Measures of word commonness

Petr Savický[1] Jaroslava Hlaváčová[2]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou Věží 2, 182 07, Prague, Czech Republic
e-mail: savicky@cs.cas.cz

[2] Institute of the Czech National Corpus, Faculty of Arts of Charles University, nám. J. Palacha 2, 116 38, Prague, Czech Republic
e-mail: jaroslava.hlavacova@ff.cuni.cz

corresponding author: J. Hlaváčová

Abstract

The main goal of this paper is to investigate methods of how to rank words in a way that corresponds to an intuitive notion of "commonness". Since there is no formal definition of such a notion, our techniques may be considered as a suggestion for such a definition.

The commonness of words is sometimes roughly substituted with their frequency in a language corpus. In order to suggest a better measure, we define a quantity, which we call *corrected frequency*. It depends not only on the frequency of a word in a corpus, but also on its distribution within the corpus. Unlike previous solutions of the same problem, we take the corpus as an uninterrupted sequence of words with no regard to borders between files, texts, genres, or any others.

We introduce three different corrected frequencies. Their definitions are based on notions of information theory and analysis of random processes. Their values for individual words depend on the corpus. Hence, it is important to what extent they are stable with respect to the selection of the corpus. In order to investigate the suggested corrected frequencies from that point of view, we compare their values on five different subcorpora of the whole corpus.

We present several examples of words taken from the Czech National Corpus that demonstrate in which way the corrected frequencies correspond to the intuitive commonness of these words.

Introduction

This research was motivated by a practical problem - how to decide which words should be included in a universal dictionary of a specified size, and which not. At first glance, the answer is simple: take the most common words until you reach the given number. However, there is no well-defined measure for commonness of words in the language.

Recently, large corpora are taken as a basis for making dictionaries. For example, The New Oxford Dictionary of English (1998) was prepared using the British National Corpus (BNC, http://www.hcu.ox.ac.uk/BNC/) containing about 10^8 word occurrences and the Collins Cobuild English Dictionary (1995) was prepared using the Bank of English - a corpus containing more than $2 \cdot 10^8$ word occurrences at the time of its edition (http://titania.cobuild.collins.co.uk/boe_info.html). The initial selection of words which will be considered as possible dictionary entries is done according to their frequency in the corpus. In that sense, the word frequency serves as a first approximation of the word commonness.

The drawback of using frequency alone is that some words occur in one (or a few) small part(s) of the corpus only. Even if they have a high frequency in the corpus nobody would say that they are equally common in the language as words with the same frequency but distributed evenly throughout the whole corpus. In order to have an objective measure of word commonness it is necessary not only to look at frequency, but also to take note of distribution within the corpus.

The word "souvrství" can serve as an example. It is the Czech geological term whose English translation is "formation". In the Czech National Corpus (CNC, http://ucnk.ff.cuni.cz/) it has the same frequency (891) as the words "vzrušující" ("exciting") or "krůček" (diminutive of the word "step"). Everybody feels that these words are not equally common. If we look at the word "souvrství" more carefully, we find out that almost 96% of all its occurrences in the corpus (namely 855) belong to one text only - popular guide about interesting geological sites in Prague.

Lexicographers know this problem very well and make corrections to the initial word selection, but on an intuitive basis. In this article we suggest three measures that could help them make decisions more objectively. In other words, the measures allow us to rank words of a corpus in a way which corresponds to the concept of word commonness in the language.

Several different approaches have been undertaken towards measuring word dispersion in a corpus, see e.g. Carroll et. al. (1971), Králík (1978), or Oakes (1998). They have, as far as we know, one common property - they require pre-division of the corpus into genre groups. According to our experience with building the representative corpus, any trial of a text annotation brings plenty of problems, which are very difficult, if not even impossible, to resolve. It is hard to decide how many genres to take into account. Moreover, there is no strict border between genres, no matter how many of them we would have. For this reason, we developed a method of measuring dispersion of word distribution in a corpus that does not require classification of the texts.

For our method, the whole corpus is considered as one sequence of words obtained by concatenation of all the texts forming the corpus. The information about dispersion of the distribution of a concrete word is extracted from the positions of the occurrences of the word in this sequence.

Our method of measuring dispersion of words assigns to the words special values - corrected frequencies. Words that are evenly distributed, have the corrected frequency close to their absolute frequency. For unevenly distributed words, the corrected frequency is smaller. The exact definitions are in the following section, where three different types of the corrected frequency are presented.

