Glottometrics 13, 2006, 23-46

Some aspects of word frequencies

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the present article some new aspects of word frequency distributions are presented, namely the h-point, the k-point, the m-point, the n-point, the Lorenz curve and Gini's coefficient, all of which characterise some properties of texts. Their relationship to the measurement of vocabulary richness is scrutinized. It is an attempt at transferring some views from other domains of science to linguistics.

Keywords: word frequency, vocabulary richness, h-point, k-point,m-point, n-point, Lorenz curve, Gini's coefficient

1. Introduction

Calculation and presentation of word frequencies has many aspects, of which only a few have so far been taken into account: plain frequency, rank-frequency and type-token relation. Different models have been derived, and some characteristics have been set, in relation to vocabulary richness. However, the problem is much more complex. In this paper we will present a sketch of the field. Some kinds of counting are adequate for solving some problems but irrelevant for solving other ones, although in many cases no difference is apparent.

The first major distinction is between counting (a) word forms and (b) lemmas. Counting word forms is the usual procedure both for linguists and non-linguists because the written form affords the simplest and the most secure access to language. Lemmatisation is a complex job for professional linguists writing long programs, which work correctly up to about 90% of the time and thus may require some months of work correcting the program's errors made by the program. But even if the count is undertaken with pencil and paper, the identity of a lemma will remain an eternal problem and at least 10% of the result will be criticized by other linguists as incorrect. Hence, we may draw the interesting conclusion that lemmatisation programs are not so bad whatever they do. With lemma-counting, the vagueness of word identity becomes very conspicuous; while with word-form-counting the problem is only disguised, not solved. Consider for example German separable verbal prefixes which are described (by some linguists) as adverbs if they are not joined with the verb, but identified with prepositions in mechanical word-form-counting. Thus a verb can have double the number of forms in text that it actually has. Or consider the problem of suppletion existing in all European languages: we consider the cases of the German first person pronoun in German *ich, meiner, mir, mich* as forms of the same lemma but hesitate to include the plural form *wir* or the possessive form *mein*. In some languages there are two forms of *we* namely "we without you" and "we with you" (e.g. Indonesian *kami* and *kita*); do they belong to "I"? Hungarian has the lexeme *én* ("I") and forms like *nekem, hozzám, tőlem, engem,...* and a different plural. But in the third person *ő* ("he, she, it") there is a regular plural *ők*. In other languages there are other problems solvable only by using a set of criteria which are not inherent in data but applied as analytical means, i.e. our decisions.

Even if both forms of counting have their problems, for want of anything better they are used to describe some linguistic distribution problems (e.g. Zipf's law), today a very advanced discipline, some vocabulary richness problems, and type-token progressions. However, there are still other aspects that must be mentioned. (I) There are several units whose counting might be reasonable, namely (c) morphemes, accessible only to specialized linguists, and causing enormous difficulties in some languages; and (d) denotative units mostly called hrebs (c.f. e.g. Ziegler, Altmann 2002) which consist of all elements (morphemes, lexemes, phrases) denoting or referring to the same real or textual entity. They are very complex and are used for quite different purposes. Nevertheless, they are able to display a different aspect of vocabulary or style richness, namely the richness in expressing the same with different means. (II) The problem of plain word frequencies alone does in no case cover the full spectrum of frequency problems. It is merely the first step, the surface of the problem. Words of certain frequency can appear at special places in the sentence or in the text signalising their status (cf. e.g. Niemikorpi 1997; Uhlířová 1997); words of special frequency can be repeated at random or in quasi-regular distances (cf. Hřebíček 2000) that can be modelled; words of special frequency can have special forms or meanings – the association of length and meaning complexity with frequency are two of the oldest problems of quantitative linguistics (cf. Zipf 1935; Köhler 1986); and above all, frequency of words plays a basic role in the development of language and establishment of structures (cf. e.g. Bybee, Hopper 2001).

Here we shall restrict ourselves to the presentation of plain frequencies and searching for some possible characterizations. The presentation of frequencies can also be made in several ways. Up to now ten different ways have be proposed, though not all directly for word frequencies. Nevertheless, we shall discuss them all in turn.

(1) The *rank-frequency distribution of word forms* is the usual and the most popular presentation. It was originated by Zipf who supposed that *rank × frequency = constant*. Many empirical cases for which the harmonic series (truncated at the right side) holds can be found, and the theory has been strongly developed in the last decades. Here we ascribe rank 1 to the most frequent word, rank 2 to the second most frequent etc. It is believed that there is a law-like mechanism behind this order, even if in many cases it displays different forms. Some researchers have shown that even random texts (ape-typing) display the same behaviour (cf. Miller 1957; Li 1992). There is no contradiction in these conceptions because a background stochastic process working both in language and in non-language can result in the same distribution. Some researchers called the attention to the fact that synsemantics are situated at the beginning of the distribution, do not add much to the contents of the text, and have nothing to do with the problem of vocabulary richness. This has, again, two aspects: (I) There is no clear linguistic boundary between synsemantics and autosemantics (even using strict criteria); (II) in the rank-frequency presentation they are not strictly separated; there are always some autosemantics occurring at low ranks and some synsemantics occurring at high ranks. Hence, either one separates the word stock of the text and sets up two different curves for the two kinds of words, or one finds a point which approximately shows the separation line. As a matter of fact, such a point has been found by Hirsch (2005) and introduced in linguistics by Popescu (2006). It is very simple to find. In the rank-frequency presentation it is the point at which $r = f_r$, the point nearest to the origin [0,0], as illustrated in Figure 1.

The *h*-point can be set up approximately or one can find it by interpolation or by taking means of the neighbouring values. It is called *h-point* or *Hirsch-point* or, in linguistics, *Hirsch-Popescu-point*. See further below in 3.1.

Again there are several questions:

(α) has this point something to do with vocabulary richness?

(β) has it something to do with word class differentiation? and

(γ) does it depend on *N*, the text size?

The first question can be answered positively: the smaller the proportion of words occurring frequently, the higher the word proportion in the distribution tail in which infrequent words and hapax legomena occur. Hence, the smaller *h* is, the greater the richness.

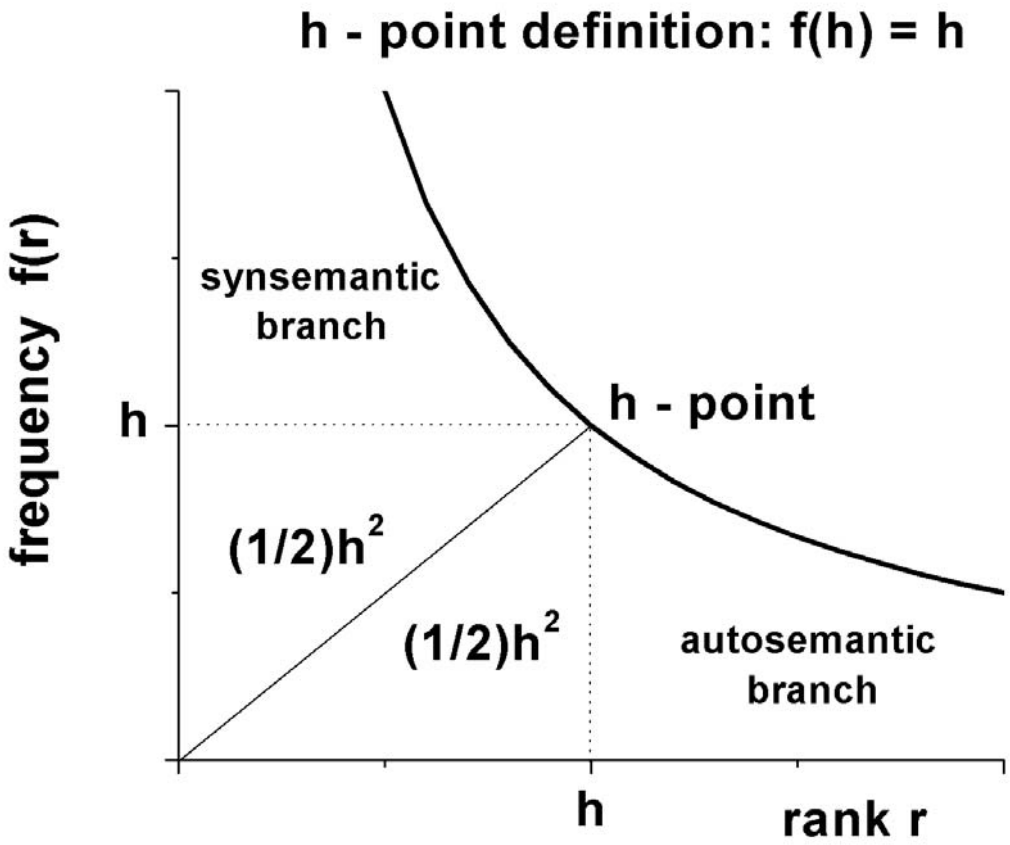


Figure 1. The *h*-point definition: the "bisector" point of the rank-frequency distribution at which rank = frequency

The second question has been analysed by several authors in different ways. Some assume the existence of two layers (auxiliary words and autosemantics); others even assume the existence of three layers, an idea which is not unrealistic; as a matter of fact, one may assume the existence of an arbitrary number of classes and in the infinity of texts some corroborating cases can always be found. We easily see that this problem can be solved by setting up confidence (or other) intervals for the *h*-point and employingh them as individual text characteristics.

The third problem is not yet solved. Theoretically, *h* must increase with increasing *N* because with increasing text, ever more words pass the $r = f(r)$-point while the addition of hapax legomena slows down. But this depends not only on the text but also on the analysed language. In highly analytic languages the situation will differ from that in highly synthetic ones. According to a private communication from B.D. Jayaram (2006), in two analysed Indian languages the share of the cumulative rank distribution up to the *h*-point does not change with increasing *N*. If this result could be ascertained in different languages, the *h*-coverage, or better the 1 - *F*(*h*) coverage, where *F*(*h*) denotes *the relative cumulative distribution up to h*, could be considered a stable characteristic of the vocabulary richness.

Since we obtain a different result (cf. 3.1), evidently this depends on the language, on the sample of texts used, on boundary conditions that are not known, and so on.

