:

$$C = \frac{1}{f(r) - r} = \frac{1}{frequency - rank}.$$



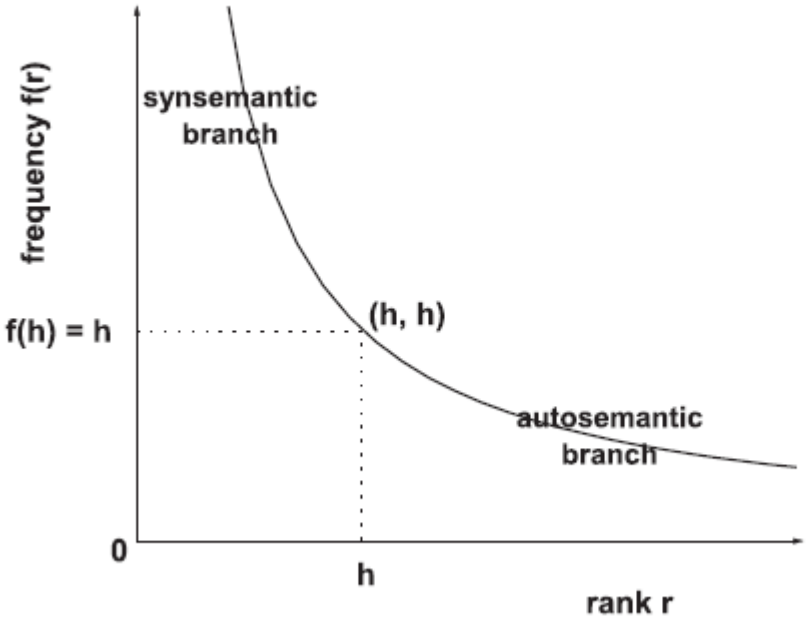
(a) Banting Nobel lecture



(b) Bellow Nobel lecture

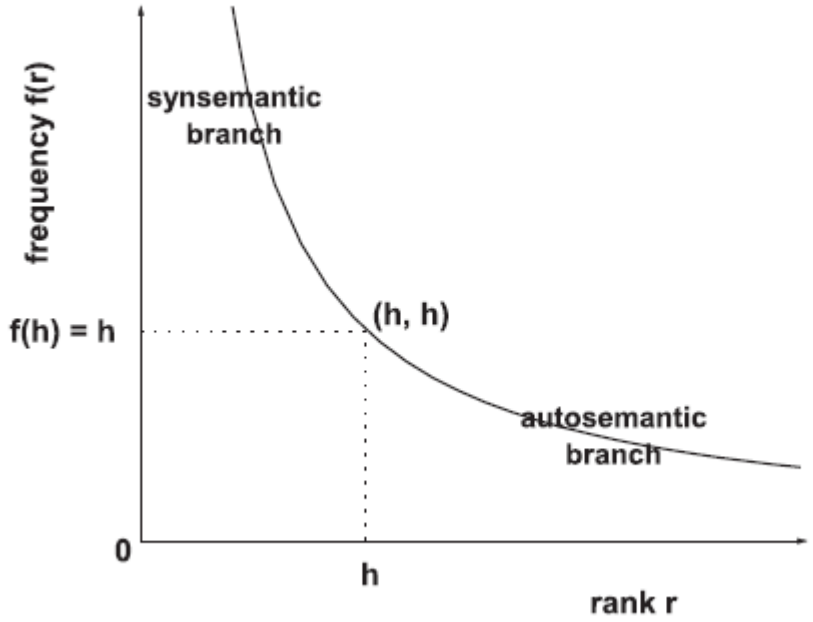
Figure 3.2: h-point determination

What is the linguistic meaning of the h-point?



- h-point has some properties which make it useful for text analysis
- auxiliaries, synsemantics etc. are usually more frequent than autosemantic
- *h-point divides the vocabulary* in **two parts**, namely in a class of magnitude *h of* ***frequent synsemantics or*** **auxiliaries** (prepositions, conjunctions, pronouns, articles, particles, etc.) and a much greater class (*V −h) of* ***autosemantics*** *which are not so frequent but* build the very vocabulary of the text.
- **rapid branch of synsemantics** and **slow branch of autosemantics**

fuzziness of *h* as a separating point?