The positions of the words depend, naturally, on the ordering of the individual texts in the corpus. Our method is based on the observation that the words that occur only in a specific type of text often occur in clusters and, hence, have corrected frequency substantially smaller than pure frequency. The chance that this happens is higher, if similar texts, for example texts from the same source, are placed consequtively in the corpus.

In the paper, we describe three possible definitions of the corrected frequency and present results of some experiments processed on the Czech National Corpus. In the experiments, we investigated the three corrected frequencies of all words in the corpus. We present some examples demonstrating that corrected frequencies are more adequate measures of commonness than pure frequency.

Further, we calculated the corrected frequencies for the same word on several different corpora. This allows us to estimate for each of the corrected frequencies, how much its values for the same word vary for different selections of the corpus. The exact results and comparison of the three corrected frequencies from this point of view are presented in the section Experimental comparison of the stability of the measures.

Notation

In the following considerations, "word" can mean word form, lemma (basic form of the word), or any other unit of text, even a morphologic tag, etc.

Let N be the length of the corpus, i.e. the number of words in it. We divide the whole corpus into N positions numbered from 1 to N. Each position is occupied by one word. Thus, the k-th word in the corpus sits in the position k.

For simplicity of notation, we assume for the whole article that a word w is selected and fixed, although it may be selected arbitrarily. This allows us not to include the word into the notation.

Let f be the frequency of the selected word in the corpus, i.e. the number of all its occurrences. For $i = 1, \dots, f$, let n_i be the position of the i-th occurrence of the word in the corpus. The word positions divide the corpus into intervals. In order to have the intervals disjoint, each interval contains the occurrence of w at its end, but not the occurrence at its beginning. The interval whose end-point is n_i will be called the left interval corresponding to the occurrence i. For $i = 2, \dots, f$, it is the interval $[n_{i-1} + 1, n_i]$. The left interval corresponding to the first occurrence n_1 is defined using the cyclic order as the union of two intervals $[n_f + 1, N] \cup [1, n_1]$.

Further, we use the following notation for the distance between two consecutive occurrences of the selected word. Namely, let $d_i = n_i - n_{i-1}$ for every $i = 2, \dots, f$ and let $d_1 = n_1 + (N - n_f)$, which is the distance between the last occurrence of the word and the first one in the cyclic order. Clearly, for all $i = 1, \dots, f$, d_i is the length of the left interval corresponding to the occurrence n_i. Notice, that

$$\sum_{i=1}^{f} d_i = N. \tag{1}$$

Corrected frequency

The corrected frequency will be defined in such a way that for an evenly distributed word, it is equal to its frequency. On the other hand, for a word occurring in one very small part of the corpus, the corrected frequency is close to 1, regardless of the pure frequency. These two requirements specify the corrected frequency in the two extreme cases. In order to specify the behaviour of the corrected frequency also in intermediate cases, we used three different techniques based on information theory and analysis of random processes, leading to three different corrected frequencies (measures of commonness). At first, in the following three sections, we introduce three different quantitative measures of a word distribution: average reduced frequency (ARF), average waiting time (AWT) and average logarithmic distance (ALD). Then, we will define the three corrected frequencies, based on these three measures.

Average reduced frequency

The first approach is the "reduced frequency" of the word defined as follows, see also Hlaváčová, Rychlý (1999), Hlaváčová (2000). If the frequency of the considered word is f, we divide all positions of the corpus into f segments of roughly equal length. If N is divisible by f, then the segments are of equal length. Otherwise, the lengths differ at most by one. If we denote $v = N/f$, the length of each segment is either $\lceil v \rceil$, the smallest integer not smaller than v, or $\lfloor v \rfloor$, the largest integer not larger than v. Reduced frequency is then the number of segments containing at least one occurrence of the word. If the word was distributed entirely evenly, its reduced frequency would be f, since each segment would contain exactly one occurrence. On the other hand, if the word occurred in one small part of the corpus only, its reduced frequency would be 1, if all the occurrences fall into one segment, or 2, if the border between two segments is amidst the cluster of the word occurrences. Reduced frequency 2 means that the word occurs in 1 or 2 clusters, but not 3. We can make similar statements for other small integers.