(2) The *rank-frequency distribution of lemmas* means a drastic change of circumstances. While in English the lemma of the article *the* is identical with its only word form, in German 7 forms of the article (of gender, case, number) belong to one lemma. Consequently, it is a very frequent word in all its forms, thus the *h*-point must lay at a greater rank, the corresponding $F(h)$ being greater and 1- $F(h)$ (which can also be considered as an index of vocabulary richness) respectively smaller. For the sake of illustration let us consider a German newspaper article (communication by R. Köhler 2006) in which the word-form rank-frequency distribution yielded $N = 1076$, $h = 13$ and $F(h) = 0.3383$, hence 1- $F(h) = 0.6617$. However, in the same text, there were found $N = 892$ lemmas with $h = 16$ and $F(h) = 0.7152$, i.e. there is a difference that must be levelled out. Thus no direct comparison is possible.

The next presentations are counted doubly, one holding for word forms, the other one for lemmas. From the linguistic point of view this difference is relevant.

(3) and (4). *The cumulative rank-frequency distribution* is usually presented in form of cumulative relative frequencies. They yield a concave curve (sequence). Since cumulative relative frequencies sum up to 1, arguing analogically we can find on this sequence again a point which is nearest to the point [0, 1]. However, here the rank itself does not play any role, nevertheless, $1 - F(h)$ could be used as an index of vocabulary richness. But again, counting lemmas we obtain a result which is not reasonably interpretable. Now, if we relativize also the ranks (calling them *rr*), i.e. divide each rank by the highest rank (V), we obtain the relation $<rr, F(r)>$, an analogon of the Lorenz curve used in economy and sociology. We shall present it below. The point we seek is given as the minimum of $[rr^2 + (1 - F(r))^2]^{1/2}$. The point of nearest distance will be called *m*-point. Its computation is presented in Figure 2.

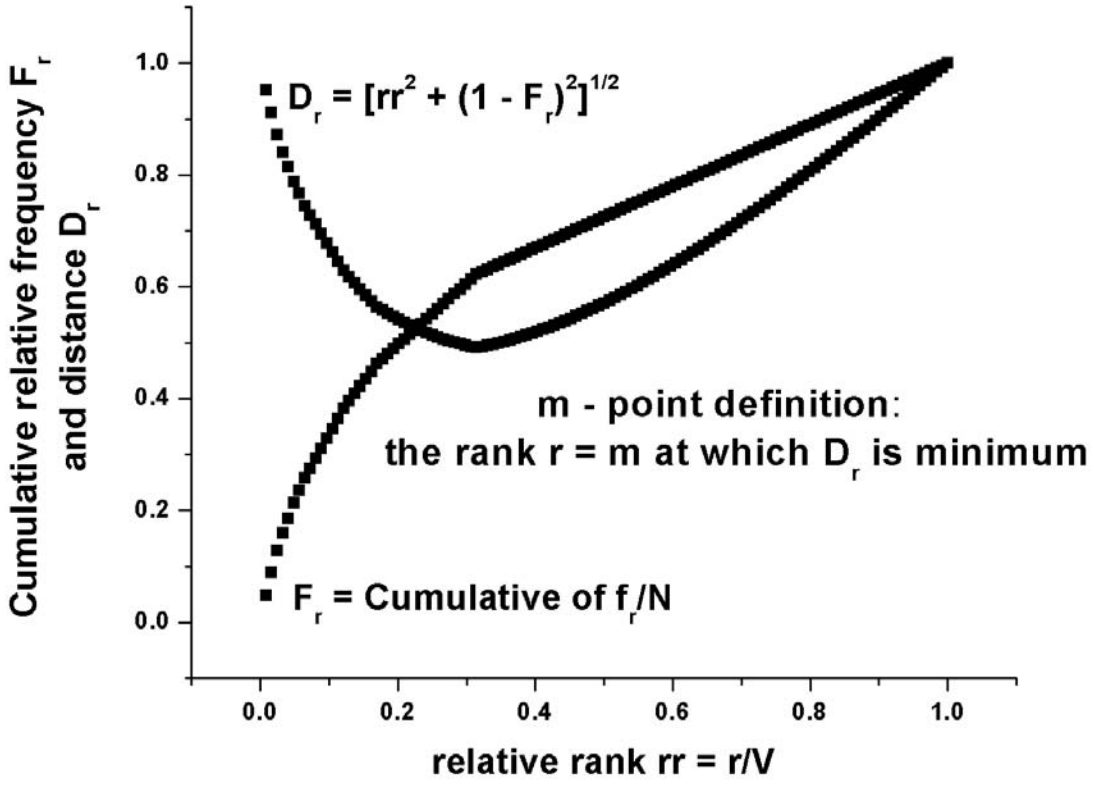


Figure 2. Computation of the *m*-point

Again, we can perform the same procedure for word forms and lemmas, obtaining two different results, and we can construct intervals for the *m*-point (minimum of the D_r curve) in Fig. 2. Further, we can approximate F_r using a continuous curve; the same can be done for D_r.

(5) and (6). *Reversed ranking*. Under some circumstances we begin to ascribe ranks in reverse order, i.e. the most infrequent word obtains rank 1, the next one rank 2, etc., and the most frequent word gets the highest rank. Then both ranks and frequencies are relativized: the ranks are divided by the highest rank yielding, say R_i ; and the relative frequencies are cumulated, yielding, say S_i (see 3.3). Thus we obtain the relation $<R_i, S_i>$ yielding a monotone convex sequence called Lorenz curve. If we add the point [0,0] to the data, then the sequence begins at zero and ends at 1. Again, we may seek the analogy to the *h*-point as that which is nearest to [1,0] using the same Euclidian distance as above; but there has been no investigation of this up to now. We do not know how text length affects the sequence even if both ranks and frequencies are relativized. Instead, we can proceed in the same way as it is done in other sciences. We start from the fact that if every word occurred in the text exactly once, i.e. if the text had maximal possible vocabulary richness, the Lorenz "curve" would be a straight line from [0,0] to [1,1], since relativized ranks would be equal to the cumulative relative frequencies, i.e. $R_i = S_i$. In this way we get at least a concept of maximal vocabulary richness. However, real texts differ from this maximum and the difference consists in the area between this straight line and the real Lorenz curve of the text. The area is usually called *Gini's coefficient* and can be computed as the sum of small trapezoids between the two lines. Since, in this way, everything is relativized, texts should be comparable independently of *N*. But even in this case, *G* can depend on *N* as will be shown below. However, in any case, the greater Gini's coefficient, the greater the disparity of word participation in text building, i.e. the smaller the vocabulary richness is. Since rank-frequency distributions of words follow a kind of Zipf law, this conclusion holds generally true; but there could be some not quite natural texts contradicting this interpretation. A wide-ranging investigation in many languages may bring a solution. Again, the difference between word-form counting and lemma counting could provide some surprises.

In long-tail distributions encompassing word counts, the point next to [1, 0] would be the point containing the hapax legomenon with the highest rank (remember that the ranking is reversed) or some of the words with frequency 2.

(7) and (8). *Ranking of differences.* This representation has been introduced by Balasubrahmanyan and Naranan (1996), as far as we know. Here the frequencies are decreasingly ordered and relativized, i.e. one obtains first all p_r. Then one defines a new variable $d_r = p_r - p_{r+1}$ and analyses its form. In this way one can obtain partial sums distributions used sporadically in linguistics and musicology (Wimmer, Altmann 2001). Again, word forms and lemmas can yield different results.

(9) and (10). *Frequency spectrum.* Every rank-frequency distribution can be easily transformed in a frequency spectrum – though the formal transformation is not always simple. But in empirical data it is easy to state that there are exactly f_1 words occurring once, f_2 words occurring twice, …, f_x words occurring *x*-times, simply by adding "from below". Some linguists call this representation as "frequency of frequency" or "lexical frequency". The random variable is "number of occurrences" and the frequency is the "number of words having the given number of occurrences". Again, there is a difference between word forms and lemma distributions. Since these monotone decreasing sequences have a hyperbolic form, the same operations can be performed as with ranked distributions. We can find the *h-point* analogue, named in the following *k-point*, and the analogous nearest point to [0, 1] for the cumulative distribution. But here the interpretation of the *k*-point is just the opposite: the greater $F(k)$ (denoting here *the relative cumulative distribution up to k*), the greater the vocabulary richness. Some authors constructed different indices based only on hapax leg-

omena ($x = 1$), other ones used the first 50 frequency classes ($x = 1,\ldots,x = 50$), but the k-point is a unique, objective point telling something about the dynamics of the sample. It is, of course, different for word forms and for lemmas but further investigation will show the relation between them based on the synthetism of language. For the time being, the study of texts is separated from the study of language but step by step we will bring them together and will be able to draw conclusions across the two fields.

2. Tasks and problems

In "plain" word frequency research different tasks have been formulated: (a) *Text characterization*, fulfilled using different statistics whose sampling properties are not always known, thus possibly rendering any conclusions irrelevant. If known statistics are used, we can draw relatively safe conclusions. Characterization is used, for example, for solving disputed authorship problems, for psychiatric purposes, for the study of language ontogenesis, for decisions about the affiliation of a text to a genre (and generally for establishing genres), for discourse analysis, for typologically characterizing a given language, and for drawing conclusions about other properties of language, e.g. vocabulary richness. The last task is so problematic that it has developed as an independent discipline encompassing a number of approaches, extensive literature and countless case studies (see point *d* below).

(b) A special part of the above task is the *comparison of texts* either by word for word comparison, global comparison using indices of some properties or comparing the frequency distributions (using e.g. chi-square statistics, information statistics, nonparametric tests etc.)

(c) *Information flow* is the study of increase of information in the course of the text. Usually we try to capture it by a curve or by an index of type-token-ratio. There are a great number of curves established by different arguments – and their number can still be indefinitely increased (they are monotonously increasing and concave) – but the background problems are so serious that there are always years of stagnation in the research. First of all, type-token relation (TTR) can be measured in three ways, of which only one can be set in connection with information flow (cf. Wimmer 2005); the others produce fractals. Second, it measures only a part of the new information because not only new words (types) furnish new information but also a new connection of old ones. Third, the type-token ratio reproduces a part of the hearer's information; that of the speaker is quite different (cf. Andersen, Altmann 2006). Fourth, there is an enormous difference between counting the increase of form-types and lemma-types. Counting form-types, strongly synthetic languages (such as Hungarian) seem to have a more rapid information flow than strongly isolating languages (such as English), which is nonsense. Thus lemma-type counting would be the only correct way; but most investigators count word-form-types because it is simpler. Fifth, the type-token curve or index is frequently used to measure the vocabulary richness of a text. On one hand, if counting forms, only intra-language comparisons are possible; on the other, all indices depend on N and their confidence intervals are enormous, hence judgements about vocabulary richness using TTR are problematic. Sixth, what *is* vocabulary richness? Does it have something in common with the velocity of increase of new types, or only their number in the text? The first drops absolutely with increasing text length, the second relatively. Does it have something to do with the extent of author's vocabulary? Surely not, because writing adults know all words of a language (except specialist terminology); it depends rather on the theme, the author's selectivity, the purpose of the text (e.g. didactic text, poetic text, newspaper text…). But how these entities can be measured? Thus discussion about the concept itself can be continued *ad infinitum* (cf. Wimmer, Altmann 1999). Intuitively we know that there should

be something like vocabulary richness but its operationalization is a matter of definitions and criteria. Below we shall show some possible ways differing from the known ones.