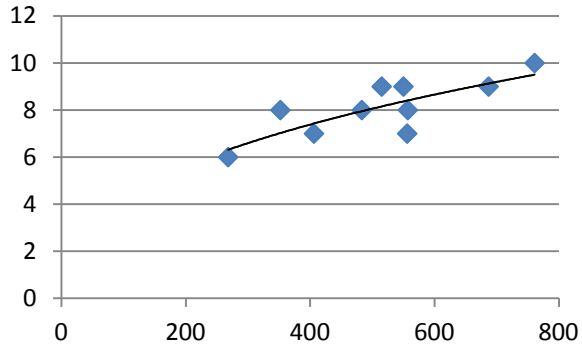


→ Of course, this separation is not clear-cut, sometimes there are autosemantics in the "rapid branch" and synsemantics in the "slow branch".
= autosemantics before the h-point and synsemantics after the h-point

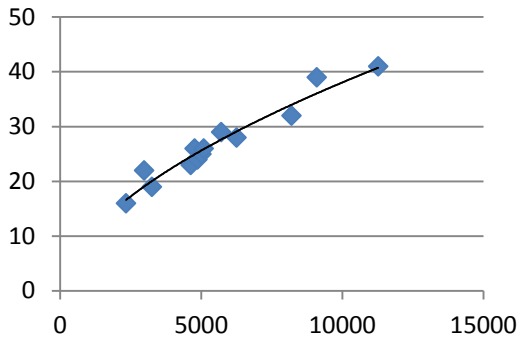
→This behaviour can be used for the analysis of the thematic concentration of texts!

Dependency of the h-point on text length: Analysis of different texts within one language:

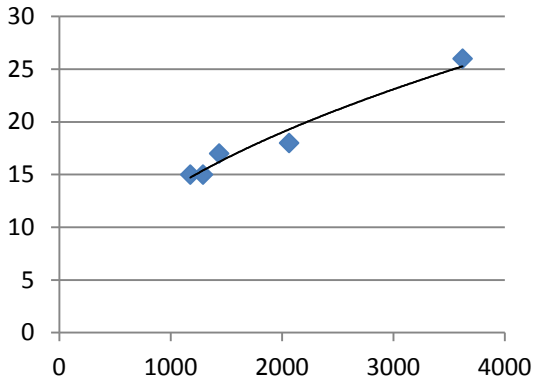
Relation of N and h-point:



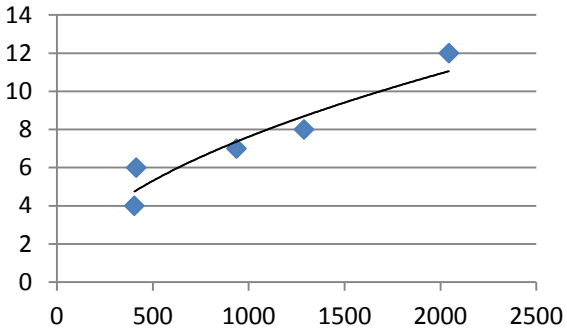
Bulgarian



English



Maori



Hungarian

h-point and text lengths

- Useless for textlogical studies ?
- Useless for cross linguistics comparision ?