As the value of the reduced frequency depends on the beginning of the first segment and there is no firm reason to start always at the first position, we use "average reduced frequency" (ARF) instead. In order to explain the definition of ARF, we assume for a moment that v is an integer. The formula derived under this assumption is then also used in the general case, when v is not an integer.

The ARF is the arithmetic average of the reduced frequencies of the word over all possible beginnings of the first segment. It is sufficient to consider only the first v positions of the corpus as a possible beginning of the first segment, because if the first segment starts at a position j, the reduced frequency would be the same as if it started at any of the positions $j + v, j + 2v, \ldots$. Thus, we can assume that the position j belongs to the first segment, in other words, $1 \leq j \leq v$. For $j = 1, \ldots, v$, let RF_j be the reduced frequency, if the first

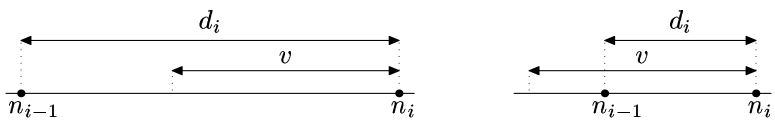


Figure 1: Long and short interval between word occurrences.

segment starts at the position j. Clearly,

$$ARF = \frac{1}{v}\sum_{j=1}^{v} RF_j. \tag{2}$$

In order to calculate RF_j according to its definition, we need to determine for all the segments that start at the positions $j, j+v, j+2v, \dots, j+(f-1)v$, whether they contain an occurrence of the word or not. Thus, in order to calculate $\sum_{j=1}^{v} RF_j$, we need to consider all the segments that start at all positions $1, 2, \dots, N$. For a moment, choose an occurrence n_i of the word and let us examine the group of the segments that start at every position of the left interval corresponding to n_i. There are d_i such segments, since there are d_i positions in the left interval of n_i. The contribution of these segments to $\sum_{j=1}^{v} RF_j$ depends on the distance d_i as follows. If $d_i \leq v$, then all d_i segments of the group contain the position n_i and contribute to the sum. If $d_i > v$, only v segments of the group contribute. For an illustration of a situation with $d_i > v$ and $d_i < v$, see figure 1.

Altogether,

$$\sum_{j=1}^{v} RF_j = \sum_{i=1}^{f} \min\{d_i, v\}$$

and, hence,

$$ARF = \frac{1}{v}\sum_{i=1}^{f} \min\{d_i, v\}. \tag{3}$$

As mentioned above, this formula will be used to define ARF in the general case, although the above explanation used the assumption that v is an integer. Hence, let ARF be defined by formula (3).

The basic properties of ARF are the same as those of the reduced frequency. If the word occurs in one small part of the corpus, its ARF would be slightly higher than 1, depending on the span between the first and the last occurrence of the word within the cluster. The smaller the span, the lower value of ARF. If the cluster is large, ARF would be higher. If the word occurs in 2 clusters, its ARF would be higher than 2, and so on for other integers.

Average waiting time

For every position of the corpus, let the "waiting time" be the number of positions that have to be read starting at the given position in order to hit the first consequent word occurrence. We assign the waiting time to every position of the corpus. For the positions inside the left interval corresponding to an occurrence n_i, the waiting time achieves values $d_i, d_{i-1}, \ldots, 2, 1$.

"Average waiting time" is the arithmetic average of waiting times of all corpus positions:

$$AWT = \frac{1}{N}\sum_{i=1}^{f}\sum_{j=1}^{d_j} j = \frac{1}{N}\sum_{i=1}^{f} d_i(d_i+1)/2.$$

Thus, using (1)

$$AWT = \frac{1}{2N}\left(N + \sum_{i=1}^{f} d_i^2\right) = \frac{1}{2}\left(1 + \frac{1}{N}\sum_{i=1}^{f} d_i^2\right).$$

Any word with frequency 1 has the AWT equal to $(N+1)/2$. AWT of words with higher frequency depends on their distribution within the corpus. If all the occurrences of the word are placed in a small part of the corpus, its AWT is close to $(N+1)/2$, even if its frequency is high. For more evenly distributed words, the AWT decreases.