(d) *Modeling word frequencies* is the hobby of mathematicians shared seldom by linguists unless they want to show the adequate application of a model in a language. The problem goes beyond the boundaries of linguistics and Zipf's law is a well known concept in a great number of scientific disciplines. There are different approaches leading to different models and whole families of distributions. Baayen (2001, 2005) mentions urn models, LNRE (Large number of rare events) models and power models to which one can add the proportionality model from which curves and distributions, both discrete and continuous can be derived (cf. Wimmer, Altmann 2005). All models contain some truth but the problem of the meaning of parameters and the boundary conditions are no problems for mathematicians. Sometimes, they may be put in the drawer of ceteris-paribus conditions but this is no definitive solution.

A further problem of this research is the *homogeneity* of the analysed text, seldom taken into account even by linguists. Some linguistic problems can be solved using corpora, e.g. the study of grammatical usage, but for other problems, e.g. the study of vocabulary richness, a corpus is irrelevant. It is well known that in a novel consisting of several chapters, each chapter can turn out to be different from the previous ones in general or special aspects. The boundary conditions change not only in dependence on the theme but also in dependence on the pause in writing. The disposition of the writer is changed after a pause (e.g. coffee break or night), it brings new rhythms, e.g. word length or sentence length rhythm; the theme change causes jumps in the type-token curve, i.e. the new chapter brings a new stock of types which can destroy the regularity of the distributions found in the previous chapters. The heavy problem in textology is the fact that there are no populations having fixed values of properties (cf. Orlov, Boroda, Nadarejšvili 1982). There is nothing like "Goethe's sentence length" or "Shakespeare's word-frequency distribution". If such populations existed, then every sample from them should reflect the given property within an admitted confidence interval. That means that samples from that population must not differ significantly from each another and a pooled sample must still reflect the property of the population. It has been shown that in textology this is not true. Though some properties may remain constant in the course of a novel – and in that case it must be demonstrated – other ones may change, building a time series, a chaotic sequence, quasi-regular runs, a complex oscillating sequence or something else. In a drama even the speech of individual persons may display properties which are not in conformity with those of the drama as a whole (which is a very problematic population). Hence it is safer to study some textual properties in sufficiently large self-contained "natural" parts and if they are homogeneous (within the admitted confidence interval) to pool them and make statements about the text as a whole. Unfortunately, studies of the sequential character of texts are not popular and we still do not know which of the infinite number of text properties may be considered to be the stable ones (cf. e.g. Hřebíček 1997, 2000).

3. Characterizations

In the following sections we shall take into account word forms, allowing us an access to texts in different languages without the necessity of insecure lemmatisation; we shall strive for text characterization using some new methods and we shall try to draw conclusions concerning other text properties, e.g. vocabulary richness.

In the following we shall investigate four characteristic points of word frequency distribution:

1. The *h*-point, defined as that point on the rank-frequency distribution which is the nearest to the [0, 0], i.e. to the origin (see Figure 1)

2. The *k*-point which is defined in the same way but for the frequency spectrum. It differs from the *h*-point in its interpretation in connection with vocabulary richness.

3. The *m*-point concerns the cumulative distribution $F(m)$ of frequencies and one can, again distinguish two points, one for the rank-frequencies, the other for frequency spectra. A further differentiating property is taking ranks in absolute or in relative values. The *m*-point, in any form, is the nearest to the [0,1] point.

4. The *n*-point is computed only for rankings but the ranking is performed in reverse order, i.e. the smallest frequency has rank 1, the second smallest rank 2 etc. The ranks themselves are considered in relative values, i.e. $rr = r/V$. The *n*-point is the nearest to [1,0]. The cumulative values of relative frequencies yield the Lorenz curve, and the area between the bisector and the Lorenz curve is called Gini's coeffcient.

3.1. The *h*-point

The history of the *h*-index is short. It has been recently introduced by Hirsch (2005) in scientometrics as a tool for evaluating individual scientific output. Subsequently, one of us proposed its extension to linguistic analysis (Popescu 2006). The main reason requiring the introduction of the *h*-index in text evaluation matters may be summarized as follows. Let us consider a (more or less) Zipfian, hyperbolic rank-frequency distribution, as schematically shown in Fig.1. The area covered by the distribution curve represents the total word count or the text length (size), *N*, while the maximal rank gives the total number of unique, distinct, different words, that is the text vocabulary, *V*. It is obvious that there always exists a very special (rank, frequency) point, a "crossing point", at which the frequency is nearest to its rank, a value denoted by *h*. This is the definition of the *h*-point having the extraordinary property that it splits any vocabulary (*V*) into two word classes, namely in (1) a number of the order of magnitude of *h* of highly frequent synsemantic or auxiliary words (such as prepositions, conjunctions, pronouns, articles, adverbs, and others that are likewise meaningful only in the company of or reference to other words) and (2) usually a much larger number ($V - h$) of lowly frequent, seldom, autosemantic words, with $V >> h$, actually building the very vocabulary of the text. In other words, the *h*-point appears as a natural cutting point of the word-frequency distribution into two very distinct branches, namely the "rapid" branch, of relatively few (about *h*) unique words that are highly frequent, and the "slow" branch, of many (about $V - h$) unique words that are of low frequency. The role played by the *h*-point in separating autosemantics from synsemantics of a text reminds us of the tole of "Maxwell's demon" in physics in separating high velocity from low velocity gas molecules. In this connection, of particular interest in text analysis are relative cumulative frequencies up to the *h*-point, denoted by $F(h)$, and defined as the ratio of the area under the distribution curve from rank = 1 up to the rank = *h*, and the total area under this curve (the text length *N*). Let us illustrate this point by a simple example using the rank-frequency distribution of word forms in J.W.v.Goethe's poem "Erlkönig" as shown in Table 1.

As can easily be seen, the *h*-point is at rank $r = 6$ because $f(6) = 6$, too. Since $N = 225$, the $F(h) = (11 + 9 + 9 + 7 + 6 + 6)/225 = 0.2133$. If we want to express very approximately the vocabulary richness of the text, we must use $1 - F(h)$ yielding $1 - 0.2133 = 0.7867$ and add (or subtract) some index of synthetism constructed specially for this purpose. (For some synthetism indices see Altmann, Lehfeldt 1973).

Table 1
Rank-frequency distribution of word forms in Goethe's "Erlkönig"

Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency
1	11	11	4	21	3	31	2
2	9	12	4	22	2	32	2
3	9	13	4	23	2	33	2
4	7	14	4	24	2	34	2
5	6	15	4	25	2	35	2
6	*6*	16	3	26	2	36	2
7	5	17	3	27	2	37	2
8	5	18	3	28	2	38	2
9	4	19	3	29	2	39	2
10	4	20	3	30	2	40-124*	1

* The ranks 40 to 124 have frequency 1

As already Hirsch argued, there exists a simple relationship between h and the total area under the rank–frequency curve, namely

(1) $$N = a\, h^2$$

a being a constant of the word distribution and N the total word count of the considered text (the text length). From this it follows the difference equation

(2) $$\Delta N = 2ah\, \Delta h$$

which gives how the location of the h-point changes with increasing text length. Hence, by division, we eliminate the constant a and obtain a quite general relationship, namely

(3) $$\Delta N/N = 2\, \Delta h/h$$

In particular, in order to obtain an increase of $\Delta h = 1$, we need an increase of N given by $(\Delta N)_{\Delta h=1} = 2\, N/h$.

In Table 2 we show some results from self-contained complete Nobel lecture English texts of which, perhaps, only some were written in genuine English (see reference on *The Nobel Lecture*). However, a good sign that the selected text sample is quite homogeneous can be considered the stability of the ratio $a = N/h^2$ at an average value of 7.5 and a standard deviation of only 0.74. This quality can be appreciated by taking into consideration the large interval of the a-value, ranging from 4.5 to 9.5, for a great variety of literary texts (Popescu, 2006).

As discussed above, a reliable measure of vocabulary richness appears to be the quantity 1- $F(h)$, where $F(h)$ is *the relative cumulative distribution up to h*, and this value is given in Table 2. However, a fundamental correction should be made to the F-values as defined previously. This is related to the well known fact that there is always some ambiguity in separating synsemantics from autosemantics. The effective overlapping area of this genuine "linguistic Gordian knot" is of the order h^2, as it can be evaluated by a naked eye inspection of the rank-frequency distribution. It is clear that we overestimated so far the $F(h)$ synsemantics area and, correspondingly, underestimated the 1 - $F(h)$ autosemantics area. As a basic correction of our former $F(h)$ data, it appears quite naturally to cut the h^2 "knot" exactly into two parts, as suggested in Fig.1, one part to be subtracted from synsemantics and one part to be

added to autosemantics. Consequently, we finally have to correct the $F(h)$ of the rank-frequency distribution as follows