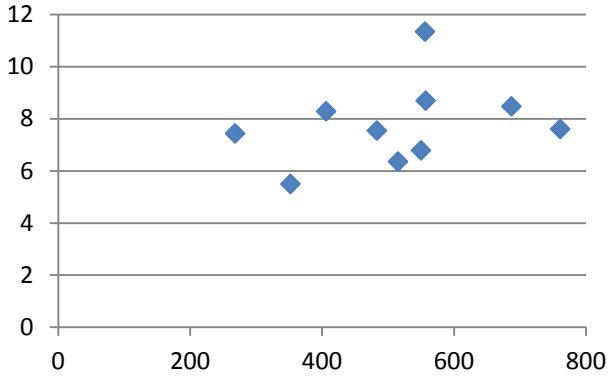
Hirsch (2005) showed that there is a relationship between the h-point and text length N, represented by the total area below the rank frequency curve, namley

$$N = ah^2 .$$

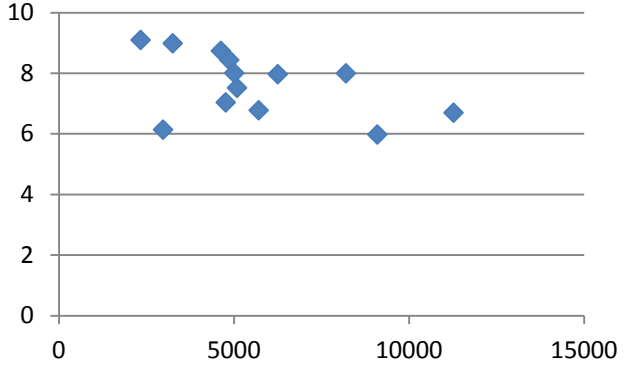
Usually the dependence of a textological index on text length is very detrimental to any further discussion but in this case the parameter *a* shows the partitioning of the text in parts whose size is adapted to the text length. Thus we get the index

$$a = \frac{N}{h^2}$$

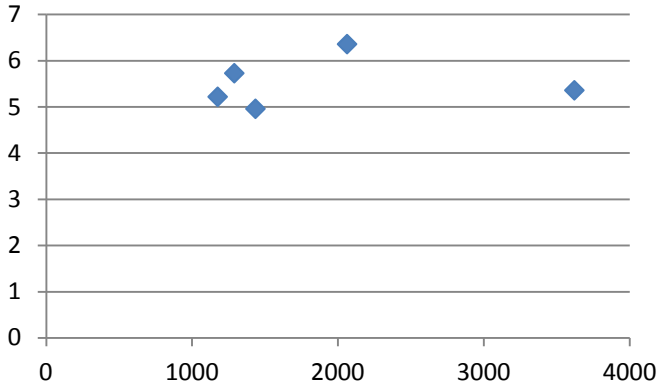
Dependency of text length and parameter *a* ?



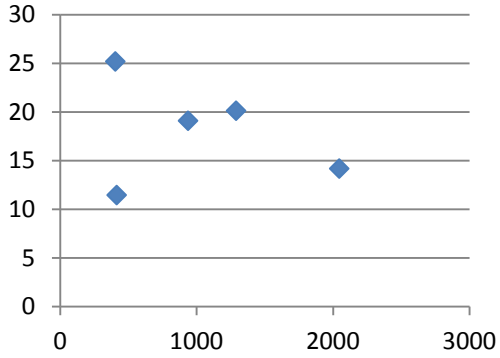
Bulgarian



English



Maori



Hungarian

Linguistic meaning of *h* and *a*:

- Greater *h* (smaller *a*) is a sign of analytism, i.e. the number of word forms is smaller, the synthetic elements are replaced by synsemantics.
- *h* and *a* are at the same time both characteristics of a text (within the given language)
- *h* and *a* signs of analytism/synthetism in cross-linguistic comparison.
- Using the index ***a*** we **get rid of the dependence on *N***.

- Further advantage: index *a* can be used for statistical testing of the significance of differences

Table 3.2: Mean values of quantity *a* in 20 languages

Language	Mean *a*	Language	Mean *a*
Samoan	4.56	Italian	8.41
Rarotongan	5.02	Romanian	9.15
Hawaiian	5.37	Slovenian	9.19
Maori	5.53	Indonesian	9.58
Lakota	5.69	Russian	10.10
Marquesan	5.69	Czech	10.33
Tagalog	7.24	Marathi	11.82
English	7.65	Kannada	16.58
Bulgarian	7.81	Hungarian	18.02
German	8.39	Latin	19.56