Average logarithmic distance

Contrary to the definition of waiting time, where we assigned a different value to every position of the left interval corresponding to a word occurrence, in this case we assign the same value – "logarithmic distance" $\log_{10} d_i$ – to all the positions of the left interval corresponding to the occurrence n_i. "Average logarithmic distance" ALD is then the arithmetic average of logarithmic distances of all the positions within the corpus:

$$ALD = \frac{1}{N}\sum_{i=1}^{f} d_i \log_{10} d_i.$$

Any word with frequency 1 has the ALD equal to $\log_{10} N$. For more frequent words the value of ALD depends again on clustering of their occurrences in the corpus. A word that occurs in one small cluster has the ALD close to $\log_{10} N$, even if it has quite high frequency. More evenly distributed words have the ALD smaller.

The formula for ALD resembles a formula for the entropy. Indeed, ALD may be defined using entropy as follows. Consider a probability distribution on the f occurrences of the word such that for $i = 1, \ldots, f$ the probability of n_i is $p_i = d_i/N$. The entropy of this distribution is

$$H = -\sum_{i=1}^{f} p_i \log_2 p_i \tag{4}$$

and we have $ALD = \log_{10} N - H/\log_2 10$. The entropy H increases, if the distribution with the probabilities p_i gets closer to the uniform distribution. Hence, ALD decreases, if the word is distributed more evenly.

Definition of the corrected frequencies

In this section, we define three corrected frequencies corresponding to ARF, AWT, ALD. Let w be any word and let M denote any of the measures ARF, AWT, ALD for the word w. Then, let f_M for w, the corrected frequency of w with respect to M, be the "frequency" of an evenly distributed word that has the same value of M as w. We have put the word frequency into quotation marks, since we allow this quantity to be a non-integer in order to obtain smooth functions in the formulas. For each possible M, the corrected frequency f_M may easily be expressed using the value of M. Let us present the formulas for the individual cases.

For any word w, let $ARF(w)$, $AWT(w)$, and $ALD(w)$ be the values of the corresponding measures for the word w. Let a word w with frequency f in a corpus of length N be given. We are looking for the "frequency" f' of an evenly distributed word w' such that its respective measure has the same value as the same measure of the word w. The following equalities should be fulfilled.

$$\begin{aligned} ARF(w') &= f' \\ AWT(w') &= \frac{1}{2}(N/f' + 1) \\ ALD(w') &= \log_{10}(N/f') \end{aligned}$$

Using these formulas, one can solve the equations $ARF(w') = ARF(w)$, $AWT(w') = AWT(w)$, $ALD(w') = ALD(w)$ in the unknown f' and obtain the formula for f_{ARF}, f_{AWT}, and f_{ALD}. We come to the following definition, based on formulas obtained in this way.

Definition 1 For any word w let $f_{ARF}(w)$, $f_{AWT}(w)$ and $f_{ALD}(w)$ be defined as follows

$$\begin{aligned} f_{ARF}(w) &= ARF(w), \\ f_{AWT}(w) &= \frac{N}{2\,AWT(w) - 1}, \\ f_{ALD}(w) &= N \cdot 10^{-ALD(w)}. \end{aligned}$$

Sometimes we omit the word w from the notation, if the word follows from the context. Moreover, we use the notation ARF instead of f_{ARF}. Using (3), one can express ARF directly from d_i. In order to express f_{AWT} and f_{ALD} directly from the distances d_i, one can use the formulas

$$f_{AWT} = \frac{N^2}{\sum_{i=1}^{f} d_i^2} \tag{5}$$

and

$$f_{ALD} = \exp\left(-\sum_{i=1}^{f} \frac{d_i}{N} \ln \frac{d_i}{N}\right). \tag{6}$$

Note that we also have $f_{ALD} = 2^H$, where H is the entropy defined by (4).

The Czech National Corpus and the new measures

We calculated all three characteristics on the real data from the Czech National Corpus, which contains 100,054,133 tokens. As Czech is a flective language with a great number of word forms creating a lot of lemmas, we usually work with lemmas rather than word forms themselves. So did we in this case too. In the rest of the article, word will always mean lemma.

There are more than 330,000 different lemmas with frequency greater than 1. For the calculations we took only lemmas with frequency at least 5. This reduced the number of lemmas to 174,313.

Graphs 1,2,3 in Fig. 2 show the relationships between frequency f of words and their corrected frequencies f_{ARF}, f_{AWT}, f_{ALD} respectively. For every measure M among ARF, AWT, ALD, the corresponding graph consists of a set of points corresponding to individual words in the corpus. For a word with frequency f and corrected frequency f_M, the corresponding point has horizontal coordinate $\log_{10} f$ and vertical coordinate $\log_{10} f_M$.