$$\underline{F(h)} = [Areah - h^2/2] / N = F(h) - h^2/2N \quad (4)$$

and, similarly, the $F(k)$ of the frequency spectrum distribution

$$\underline{F(k)} = [Areak - k^2/2] / V = F(k) - k^2/2V \quad (5)$$

Table 2
A sample of 33 Nobel lectures sorted by $1 - \underline{F(h)}$

Year	Field	Nobel Awardee	*N*	*V*	*h*	*k*	1 - *F*(*h*)	1 - $\underline{F(h)}$	*F*(*k*)	$\underline{F(k)}$	*a* = N/h^2
1996	Lit	Wislawa Szymborska	1982	826	16	6	0.7321	**0.7967**	0.9395	0.9177	7.74
2002	Peace	Jimmy Carter	2330	939	16	6	0.6996	**0.7545**	0.9414	0.9222	9.10
1986	Peace	Elie Wiesel	2693	945	19	6	0.6755	**0.7425**	0.9280	0.9090	7.46
1935	Chem	Irène Joliot-Curie	1103	390	12	6	0.6745	**0.7398**	0.9333	0.8871	7.66
1993	Lit	Toni Morrison	2971	1017	22	7	0.6382	**0.7197**	0.9351	0.9110	6.14
1976	Lit	Saul Bellow	4760	1495	26	7	0.6317	**0.7027**	0.9472	0.9308	7.04
1975	Med	Renato Dulbecco	3674	1005	22	8	0.6353	**0.7012**	0.9284	0.8966	7.59
1930	Lit	Sinclair Lewis	5004	1597	25	7	0.6325	**0.6950**	0.9474	0.9321	8.01
1959	Lit	Salvatore Quasimodo	3695	1255	21	7	0.6327	**0.6924**	0.9474	0.9279	8.38
1989	Econ	Trygve Haavelmo	3184	830	21	9	0.6209	**0.6902**	0.9373	0.8885	7.22
1986	Econ	James M. Buchanan Jr.	4622	1232	23	7	0.6326	**0.6898**	0.9221	0.9022	8.74
1989	Peace	Dalai Lama	3597	1030	23	7	0.6122	**0.6857**	0.9388	0.9150	6.80
1950	Lit	Bertrand Russell	5701	1574	29	8	0.6102	**0.6840**	0.9428	0.9225	6.78
1905	Med	Robert Koch	4281	1066	24	9	0.6157	**0.6830**	0.9306	0.8926	7.43
1953	Peace	George C. Marshall	3247	1001	19	8	0.6255	**0.6811**	0.952	0.9200	8.99
1970	Lit	Alexandr Solzhenitsyn	6512	1890	32	9	0.6023	**0.6809**	0.9524	0.9310	6.36
1975	Econ	Leonid V. Kantorovich	3923	1042	22	8	0.6179	**0.6796**	0.9367	0.9060	8.11
1983	Peace	Lech Walesa	2586	769	19	6	0.6079	**0.6777**	0.9129	0.8895	7.16
1902	Phys	Pieter Zeeman	3480	908	21	8	0.6118	**0.6752**	0.9273	0.8921	7.89
1973	Lit	Heinrich Böll	6088	1672	28	9	0.6107	**0.6751**	0.9474	0.9232	7.77
1991	Peace	Mikhail Gorbachev	5690	1546	26	11	0.6062	**0.6656**	0.9592	0.9201	8.42
1920	Phys	Max Planck	5200	1342	24	10	0.6002	**0.6556**	0.9508	0.9135	9.03
1984	Lit	Jaroslav Seifert	5241	1325	26	9	0.5903	**0.6548**	0.9404	0.9098	7.75
1963	Peace	Linus Pauling	6246	1333	28	10	0.5908	**0.6536**	0.9347	0.8972	7.97
1925	Med	John Macleod	4862	1176	24	9	0.5907	**0.6499**	0.9379	0.9035	8.44
1925	Med	Frederick G. Banting	8193	1669	32	11	0.5871	**0.6496**	0.9401	0.9039	8.00
2004	Lit	Elfriede Jelinek	5746	1038	33	8	0.5522	**0.6470**	0.8863	0.8555	5.28
1979	Peace	Mother Teresa	3820	636	26	9	0.5571	**0.6456**	0.8789	0.8152	5.65
1911	Chem	Marie Curie	4317	1016	25	9	0.5691	**0.6415**	0.9409	0.9010	6.91
1902	Phys	Hendrik A. Lorentz	7301	1423	31	9	0.565	**0.6308**	0.9178	0.8893	7.60
1938	Lit	Pearl Buck	9088	1825	39	10	0.5453	**0.6290**	0.9326	0.9052	5.98
1908	Chem	Ernest Rutherford	5083	985	26	12	0.5448	**0.6113**	0.9442	0.8711	7.52
1965	Phys	Richard P. Feynman	11265	1659	41	11	0.5337	**0.6083**	0.9066	0.8701	6.70

where N = text size, V = text vocabulary, and the wording *Areah* and *Areak* stands for the area under the distribution curve up to the h-point and k-point respectively. The correctness of this procedure can be appreciated by its consequences, that is by the excellent ranking rearrangement of the tabulated data, Table 2, as shown in the graphs of Fig. 3. The important conclusion is that the $F(k)$ index does not depend on N, while the $1 - F(h)$ index varies monotonously in the expected direction (downwards) in dependence on N (see Fig. 4) as can easily be computed from the data in Table 2.

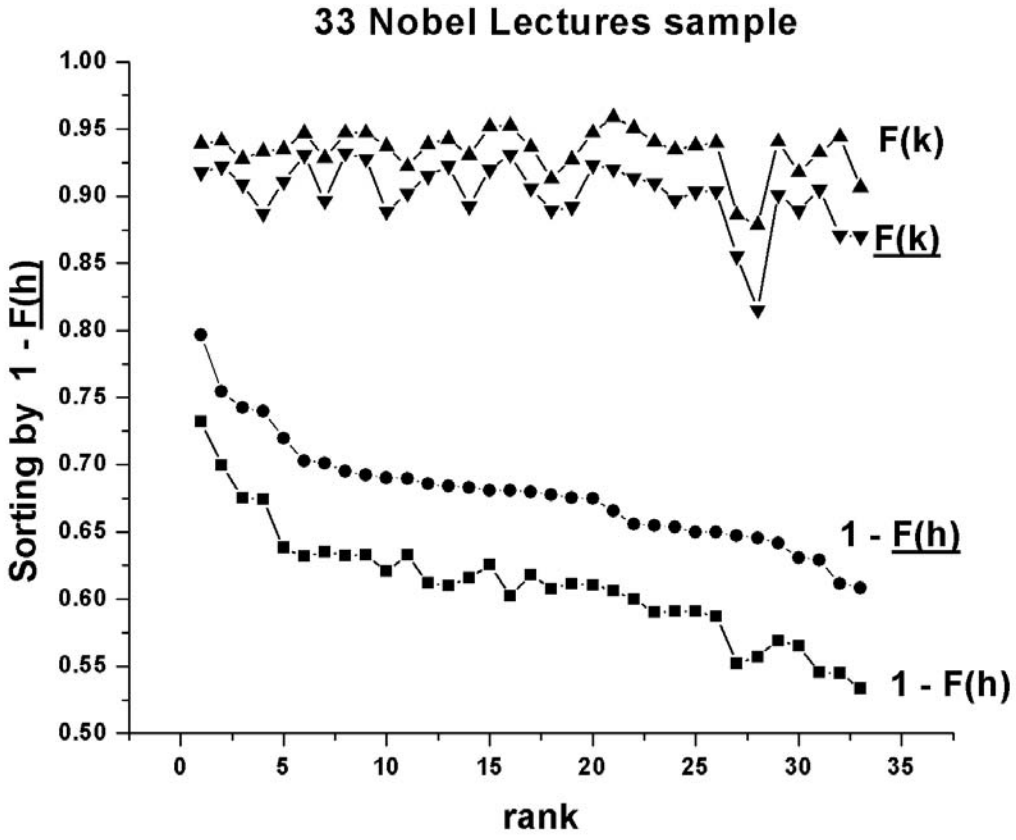


Figure 3. Some characteristics and ranking of a 33 Nobel lectures text sample

Generally, we shall point out that the boundary between auxiliary words and content words in actual texts is a fuzzy one rather than a sharp one, as it would appear from the above definition of the h-point at $f(h) = h$, with the corresponding relative cumulative frequency $F(h)$ up to h. Thus, for instance, for Goethe's "Erlkönig" rank-frequency distribution we found $h = 6$, $f(6) = 6$, and $F(6) = 48/225 = 0.2133$; while summing up all extant, real auxiliary words of this distribution, we get a value close to $F(17) = 92/225 = 0.4089$. Consequently, we have an ideal h-point, $h = 6$, defined merely on ideal symmetry grounds, and an effective H-point, $H = 17$, expressing the real boundary location and self-diffusion of synsemantics and autosemantics. Actually, any word frequency analysis in these terms should establish the relationship between these two cardinal points, the ideal, symmetric one, h-point, and the effective, real one, H-point.

The problem can be solved in different ways, e.g. setting up asymmetrical intervals, using order statistics or considering the two domains as fuzzy sets whose elements have different degrees of belonging in one of the domains. Here we shall consider only the problem of vocabulary richness. While $F(k)$ and $\underline{F(k)}$ do not depend (in our sample) of N, they can be

directly used as richness indices. In Table 3 we present an ordering of 33 Nobel lectures by *F(k)*. Though many natural scientists are in the first half and writers in the second, no conclusions can be drawn from this observation. In the same way, there is no relation of vocabulary richness to the year of origin. *F(k)* seems to be a quite independent index.

Table 3
33 Nobel lectures sorted by *F(k)*

Year	Field	Nobel Awardee	F(k)	N	Year	Field	Nobel Awardee	F(k)	N
1979	Peace	Mother Teresa	0,8152	3820	1975	Econ	Leonid V. Kantorovich	0,9060	3923
2004	Lit	Elfriede Jelinek	0,8555	5746	1986	Peace	Elie Wiesel	0,9090	2693
1965	Phys	Richard P. Feynman	0,8701	11265	1984	Lit	Jaroslav Seifert	0,9098	5241
1908	Chem	Ernest Rutherford	0,8711	5083	1993	Lit	Toni Morrison	0,9110	2971
1935	Chem	Irène Joliot-Curie	0,8871	1103	1920	Phys	Max Planck	0,9135	5200
1989	Econ	Trygve Haavelmo	0,8885	3184	1989	Peace	Dalai Lama	0,9150	3597
1902	Phys	Hendrik A. Lorentz	0,8893	7301	1996	Lit	Wislawa Szymborska	0,9177	1982
1983	Peace	Lech Walesa	0,8895	2586	1953	Peace	George C. Marshall	0,9200	3247
1902	Phys	Pieter Zeeman	0,8921	3480	1991	Peace	Mikhail Gorbachev	0,9201	5690
1905	Med	Robert Koch	0,8926	4281	2002	Peace	Jimmy Carter	0,9222	2330
1975	Med	Renato Dulbecco	0,8966	3674	1950	Lit	Bertrand Russell	0,9225	5701
1963	Peace	Linus Pauling	0,8972	6246	1973	Lit	Heinrich Böell	0,9232	6088
1911	Chem	Marie Curie	0,9010	4317	1959	Lit	Salvatore Quasimodo	0,9279	3695
1986	Econ	James M. Buchanan Jr.	0,9022	4622	1976	Lit	Saul Bellow	0,9308	4760
1925	Med	John Macleod	0,9035	4862	1970	Lit	Alexandr Solzhenitsyn	0,9310	6512
1925	Peace	Frederick G. Banting	0,9039	8193	1930	Lit	Sinclair Lewis	0,9321	5004
1938	Lit	Pearl Buck	0,9052	9088					

On the other hand, the quantity *F(h)* depends on N as shown in Fig. 4. The trend can be captured by the curve $1\text{-}F(h) = 1.5467N^{-0.0985}$ which, though yielding only a low determination coefficient $R = 0.60$, shows highly significant F and t values. Though this trend will surely be modified by adding several other texts in different languages, knowing the dependence we can use it for estimating the vocabulary richness of a text in a very simple way. Ignoring the absolute value of 1 – *F(h)* we consider only its difference to the computed curve which is a kind of norm (a kind of mean) for Nobel lectures. The results are shown in Table 4.