Is there a significance difference between Tagalog and Indonesian?

$$t = \frac{|\bar{a}_1 - \bar{a}_2|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3.4)$$

where

$$s^2 = \frac{\sum_{i=1}^{n_1}(a_{i1} - \bar{a}_1)^2 + \sum_{i=1}^{n_1}(a_{i2} - \bar{a}_2)^2}{n_1 + n_2 - 2}$$

Tagalog $(7.91 - 7.24)^2 + (8.12 - 7.24)^2 + (5.69 - 7.24)^2 = 3.6258; n = 3$

Indonesian $(10.44 - 9.27)^2 + (7.61 - 9.27)^2 + (9.64 - 9.27)^2 + +(13.72 - 9.27)^2 + (6.47 - 9.27)^2 = 31.9039; n = 5$

hence $s^2 = (3.6258 + 31.9039)/(5 + 3 - 2) = 5.9216$, and $s = \sqrt{5.9216} = 2.4334$. Inserting in (3.4) we obtain

$$t = \frac{|7.24 - 9.27|}{2.4334\sqrt{\frac{1}{3} + \frac{1}{5}}} = 1.14\,. \qquad (3.5)$$

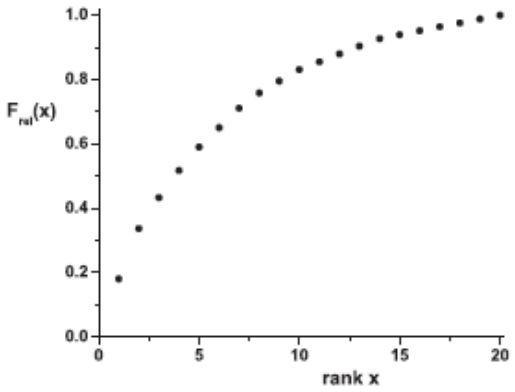
Since for a two-sided test $t_{0.05}(6) = 2.45$, the difference is not significant.

→ can be used for cross linguistic comparison (mean values)

→ for comparison of texts other test must be used! (see Popescu 2009)

A first look at vocabulary richness:

h point is seperating two different areas of the rank frequency distribution:
synsemantic and **autosemantic** branch



2.3: Cumulative ranking of relative frequencies ($F(x)$)

- cumulative relative frequency up to the *h*-point, i.e. *F*(*h*), represents the *h*-coverage of the text.
- F(h) is the coverage by auxiliaries
- Fuzziness of the border – empirical correction

$$\underline{F(h)} = F(h) - \frac{h^2}{2N} .$$

Now, the full area of the distribution from which *F*(*h*) is subtracted, i.e.
1–*F*(*h*) = one aspect of the vocabulary richness of the texts.

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right) .$$

→ NO dependency of text length!

→ an acceptable coefficient of vocabulary richness

→ Again: tests of significance are possible

→ Many particular investigations within one language are necessary in order to find all factors contributing to this kind of richness measurement

Since only $F(h)$ is a variable representing a proportion (the rest are constants), it is easy to set up an asymptotic test. Since $Var(R1) = F(h)[1 - F(h)]/N$, the difference between two texts can be tested using the familiar normal variable yielding

$$z = \frac{R_1 - R_2}{\sqrt{Var(R_1) + Var(R_2)}} . \qquad (3.9)$$

For the sake of ease, the proportion $F(h)$ was included in the table. Consider e.g. the difference between Latin *Lt-04* and Tagalog *T-03* in the last row of Table 3.6. Using the numbers in the table we can directly insert the values in formula (3.9) and obtain

$$z = \frac{0.8469 - 0.6132}{\sqrt{\frac{0.1998(1-0.1998)}{4285} + \frac{0.4747(1-0.4747)}{2054}}} = 18.55$$

a value which is highly significant. Of course, these asymptotic tests strongly depend on N but at least a preliminary classification of texts is possible.