The sets of points in all three graphs have a similar shape. There is an area containing a lot of points corresponding to small frequencies. This area is "wide" in the vertical direction, which means that in this area, one can find words with the same frequency, but quite different corrected frequencies. This demonstrates the presence of words with the same frequency but different distributions in the corpus.

More evenly distributed words are placed near the upper edge of the set of points in all three graphs. These are the words that appear in a majority of texts, not only in small clusters. Words that occur only in small clusters have smaller values of the corrected frequencies. Hence they are placed below the upper edge of the set of points. There is a "bottom line" at every graph, depicting words with the lowest values of the corrected frequencies. These are very close to 1. This means that each of these words occurs in one very small cluster of the corpus (all its occurrences fall into a section not exceeding 2% of the corpus size). The frequencies of the words at this line do not exceed 1000.

For high frequencies, the sets of points are "thin" which means that the corrected frequencies of words with the same frequency do not differ much. It can be shown analytically that this can happen only for words which are distributed quite evenly throughout the whole corpus and do not occur in clusters.

Clustering of words is best visible in graph 1, describing the relation between f and f_{ARF}. The horizontal lines depict words occurring (from the bottom) in 1, 2, 3, 4 clusters. For higher integers the lines merge with other points and are not distinguishable easily.

Experimental comparison of the stability of the measures

The corpus is just a sample of the language, which is used to make conclusions about the whole language. Hence, it is important to which extent our measures are stable with respect to the selection of the corpus. For this purpose, we have to compare the values of the corrected frequencies of the same word in different corpora. Although all three corrected frequencies of a word with pure frequency f have values in the same interval $[1, f]$, there is a systematic bias. For example, on average, ARF is 1.2 times larger than f_{ALD}. Since the intended application of the corrected frequency is to rank words, the actual values of the corrected frequencies are not important, if the ordering of the words according to these values is known. In order to eliminate the influence of the bias, we compare the ordering of words instead of the concrete values of their corrected frequencies.

In order to estimate the stability of the ordering, we have split the whole corpus into five disjoint subcorpora. We tried to preserve approximately the proportion of different styles in the subcorpora. For each of the five subcorpora and each of the three corrected frequencies, we consider a list of words from the subcorpus sorted in decreasing order according to the value of the selected corrected frequency. We considered only words having at least 2 occurrences in each subcorpus. For each of these words and each of the three corrected frequencies, we have five indices. We denote them $\text{index}_{ARF,i}(w)$, $\text{index}_{AWT,i}(w)$, $\text{index}_{ALD,i}(w)$, where $i = 1, \dots, 5$ is the number of the subcorpus.

The stability of a given corrected frequency on a given word is measured as the difference between the maximum and the minimum among the five indices of the word. We call this difference the range. The range of ARF for a word w is

$$\text{range}_{ARF}(w) = \max_{i=1..5} \text{index}_{ARF,i}(w) - \min_{i=1..5} \text{index}_{ARF,i}(w).$$

We define $\text{range}_{AWT}(w)$ and $\text{range}_{ALD}(w)$ in an analogous way.

We divide the words into several groups and compare the ranges of the corrected frequencies in each group separately. The words are divided into the groups depending on their pure frequency in the corpus and the five subcorpora as follows. Since the subcorpora have slightly different sizes, we use the pure frequency normalized to a million tokens. The pure frequency per million tokens in the whole corpus will be denoted $\bar{f}(w)$, and the pure frequency per million tokens in each of the five subcorpora will be denoted $\bar{f}_i(w)$ for $i = 1, \dots, 5$. Moreover, let $\bar{f}_{max}(w) = \max_{i=1...5} \bar{f}_i(w)$ and $\bar{f}_{min}(w) = \min_{i=1...5} \bar{f}_i(w)$. The words were grouped together if they have similar values of both $\log \bar{f}(w)$ and $\log \bar{f}_{max}(w) - \log \bar{f}_{min}(w)$. More exactly, the interval containing the values of $\log \bar{f}(w)$ for all considered words was split into 19 subintervals of equal length. Independently, the interval of the values of $\log \bar{f}_{max}(w) - \log \bar{f}_{min}(w)$ for these words was split into 14 subintervals of equal length. Two words belong to the same group if they fall into the same interval in both considered parameters.