Table 4
Vocabulary richness measured by 1- *F(h)*

Nobel Awardee	*N*	1-*F(h)*	Theor	Diff	Nobel Awardee	*N*	1-*F(h)*	Theor	Diff
Irène Joliot-Curie	1103	0.7398	0.7757	-0.0359	Saul Bellow	4760	0.7027	0.6716	0.0311
Wislawa Szymborska	1982	0.7967	0.7322	0.0645	John Macleod	4862	0.6499	0.6702	-0.0203
Jimmy Carter	2330	0.7545	0.7206	0.0324	Sinclair Lewis	5004	0.6950	0.6683	0.0267
Lech Walesa	2586	0.6777	0.7132	-0.0355	Ernest Rutherford	5083	0.6113	0.6673	-0.0560
Elie Wiesel	2693	0.7425	0.7103	0.0321	Max Planck	5200	0.6556	0.6658	-0.0102
Toni Morrison	2971	0.7197	0.7035	0.0162	Jaroslav Seifert	5241	0.6548	0.6653	-0.0105
Trygve Haavelmo	3184	0.6902	0.6988	-0.0086	Mikhail Gorbachev	5690	0.6656	0.6599	0.0057
George C. Marshall	3247	0.6811	0.6974	-0.0163	Bertrand Russell	5701	0.6840	0.6598	0.0242
Pieter Zeeman	3480	0.6752	0.6927	-0.0175	Elfriede Jelinek	5746	0.6470	0.6593	-0.0123
Dalai Lama	3597	0.6857	0.6904	-0.0047	Heinrich Böll	6088	0.6751	0.8555	0.0196
Renato Dulbecco	3674	0.7012	0.6890	0.0122	Linus Pauling	6246	0.6536	0.6539	-0.0029
Salvatore Quasimodo	3695	0.6924	0.6886	0.0038	Alexandr Solzhenitsyn	6512	0.6809	0.6512	0.0299
Mother Teresa	3820	0.6456	0.6863	-0.0407	Hendrik A. Lorentz	7301	0.6308	0.6439	-0.0131
Leonid V. Kantorovich	3923	0.6796	0.6845	-0.0049	Frederick G. Banting	8193	0.6496	0.6366	0.0130
Robert Koch	4281	0.6830	0.6787	0.0043	Pearl Buck	9088	0.6290	0.6312	-0.0012
Marie Curie	4317	0.6415	0.6781	-0.0366	Richard P. Feynman	11265	0.6083	0.6170	-0.0087
J.M. Buchanan Jr.	4622	0.6898	0.6736	0.0162					

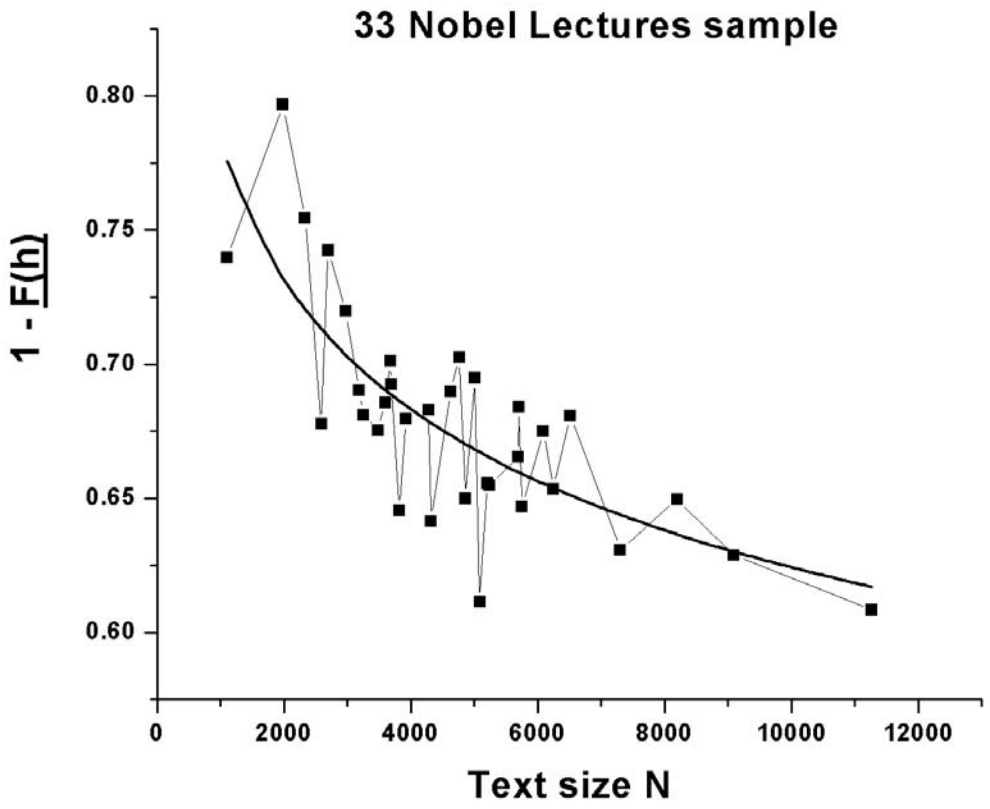


Figure 4. Dependence of $\underline{F(h)}$ on N.

The difference values can be normalized or relativized. However, the computed "theoretical" values hold only for the given data. Every new text would slightly change both the curve and the differences. Thus whatever the meaning of the above difference, vocabulary richness must be ascertained by means of several independent procedures.

3.2. The m-point

The *m*-point can be found if we directly scrutinize the cumulative rank-frequency distribution taking into account both relative frequencies and *relative* ranks. Since the sum of relative frequencies is 1 and ranks go from $1/V$ to $V/V = 1$, the curve approaches the point [1,1]. The *F*-curve has in our case a monotonously increasing concave form. Let us transform the values from Table 1 in *F* and present them as Table 5. In our case the relative ranks are $rr = r/124$ and F_r is the cumulative relative frequency.

Table 5
Cumulative distribution of ranked word forms of Goethe's "Erlkönig"

rr	F_r	rr	F_r	rr	F_r	rr	F_r
0.0081	0.0489	0.0887	0.3111	0.1693	0.4622	0.2500	0.5511
0.0161	0.0889	0.0968	0.3289	0.1774	0.4711	0.2581	0.5600
0.0242	0.1289	0.1048	0.3467	0.1855	0.4800	0.2661	0.5689
0.0322	0.1600	0.1129	0.3644	0.1935	0.4889	0.2742	0.5778
0.0403	0.1867	0.1210	0.3822	0.2016	0.4978	0.2823	0.5867
0.0484	0.2133	0.1290	0.3956	0.2097	0.5067	0.2903	0.5956
0.0565	0.2356	0.1371	0.4089	0.2177	0.5156	0.2983	0.6044
0.0645	0.2578	0.1452	0.4222	0.2258	0.5244	0.3065	0.6133
0.0726	0.2756	0.1532	0.4356	0.2339	0.5333	0.3145	0.6222
0.0807	0.2933	0.1613	0.4489	0.2419	0.5422	*	**

*from 40 up to 124 by step 1/124 = 0.0081
** by step 0.00444444

To find the given point we compute the Euclidian distance from [0, 1] as

$$D_r = \sqrt{rr^2 + (1 - F_r)^2} \qquad (6)$$

where rr is the relative rank, yielding $D_1 = [0.0081^2 + (1 - 0.0489)^2]^{1/2} = 0.9511$, $D_2 = 0.9113$, $D_3 = 0.8714$, $D_4 = 0.8406$, …, $D_{38} = 0.4934$, $D_{39} = 0.4916$, $D_{40} = 0.4934$; …, $D_{124} = 1.0000$. The smallest distance is at rank $m = 39$ where frequency 2 occurred for the last time. The $F_{39} = 0.6222$. The computation is shown in Fig. 5

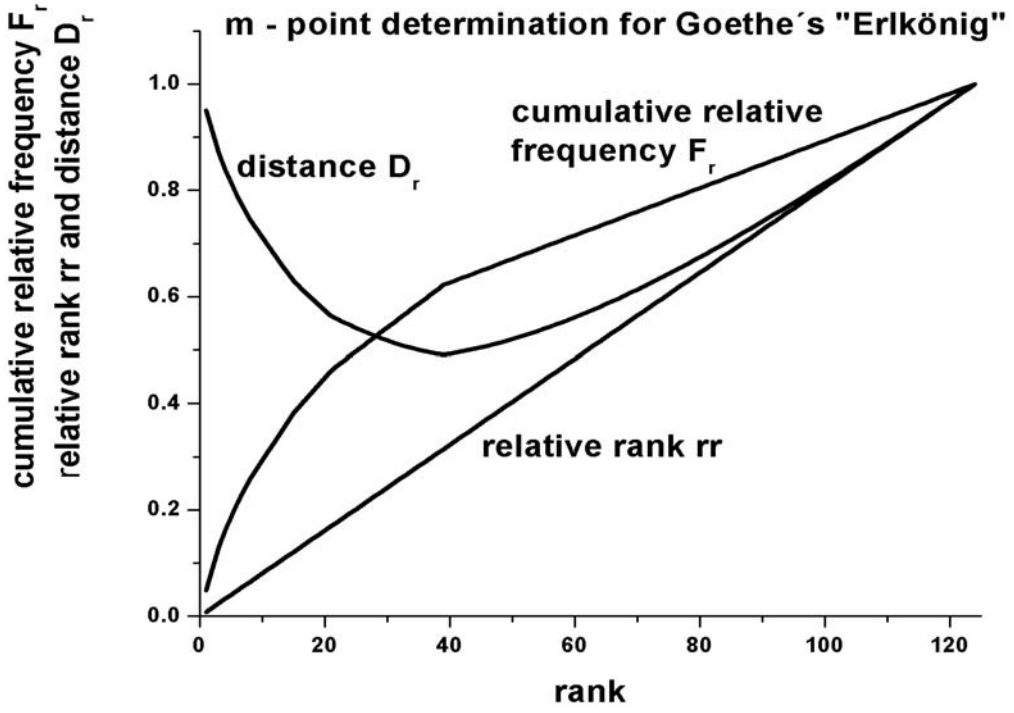


Figure 5. The m-point determination using the minimal distance to [0,1]

In Table 6, there are some results from the same texts as in Table 2, showing the *m*-point and the corresponding cumulative frequency *F*(*m*).