Open questions:

- Do synsematnics and auxiliaries really contribute to vocabulary richness?
- auxilaries are present in all texts
- specific importance of autosemantics before the h-point?
- freqency of autosemantics before h-point: thematic concentration
- what about counting hapax legomena?
- why words occuring twice (dislegomena)= or more times do not contribute to vocabulary richness?

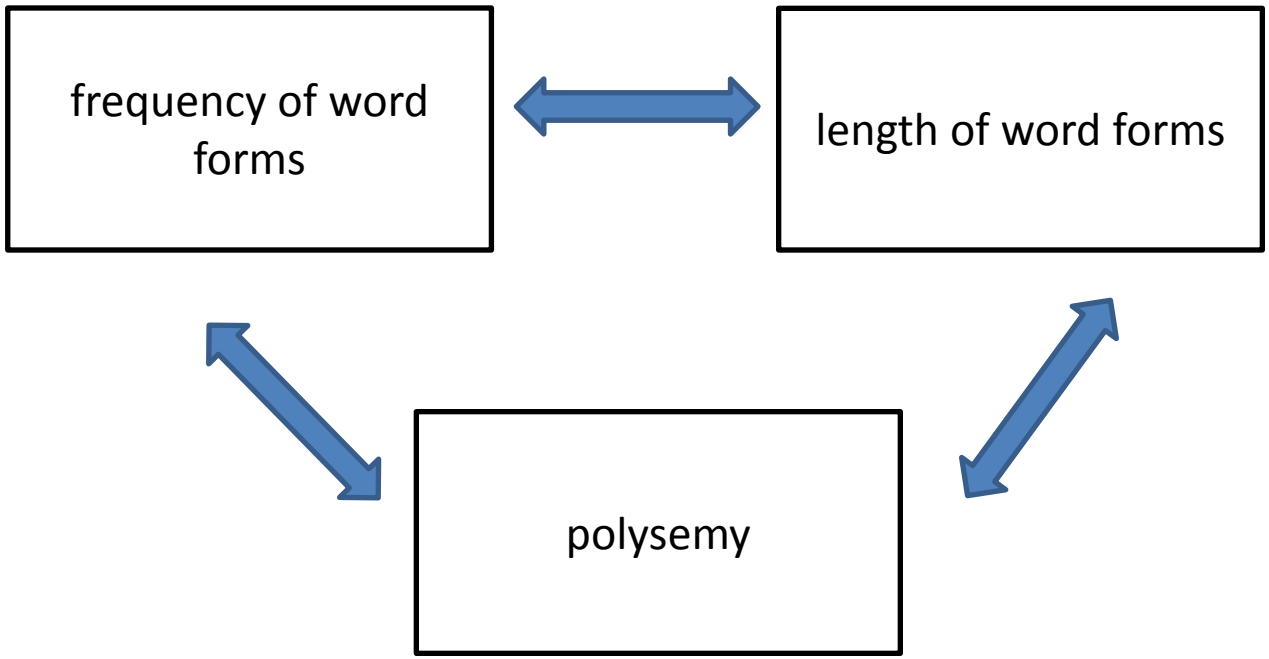
→ again dependency on N

notorious problem of stylometrics:

1. If a characteristics depends on N, then one can find a theoretical function which corresponds to this dependence
2. If an index depend on N, then two indices are not directly commensurable
3. Problem of text length and application “classical” statistical tests
4. What is determined by the language/text type and what is the creativity of a author?

Nevertheless the study of word frequencies is not a trivial task:

1. frequency of word forms – rank of word forms
2. frequency of word forms – length of word forms
3. frequency of word forms – degree of polysemy (in texts)



--> exploration of mutual interrelations and dependencies !

Intermediate conclusion:

- Word frequency studies are not a trivial task
- QL is going beyond the modelling problem (Zipf‘s law)
- *h* point and *a* as intersubjective determinable point of distribution curves
- h-point is linguistically interpretable and usefull for many descriptive purposes
- h-point is relevant for text studies and cross linguistic studies
- more systematic studies of the h-point are required
- open question of interrelation of h-point with other properties (hapax legomena, steepness of the curve, parameter of Zipf‘s law … …)