Results of the comparison are presented in graphs 4,5,6 in Fig. 3. Each of the graphs corresponds to a pair of the corrected frequencies. The graph is a map consisting of 14 times 19 squares, each of which corresponds to one group. These groups are the same in all graphs. The axes in the graphs are labeled with the two parameters used to split the words into groups.

The groups are marked by different hatching, which shows, for each graph separately, which of the two compared corrected frequencies is more stable for the words in the group. The algorithm determining the marking was as follows.

For each group, let n_1 (respectively n_2) be the number of words in the group, for which the first (respectively the second) of the compared corrected frequencies has the smaller range of the corresponding index. For example, in each group of the graph 4 comparing ARF and f_{ALD}, we have:

1. n_1 is the number of the words w for which $\text{range}_{ARF}(w) < \text{range}_{ALD}(w)$;
2. n_2 is the number of the words w for which $\text{range}_{ARF}(w) > \text{range}_{ALD}(w)$.

Empty squares in the graphs correspond to empty groups. Squares corresponding to nonempty groups are marked by the name of the corrected frequency, which is more stable for the words in the group. Using the definition of significant difference described below, each nonempty square in the graph 4 is marked by

- ARF, if n_1 is significantly larger than n_2;
- f_{ALD}, if n_2 is significantly larger than n_1;
- "indif", if the difference between n_1 and n_2 is not significant.

The other two graphs comparing ARF versus f_{AWT} and f_{ALD} versus f_{AWT} are constructed in a similar way.

The exact definition of what is a significant difference between n_1 and n_2 was inspired by a statistical test for a binomial distribution. In other words, we formally consider n_1 and n_2 as the number of positive and negative results of $n_1 + n_2$ independent coin flipping with probability p of the positive result. If the numbers n_1 and n_2 are such that it is possible to reject the hypothesis $p \leq 0.5$ at the 5% confidence level, we consider n_1 significantly larger than n_2. Similarly, if we can reject $p \geq 0.5$, we consider n_1 significantly smaller. If it is not possible to reject any of the hypotheses $p \leq 0.5$ and $p \geq 0.5$ at the 5% confidence level, we consider the difference insignificant.

Let us describe the results presented in the graphs. The graph ARF versus f_{ALD} shows that for words with small variation between pure frequency per million tokens in different subcorpora, ARF is more stable than f_{ALD}. This follows from the fact that the groups at the bottom part of the graph are marked by ARF. Since the y-coordinate in the graph corresponds to the variation of pure frequency per million tokens between subcorpora, these groups contain words with low value of this variation.

On the other hand, for words having large differences between frequency per million tokens in different subcorpora, f_{ALD} is more stable. This follows from the fact that the nonempty groups at the top part of the graph are marked by f_{ALD}. Words with large variation of the pure frequency are more problematic and the graph shows that f_{ALD} should be used for them.

The graph ARF versus f_{AWT} shows that the relationship of f_{AWT} and ARF is similar to that of f_{ALD} and ARF. However, the area where f_{AWT} is more stable than ARF is smaller than the corresponding area in the graph f_{ALD} versus ARF. Hence f_{AWT} seems to be less stable than f_{ALD}.

This is verified in the third graph. There are groups with no significant difference between f_{AWT} and f_{ALD} and there are also groups where f_{AWT} is less stable. There is no group where f_{AWT} is more stable than f_{ALD}.

Examples

In this section, we present the values of corrected frequencies for a few concrete words. Let us introduce another characteristic for every word – clustering numbers V_j. For any $j = 1, \ldots, f$, let V_j be the sum of its j largest distances d_i. We include V_j for $j = 1, \ldots, 4$ in our tables, since these parameters allow us to determine the number of clusters in which the word occurs, if it occurs in at most 4 small clusters.

If the word appears in one small cluster, then the largest distance is close to the length N of the whole corpus, since we take it cyclically. The remaining d_i's are small, because the distances within the cluster are negligible compared to the largest distance. It follows that V_1 is close to N and the remaining V_j's are only slightly larger than V_1.

For words occurring in exactly two clusters, V_1 is the greater distance between the two clusters (taken cyclically) and V_2 is close to N. The remaining V_j's are not much higher than V_2. Similar statements can be made about V_j for other values of j.