Table 6
The m-point and the cumulative frequencies

Year	Field	Nobel Awardee	*N*	*V*	*m*	1 - *F(m)*
1996	Lit	Wislawa Szymborska	1982	826	215	0,3133
2002	Peace	Jimmy Carter	2330	939	239	0,3163
1986	Peace	Elie Wiesel	2693	945	208	0,313
1935	Chem	Irène Joliot-Curie	1103	390	93	0,3191
1993	Lit	Toni Morrison	2971	1017	207	0,2975
1976	Lit	Saul Bellow	4760	1495	293	0,3008
1975	Med	Renato Dulbecco	3674	1005	225	0,2727
1930	Lit	Sinclair Lewis	5004	1597	307	0,3006
1959	Lit	Salvatore Quasimodo	3695	1255	258	0,3031
1989	Econ	Trygve Haavelmo	3184	830	179	0,2754
1986	Econ	James M. Buchanan Jr.	4622	1232	271	0,2746
1989	Peace	Dalai Lama	3597	1030	240	0,2722
1950	Lit	Bertrand Russell	5701	1574	325	0,2619
1905	Med	Robert Koch	4281	1066	213	0,2684
1953	Peace	George C. Marshall	3247	1001	207	0,2978
1970	Lit	Alexandr Solzhenitsyn	6512	1890	341	0,2816
1975	Econ	Leonid V. Kantorovich	3923	1042	229	0,2763
1983	Peace	Lech Walesa	2586	769	165	0,2769
1902	Phys	Pieter Zeeman	3480	908	191	0,269
1973	Lit	Heinrich Böell	6088	1672	337	0,2654
1991	Peace	Mikhail Gorbachev	5690	1546	334	0,265
1920	Phys	Max Planck	5200	1342	278	0,2671
1984	Lit	Jaroslav Seifert	5241	1325	264	0,2624
1963	Peace	Linus Pauling	6246	1333	278	0,2448
1925	Med	John Macleod	4862	1176	234	0,2742
1925	Med	Frederick G. Banting	8193	1669	334	0,2457
2004	Lit	Elfriede Jelinek	5746	1038	198	0,2183
1979	Peace	Mother Teresa	3820	636	129	0,2427
1911	Chem	Marie Curie	4317	1016	193	0,2699
1902	Phys	Hendrik A. Lorentz	7301	1423	273	0,2461
1938	Lit	Pearl Buck	9088	1825	337	0,2294
1908	Chem	Ernest Rutherford	5083	985	194	0,2506
1965	Phys	Richard P. Feynman	11265	1659	298	0,2274

It can easily be shown that *m* and 1-*F*(*m*) depend on *N* even if the determination coefficient is not too great. Hence it can be used in the same way as $\underline{F(h)}$ for estimating the vocabulary richness taking the difference between the computed and the observed points. However, here a theoretical approach would be more appropriate.

3.3. Gini's coefficient

The overall characterisation of word frequency distribution can be further performed by means of the Gini-coefficient used mostly in economics. As a matter of fact it shows us the deviation of the maximum vocabulary richness which would be attained if all words occurred exactly once. There are other indices doing this work, e.g. Herfindahl's repeat rate or Shannon's entropy or simply the skewness of the frequency distribution, etc.; but Gini's coefficient has not been used up to now.

If we present the distribution in cumulative form such that all frequencies are 1 (maximum vocabulary richness), then the cumulative frequencies would equal to the relative ranks, i.e. $rr_i = F_i$. In Cartesian coordinates this would yield a straight line of 45^o from [0,0] to [1,1]. Since in other sciences the ranking is performed in reversed form, we shall show it here, too. In programming this can be done mechanically, we shall present it explicitly. Starting from Table 1 we leave the ranks as they are but begin to rank frequencies "from below", i.e. the smallest frequency obtains rank 1, the second smallest rank 2 etc. At the same time we take both the relative values of ranks and the cumulative relative values of frequencies. The beginning of such a table is given in Table 7.

Table 7
Reverse ranking of frequencies in Goethe's "Erlkönig"

Rank r	Frequency f_r	Relative rank rr	Relative frequency p_r	Relative cumulative frequency F_r
1	1	1/124 = 0.00806	1/225 = 0.00444	0.00444
2	1	0.01613	0.00444	0.00888
3	1	0.02419	0.00444	0.01333
4	1	0.03226	0.00444	0.01778
5	1	0.04032	0.00444	0.02222
.....				
124	11	1.0000	11/225 = 0.04889	1.00000

As can be seen, the empirical values (of F_r) are positioned below this line (represented by rr). Hence Gini's coefficient is defined as the proportion of the area between the empirical F_r-values (last column) and the straight line between [0,0] and [1,1] to the whole area under the straight line, as shown in Fig. 3. The sequence of small straight lines $<rr_i, F_r>$ is usually called Lorenz curve. The whole area under the straight line is 0.5 and in our terms it means maximal vocabulary poverty (there is only one word steadily repeated in the text).

The area between the Lorenz curve and the straight line consists of small trapezoids. The area of a trapezoid is given as

$$A = \frac{a+b}{2}h \tag{7}$$

where a and b are the unequal sides and h is the height. The height $h = rr_{i+1} - rr_i$ while the two sides are given as

$$a = rr_i - F_i$$
$$b = rr_{i+1} - F_{i+1}$$

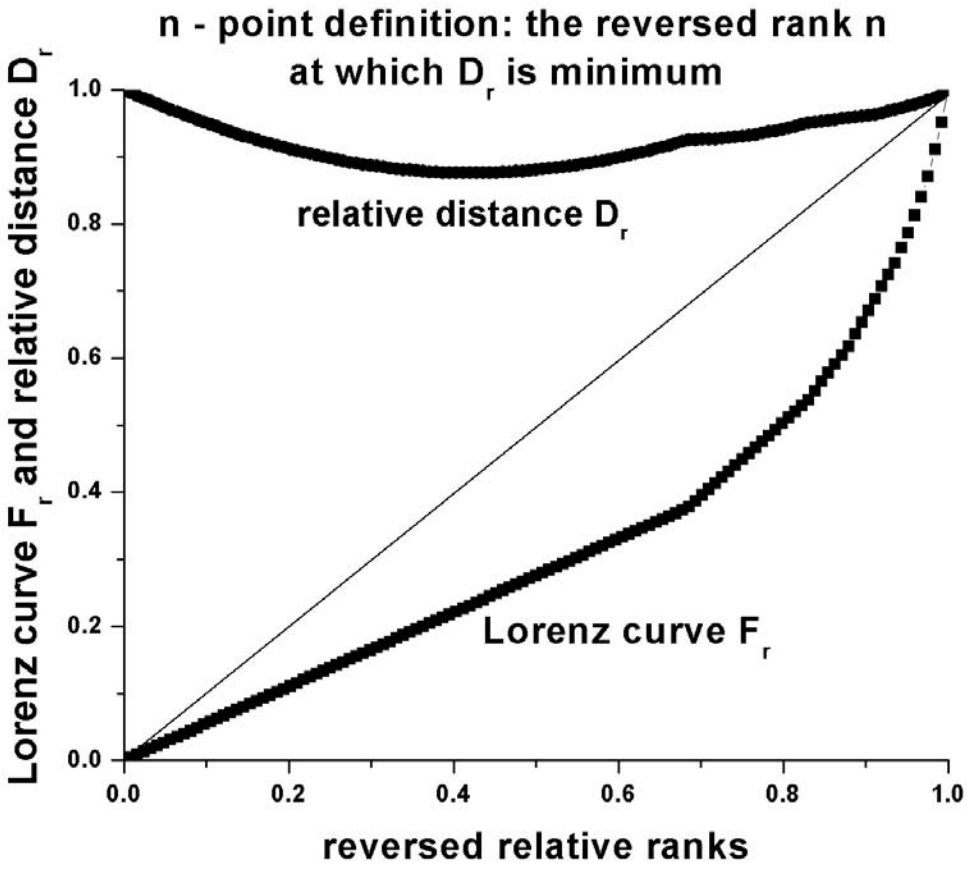


Figure 6. Lorenz curve. Reversed relative ranks ($rr = x$) against cumulative relative frequencies (y)

hence

$$(8) \quad A = \frac{(rr_i - F_i) + (r_{i+1} - F_{i+1})}{2}(rr_{i+1} - rr_i)$$

Consider the first trapezoid in the above scheme. We have

$$A_1 = \frac{(rr_1 - F_1) + (rr_2 - F_2)}{2}(rr_2 - rr_1)$$

$$= [(0.00806 - 0.00444) + (0.0161 - 0.00888)](0.0161 - 0.00806)/2 = 0.0000435.$$

and Gini's coefficients will be computed as follows. First we compute the sum of the trapezoids

$$(9) \quad G_1 = \sum_{i=1}^{V-1} \frac{(rr_i - F_i) + (r_{i+1} - F_{i+1})}{2}(rr_{i+1} - rr_i)$$

yielding $G_1 = 0.1828$. Then we compute the proportion of this area to the whole area under the straight line, i.e.

$G = 0.1828/0.5 = 0.3656.$

Fortunately, there are other methods for computing G without the necessity of reversing the frequencies and computing relative frequencies and cumulative frequencies. Consider V as the highest rank and N as text length, i.e. sum of all frequencies as given in Table 1 ($V = 124$, $N = 225$). Then we obtain G directly as

$$G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} rf_r \right) \tag{10}$$

yielding exactly the same value as the first procedure. Still other variants are known. For $V >> 1$ it is approximately

$$G = 1 - \frac{2}{VN}\sum_{r=1}^{V} rf_r \,. \tag{11}$$

Gini's coefficient shows the position of the text between maximal and minimal vocabulary richness. Of course, the situation will differ with lemmatized texts and with the frequency spectrum. Some values for selected texts are given in Table 8.