Clustering of words is (naturally) typical especially for words with low frequency. However, there are words with quite high corpus frequency that occur in clusters, too. Those are very often proper names (for instance names of novel heroes) or special terms.

Let us have a look at the characteristics of the three examples from the introduction to this article - see Table 1.

word (in Czech)	f_{ARF}	f_{AWT}	f_{ALD}	V_1 (%)	V_2 (%)	V_3 (%)	V_4 (%)
souvrství	12.01	3.03	4.01	44.98	79.05	86.03	92.42
vzrušující	412.41	229.97	358.60	1.65	3.13	4.61	5.75
krůček	490.80	338.14	469.11	1.12	2.15	3.08	3.93

Table 1.

Table 1 presents the corrected frequencies and the numbers $V_1, \ldots, V_4$ for the three words mentioned in the introduction. Recall that these words have the same frequency, 891. The numbers V_j are expressed in percent of the size of the whole corpus.

The word "souvrství" has remarkably low value of all three corrected frequencies, since it is present in several (not more than 12, because f_{ARF} is slightly greater than 12) limited sections of the corpus. This uneven distribution is clearly visible from the values of V_j. For example, the value V_1 means that there is a continuous section of the corpus of size 44.98% which contains no occurrence of the word. The other two words have corrected frequencies much greater, which means that they are distributed more evenly. Correspondingly the V_j are much smaller.

Table 2 shows another example of a collection of words with the same frequency, but very different distribution in the corpus. The table presents the corrected frequencies and the numbers V_j expressed in percent of the size of the whole corpus for the words "gliom" (gliom), "taoista" (taoist), "kočenka" (a special word, see below), "akupresura" (acupressure), "meruňkový" (apricot as an adjective), "stojánek" (small stand), "mlhavě" (misty, foggy), and "martyrium" (agony, suffering), which all have the same frequency 137.

word (in Czech)	f_{ARF}	f_{AWT}	f_{ALD}	V_1 (%)	V_2 (%)	V_3 (%)	V_4 (%)
gliom	1.03	1.00	1.00	99.97	99.98	99.98	99.98
taoista	3.03	1.94	2.28	66.95	91.54	99.97	99.98
kočenka	4.64	1.22	1.55	90.16	94.70	98.63	99.53
akupresura	19.41	11.51	14.09	16.16	27.31	38.24	48.40
meruňkový	35.50	21.17	27.91	9.50	18.70	25.30	31.73
stojánek	51.51	29.23	42.09	7.71	14.72	21.10	26.50
mlhavě	74.45	55.32	74.34	4.86	9.11	12.96	15.84
martyrium	80.22	51.25	75.55	6.44	10.96	14.89	18.69

Table 2.

All the words in the table 2 are ordered according to f_{ARF}. "Gliom" is a very special medical term, which is present in one text only. Note that the corresponding corrected frequencies are close to 1 and the values V_j do not differ much from the length of the whole corpus. The word "taoista" occurs in three texts only. Its uneven distribution is again easily distinguishable from the values of corrected frequencies and also V_j. The word "kočenka" means a small cat in a local dialect. As the Czech National Corpus contains a lot of novels and stories of Bohumil Hrabal, who liked to use this word, "kočenka" has quite high frequency. However, our characteristics reveal immediately that it is not common at all. For a comparison, we included in the table five other words, which are more common than the first three. This fact can easily be recognized on the basis of any of the three corrected frequencies as well as the numbers V_j.

Acknowledgements. The first author was supported by the Grant Agency of the Czech Republic, grant No. 201/00/1482. The second author was partially supported by the grant of Ministry of Education of the Czech Republic No. MSM 112100002.

References

Collins Cobuild English Dictionary (1995). HarperCollins Publishers Ltd.

The New Oxford Dictionary of English (1998). Oxford University Press.

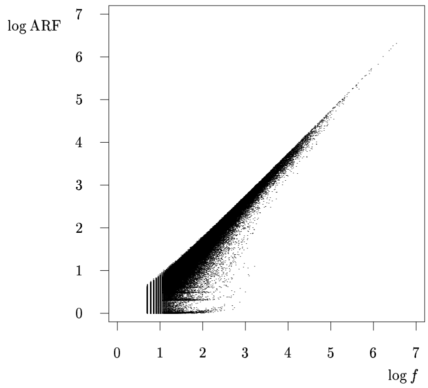
Carroll, J.B. & Davis, P. & Richman, B. (1971). *The American Heritage Word Frequency Book.* Boston, MA: Houghton Mifflin.

Hlaváčová, J. & Rychlý, P. (1999). "Dispersion of Words in a Language Corpus". In *Proc. TSD'99*, Springer-Verlag Berlin Heidelberg, 321–324.

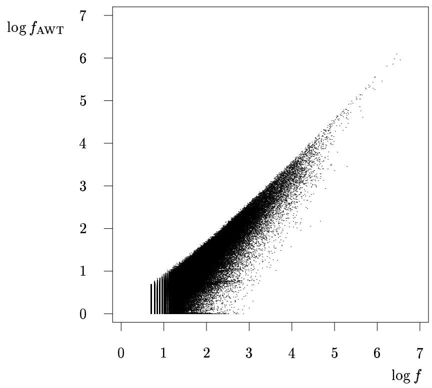
Hlaváčová, J. (2000). "Rarity of words in a language and in a corpus". In *Proc. LREC 2000*, Athens, Greece, 1595–1598.

Králík, J. (1978). "On the dispersion and its computation". In *Prague Studies in Mathematical Linguistics*, Prague, Academia, 149–158.

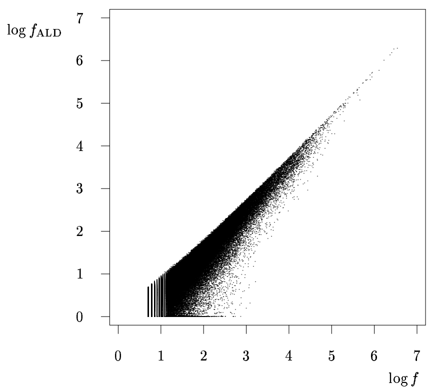
Oakes, M.P. (1998). *Statistics for Corpus Linguistics*, Edinburgh University Press.



Graph 1: $\log ARF$ versus $\log f$

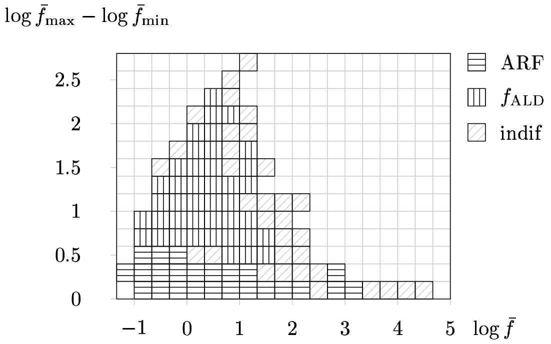


Graph 2: $\log f_{AWT}$ versus $\log f$

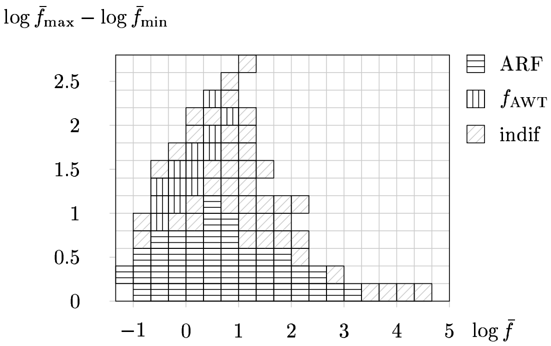


Graph 3: $\log f_{ALD}$ versus $\log f$

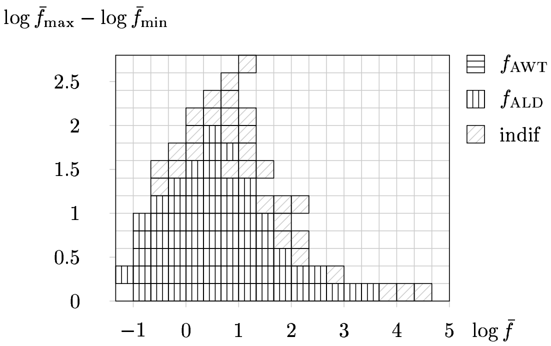
Figure 2: Scatter plots of ARF, f_{AWT}, and f_{ALD} versus f in logarithmic scale.



Graph 4: ARF versus f_{ALD}



Graph 5: ARF versus f_{AWT}



Graph 6: f_{AWT} versus f_{ALD}

Figure 3: Comparison of stability of ARF, f_{AWT}, and f_{ALD}.