Table 8
Gini's coefficient for 33 Nobel texts (rank-frequency) ordered according to N

N	1-G	N	1-G	N	1-G
1103	0.4479	3695	0.3959	5241	0.3337
1982	0.4786	3820	0.3067	5690	0.3504
2330	0.4605	3923	0.3610	5701	0.3453
2586	0.3741	4281	0.3415	5746	0.2823
2693	0.4216	4317	0.3281	6088	0.3455
2971	0.3975	4622	0.3615	6246	0.3163
3184	0.3629	4760	0.3826	6512	0.3514
3247	0.3846	4862	0.3372	7301	0.2982
3480	0.3517	5004	0.3796	8193	0.3047
3597	0.3740	5083	0.3068	9088	0.2826
3674	0.3579	5200	0.3422	11265	0.2640

Formula (9) is very practical for further computation and evaluation. Since the last expression in the formula is the arithmetic mean of the rank frequency distribution, one can derive the variance of G in a straightforward way as

$$Var(G) = \frac{4\sigma^2}{V^2 N} \tag{11}$$

and use it, for example, for setting up confidence intervals, comparisons with other texts, classifications or for studying text evolution, etc.

Needless to say, the Lorenz curve can be approximated by a continuous curve whose parameters can be used for comparisons, too.

In Table 8 we can see that 1-G depends in a high degree on N. A simple dependence yields 1-G = 2.68069$N^{-0.24196}$ with a determination coefficient R = 0.70 (and highly significant F and t tests) which can be improved by an additive constant; but since the computation is preliminary we show only the evaluation method. Again, we consider not the absolute value of 1-G but its difference to the "expected" value given by the above formula. Thus, e.g. for N = 3820 yielding 1-G = 0.3067 we obtain 1-G_t = 2.68069(3820)$^{-0.24196}$ = 0.3643. The difference (1-G) - (1-G_t) = G_t - G = 0.3067 - 0.3643 = -0.0576, i.e. the vocabulary richness of this text is smaller than expected. In the same way we can estimate the other texts and obtain the results in Table 9. Another possibility would be setting up confidence intervals for the above curve but its incessant change when adding new texts would force us to unending evaluations. Hence we show only the method.

Table 9
Evaluation of Gini's coefficient for 33 Nobel lectures
(ordered according text length N)

Year	Field	Awardee	N	1-G	1-G_t	Diff(G)
1935	Chem	Irène Joliot-Curie	1103	0.4479	0.4921	-0.0442
1966	Lit	W. Szymborska	1982	0.4786	0.4270	0.0516
2002	Peace	Jimmy Carter	2330	0.4605	0.4106	0.0499
1983	Peace	Lech Walesa	2586	0.3741	0.4004	-0.0263
1986	Peace	Elie Wiesel	2693	0.4216	0.3965	0.0251
1993	Lit	Toni Morrison	2971	0.3975	0.3872	0.0103
1989	Econ	Trygve Haavelmo	3184	0.3629	0.3808	-0.0179
1953	Peace	G.W. Marshall	3247	0.3846	0.3790	0.0056
1902	Ohys	Pieter Zeeman	3480	0.3517	0.3727	-0.0210
1989	Peace	Dalai Lama	3597	0.3740	0.3697	0.0043
1975	Med	Renato Dulbecco	3674	0.3579	0.3678	-0.0099
1959	Lit	S. Quasimodo	3695	0.3959	0.3673	0.0286
1979	Peace	Mother Teresa	3820	0.3067	0.3643	-0.0576
1975	Econ	L.V. Kantorovich	3923	0.3610	0.3620	-0.0010
1905	Med	Robert Koch	4281	0.3415	0.3544	-0.0129
1911	Chem	Marie Curie	4317	0.3281	0.3537	-0.0256
1986	Econ	J.M.Buchanan Jr.	4622	0.3615	0.3479	0.0136
1976	Lit	Saul Bellow	4760	0.3826	0.3455	0.0371
1925	Med	John Macleod	4862	0.3372	0.3437	-0.0065
1930	Lit	Sinclair Lewis	5004	0.3796	0.3413	0.0383
1908	Chem	E. Rutherford	5083	0.3068	0.3400	-0.0332
1920	Phys	Max Planck	5200	0.3422	0.3381	0.0041
1984	Lit	Jaroslav Seifert	5241	0.3337	0.3375	-0.0038
1991	Peace	M. Gorbachev	5690	0.3504	0.3309	0.0195
1950	Lit	Bertrand Russell	5701	0.3453	0.3307	0.0146
2004	Lit	Elfriede Jelinek	5746	0.2823	0.3301	0.0478
1973	Lit	Heinrich Böll	6088	0.3455	0.3255	0.0200
1963	Peace	Linus Pauling	6246	0.3163	0.3235	-0.0072
1970	Lit	A. Solzhenitsyn	6512	0.3514	0.3202	0.0312
1902	Phys	H.A. Lorentz	7301	0.2982	0.3115	-0.0133
1925	Med	F. G. Banting	8193	0.3047	0.3029	-0.0018
1938	Lit	Pearl Buck	9088	0.2826	0.2954	-0.0128
1965	Phys	R.P. Feynman	11265	0.2640	0.2805	-0.0165

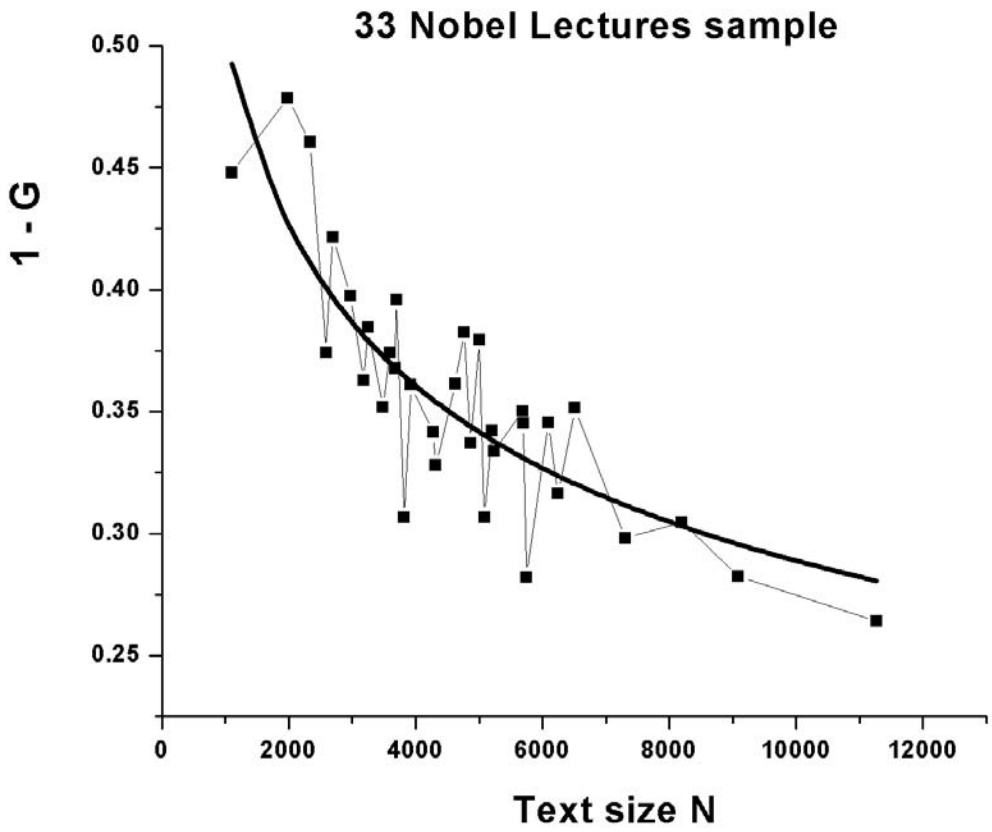


Figure 7. 1 – G in terms of text size N

As can be seen, the differences are not identical with those measured with F(h) but display the same trend. This is probably caused by the different expected trend. If we compare the two Figures (4 and 7), we can see the same course of observed values, hence one of the two coefficients is sufficient to characterize the vocabulary.

A complete table of all results and a graph are shown in the Appendix.

4. Conclusions

Our results, which may be considered first steps in a slightly different characterization of word frequencies, can be summarised as follows.

The possibility exists of distinguishing auxiliary words from content words using the *h*-point. With the aid of intervals, we could for every text ascertain the domain of pure auxiliary words, the mixed domain, and the domain of content words. On the other hand, we could use this for typological purposes, showing these domains in strongly analytical or synthetic languages. Though there will be always a difference between individual texts, it must be possible to ascertain general tendencies. Since h and $F(h)$ change with increasing N, this increase can be different in different languages.

As to vocabulary richness, the *k*-point of the frequency spectrum seems to be a relatively stable index not changing with increasing N; hence, $F(k)$ can be used for this purpose. It shows approximately the proportion of content words in the frequency spectrum. It would be interesting to compare texts of parents with those of their children, texts of people with

psychiatric disorders with those of "healthy" individuals, the texts of primitive literary forms with those of modern novels, poetic texts with scientific texts, texts of the same sort from different diachronic periods, or texts of a single writer at different points in his development, and so on.

The dependence of different text indices on N is a pitch for mathematicians as well as linguists. While researchers trying to characterize vocabulary richness strive for indices which are independent of N, other researchers want to see which characteristics are dependent on N. If we know how something changes in dependence on something else, we can eliminate the influence in an adequate mathematical way. But this is rather a task for mathematicians. Linguists usually take as many logarithms as necessary until they obtain curves whose differences look very insignificant. It would be better to study the sampling behaviour of indices, characterizing a well defined property. But how is vocabulary richness defined at all? No definition which could be directly operationalized can be found. While probability distributions of words can easily be compared, it is not so easy to compare the existing indices of vocabulary richness. Though the quantity $F(k)$ above seem to be very stable, and though we have a well-relativized variance of Gini's coefficient so that texts would be comparable in spite of their different sizes, we postpone this task, hoping to have inspired other researchers.

For typological purposes, all these indices are very useful. One might proceed as follows. Take a text, translate it into several languages, and compare the above indices at least by ordering the languages according to their magnitude. Then take other properties, and study their relation to the textual properties. Try to draw conclusions about the behaviour of $F(h)$, $F(k)$, G, etc. based on other properties of language.

References

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Andersen, S., Altmann, G. (2006). Information content of words in text. In: Grzybek, P. (ed.), *Contributions to the science of language: Word length studies and related issues: 93-117*. Boston: Kluwer.

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht/Boston/London: Kluwer.

Baayen, R.H. (2005). Word frequency distributions. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 397-409*. Berlin/New York: de Gruyter.

Balasubrahmanyan, V.K., Naranan, S. (1996). Quantitative linguistics and complex *s*ystem studies*. Journal of Quantitative Linguistics 3(3): 177-228.*

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: Benjamins.

Gini, C. (1912). "Variabilità e mutabilità" Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955)., see at http://en.wikipedia.org/wiki/Gini_coefficient

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf, see more at http://en.wikipedia.org/wiki/Hirsch_number

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Jayaram, B.D. (2006), private communication.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2006), private communication.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory 38, 1842-1845.*

Miller, G.A. (1957). Some effects of intermittent silence. *The American J. of Psychology 70, 311-314.*

Niemikorpi, A. (1997). Equilibrium of words in the Finnish frequency dictionary. *J. of Quantitative Linguistics 4(1-3), 190-196.*

Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Text, Sprache Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Popescu, I.-Iovitz (2006). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 553-562.* Berlin/New York: de Gruyter.

The Nobel Lectures, http://nobelprize.org/nobel/

Uhlířová, L. (1997). Length vs. order: word length and clause length from the perspective of word order. *J. of Quantitative Linguistics 4(1-3), 266-275*

Wimmer, G. (2005). The type-toke-relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 361-368*. Berlin/New York: de Gruyter.

Wimmer, G., Altmann, G. (1999). On vocabulary richness. *Journal of Quantitative Linguistics 6(1), 1-9.*

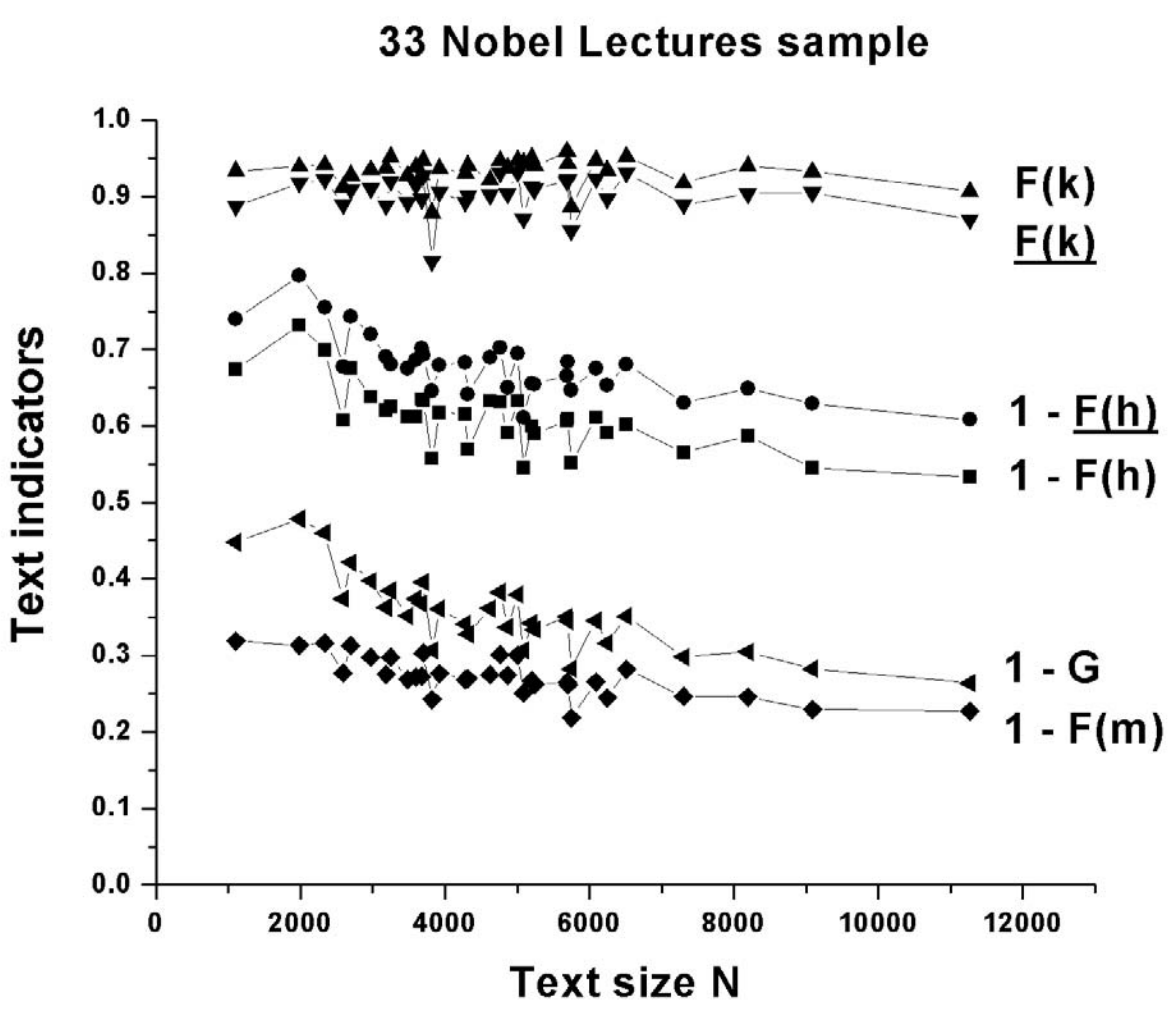
Wimmer, G., Altmann, G. (2001). Models of rank-frequency distributions in language and music. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 283-294*. Trier: WVT.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807*. Berlin/New York: de Gruyter.

Ziegler, A., **Altmann, G.** (2002). *Denotative Textanalyse. Ein textlinguistisches Arbeitsbuch*. Wien: Praesens.

Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton Mifflin.

Indicators of 33 Nobel lectures sorted by 1 - *F(h)*													
Year	Field	Nobel Awardee	*N*	*V*	*h*	*k*	*m*	1 - *F(h)*	1 - *F(h)*	*F(k)*	*F(k)*	*1 - F(m)*	*1 - G*
1996	Lit	Wislawa Szymborska	1982	826	16	6	215	0,7321	0,7967	0,9395	0,9177	0,3133	0,4786
2002	Peace	Jimmy Carter	2330	939	16	6	239	0,6996	0,7545	0,9414	0,9222	0,3163	0,4605
1986	Peace	Elie Wiesel	2693	945	19	6	208	0,6755	0,7425	0,928	0,9090	0,313	0,4216
1935	Chem	Irène Joliot-Curie	1103	390	12	6	93	0,6745	0,7398	0,9333	0,8871	0,3191	0,4479
1993	Lit	Toni Morrison	2971	1017	22	7	207	0,6382	0,7197	0,9351	0,9110	0,2975	0,3975
1976	Lit	Saul Bellow	4760	1495	26	7	293	0,6317	0,7027	0,9472	0,9308	0,3008	0,3826
1975	Med	Renato Dulbecco	3674	1005	22	8	225	0,6353	0,7012	0,9284	0,8966	0,2727	0,3679
1930	Lit	Sinclair Lewis	5004	1597	25	7	307	0,6325	0,6950	0,9474	0,9321	0,3006	0,3796
1959	Lit	Salvatore Quasimodo	3695	1255	21	7	258	0,6327	0,6924	0,9474	0,9279	0,3031	0,3959
1989	Econ	Trygve Haavelmo	3184	830	21	9	179	0,6209	0,6902	0,9373	0,8885	0,2754	0,3629
1986	Econ	James M. Buchanan Jr.	4622	1232	23	7	271	0,6326	0,6898	0,9221	0,9022	0,2746	0,3615
1989	Peace	Dalai Lama	3597	1030	23	7	240	0,6122	0,6857	0,9388	0,9150	0,2722	0,374
1950	Lit	Bertrand Russell	5701	1574	29	8	325	0,6102	0,6840	0,9428	0,9225	0,2619	0,3453
1905	Med	Robert Koch	4281	1066	24	9	213	0,6157	0,6830	0,9306	0,8926	0,2684	0,3415
1953	Peace	George C. Marshall	3247	1001	19	8	207	0,6255	0,6811	0,952	0,9200	0,2978	0,3846
1970	Lit	Alexandr Solzhenitsyn	6512	1890	32	9	341	0,6023	0,6809	0,9524	0,9310	0,2816	0,3514
1975	Econ	Leonid V. Kantorovich	3923	1042	22	8	229	0,6179	0,6796	0,9367	0,9060	0,2763	0,3610
1983	Peace	Lech Walesa	2586	769	19	6	165	0,6079	0,6777	0,9129	0,8895	0,2769	0,3741
1902	Phys	Pieter Zeeman	3480	908	21	8	191	0,6118	0,6752	0,9273	0,8921	0,269	0,3517
1973	Lit	Heinrich Böell	6088	1672	28	9	337	0,6107	0,6751	0,9474	0,9232	0,2654	0,3455
1991	Peace	Mikhail Gorbachev	5690	1546	26	11	334	0,6062	0,6656	0,9592	0,9201	0,265	0,3504
1920	Phys	Max Planck	5200	1342	24	10	278	0,6002	0,6556	0,9508	0,9135	0,2671	0,3422
1984	Lit	Jaroslav Seifert	5241	1325	26	9	264	0,5903	0,6548	0,9404	0,9098	0,2624	0,3337
1963	Peace	Linus Pauling	6246	1333	28	10	278	0,5908	0,6536	0,9347	0,8972	0,2448	0,3163
1925	Med	John Macleod	4862	1176	24	9	234	0,5907	0,6499	0,9379	0,9035	0,2742	0,3372
1925	Med	Frederick G. Banting	8193	1669	32	11	334	0,5871	0,6496	0,9401	0,9039	0,2457	0,3047
2004	Lit	Elfriede Jelinek	5746	1038	33	8	198	0,5522	0,6470	0,8863	0,8555	0,2183	0,2823
1979	Peace	Mother Teresa	3820	636	26	9	129	0,5571	0,6456	0,8789	0,8152	0,2427	0,3067
1911	Chem	Marie Curie	4317	1016	25	9	193	0,5691	0,6415	0,9409	0,9010	0,2699	0,3281
1902	Phys	Hendrik A. Lorentz	7301	1423	31	9	273	0,565	0,6308	0,9178	0,8893	0,2461	0,2982
1938	Lit	Pearl Buck	9088	1825	39	10	337	0,5453	0,6290	0,9326	0,9052	0,2294	0,2826
1908	Chem	Ernest Rutherford	5083	985	26	12	194	0,5448	0,6113	0,9442	0,8711	0,2506	0,3068
1965	Phys	Richard P. Feynman	11265	1659	41	11	298	0,5337	0,6083	0,9066	0,8701	0,2274	0,2640


33 Nobel Lectures sample
Text indicators
Text size N
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2000
4000
6000
8000
10000
12000
F(k)
F(k)
1 - F(h)
1 - F(h)
1 - G
1 - F(m)