

Rarity of words in a language and in a corpus

Jaroslava Hlaváčová

Institute of the Czech National Corpus

Faculty of Arts, nám. J. Palacha 2, Prague, Czech republic

jaroslava.hlavacova@ff.cuni.cz

Abstract

A simple method was presented last year (Hlaváčová & Rychlý, 1999) allowing to distinguish automatically between rare and common words having the same frequency in a language corpus. The method operates with two new terms: reduced frequency and rarity. The rarity was proposed as a measure of word rareness or commonness in a language.

This article deals with the rarity a bit more deeply. Its value was calculated for several different corpora and compared. Two experiments were done on the real data taken from the Czech National Corpus. Results of the first one prove that reordering of texts in the corpus does not influence the rarity of words with a high frequency in the corpus. In the second experiment, rarity of the same words in two corpora of different sizes is compared.

1. Introduction

It is difficult to distinguish which words in a language are common and which are rare. We can help with a language corpus and compare frequency of words in it. Those with a low frequency in the corpus are very probably rare in the language. However, not all words that have high frequency in the corpus have also high frequency in the language. It can happen that there is a text included in the corpus that contains extraordinary many instances of a rare (in the language) word. It is the case of some terms, proper names or special words invented by an author that were used in one literary work only.

2. Definitions

We can assign to every word of a corpus the unique number - its position within the whole corpus.

Let N designate number of words in the corpus. In other words, N is the position of the last word in the corpus.

Let $f(x)$ be the frequency of the word x in the corpus. For the next considerations I will call it **pure frequency**.

Let us for every word x divide positions of the whole corpus into $f(x)$ equal intervals.

The first interval includes positions from the beginning of the corpus to the number $[N / f(x)]$ (where $[k]$ designate the whole part of the expression k). We can express the first interval by means of mathematical notation:

$$\langle 1, [N / f(x)] \rangle.$$

The second interval is:

$$\langle [N / f(x)] + 1, [2N / f(x)] \rangle.$$

...

And the last $f(x)$ -th interval:

$$\langle [(f(x) - 1) N / f(x)] + 1, N \rangle.$$

In general the i -th interval is

$$\langle [(i - 1) N / f(x)] + 1, [iN / f(x)] \rangle$$

for all $i = 1, \dots, f(x)$.

Let us define the partial frequencies of the word x :

$Fx(i) = 1$, if the word x occurs (at least once) in the i -th interval,

$Fx(i) = 0$, otherwise (if the word x does not occur in the i -th interval).

Now we define the **reduced frequency** of the word x very simply as the sum of all the $f(x)$ partial frequencies:

$$r(x) = \sum Fx(i) \quad \text{for } i=1 \dots f(x).$$

Rarity $R(x)$ of the word x is defined as the quotient of the pure and reduced frequencies:

$$R(x) = f(x) / r(x).$$

3. Discussion – what the rarity can tell us

If a word was distributed entirely evenly in the whole corpus, its reduced frequency would be equal to its pure frequency. The special case are words with the frequency $f(x) = 1$.

The possible value of a word rarity depends on its frequency. The rarity of a word is always at least 1 and at most equal to its frequency. In mathematical notation: $R(x) \in \langle 1, f(x) \rangle$.

Thus, low frequency words have always low rarity. Otherwise it is not true, frequent words can have a low rarity. Low rarity of a frequent word implies even distribution of the word within the corpus. If the corpus was representative, the combination of high frequency and low rarity means that the word belongs to a common vocabulary, in other words that it is not rare in the language.

This conclusion is impossible to make on the basis of frequency only, because there can be (and there are) words with high frequency, that are present in a small section of the corpus only. Their distribution is uneven, so they have low reduced frequency, and so their rarity is high.

A high rarity of a word always implies very uneven distribution of the word within the corpus. Such words, even if they are frequent in the corpus, are not common in the language.

For examples from the Czech National Corpus see (Hlaváčová & Rychlý, 1999).

4. Experiments

If we think about the rarity, a number of questions can arise concerning the legitimacy of such a measure. Here are two of them:

- How much will reordering of the texts within the corpus influence the rarity value for individual words?
- How large must the corpus be in order that the rarity calculated on it could describe properly distribution of words within it?

I have made two experiments with the textual data from the Czech National Corpus that will answer the above questions:

- 1) I calculated rarity of words for different corpora that differed in the order of involved texts only and compared them.
- 2) I compared rarity of words between two Czech corpora of very different sizes (100 mil. word forms vs. 1 mil. word forms).

4.1. Reordering of texts

I took a number of texts and calculated the rarity of all the word forms from the texts. Then I made two other corpora out of the same texts, but always in different order, and calculated the rarities of the word forms.

I have applied this method twice, for fiction and for newspaper texts. (I shall refer to the fiction corpus by letter F, to the newspaper corpus by N.)

The both corpora have approximately the same number of word forms. The table 1 shows some statistics about them.

corpus	F	N
number of word forms	15 122 793	15 750 084
number of different words	484 717	453 896
number of words with $f(x)=1$	209 132	197 347

Table 1: description of the two experimental corpora

The table 2 shows 12 most frequent word forms from the corpus F and their three rarities for the both corpora. We can see at least two interesting things:

1. The rarities within one corpus do not practically differ.

word form	in English	F - corpus				N - corpus			
		frequency	r_1	r_2	r_3	frequency	r_1	r_2	r_3
a	and	590 034	1.51	1.51	1.51	369 575	1.54	1.54	1.54
se	reflexive pronoun or preposition with	451 515	1.56	1.56	1.56	281 397	1.60	1.60	1.60
na	at	260 525	1.62	1.62	1.62	272 233	1.61	1.61	1.61
v	in	213 923	1.64	1.65	1.64	367 705	1.63	1.63	1.63
to	it	213 424	1.76	1.76	1.76	86 844	1.81	1.80	1.81
že	that	187 429	1.72	1.72	1.72	148 678	1.80	1.80	1.80
jsem	(I) am	156 369	2.53	2.53	2.53	26 657	2.79	2.79	2.79
je	is	156 332	1.78	1.78	1.78	142 380	1.74	1.74	1.74
si	reflexive pronoun	122 594	1.71	1.71	1.71	53 020	1.77	1.77	1.77
s	with	112 318	1.67	1.66	1.67	115 447	1.66	1.66	1.66
do	into	112 199	1.68	1.68	1.68	99 291	1.72	1.71	1.72
ale	but	104 661	1.62	1.63	1.62	56 796	1.74	1.73	1.74

Table 2: 12 most frequent Czech words and their rarities in the two experimental corpora

2. Even though there are quite great differences between the frequencies of word forms in the both corpora (this could be also interesting topic, but it does not belong to this discussion), the rarities do not differ much.

However, these observations are true only for very frequent words in the corpora.

Let r_1 , r_2 and r_3 are the three rarities calculated on one corpus. I calculated absolute values of differences between every pair of them: $|r_1 - r_2|$, $|r_2 - r_3|$ and $|r_1 - r_3|$ for every word form of the corpora. I will call them r-differences. In the corpus F, among the word forms with frequency > 400 (this means among the 3 453 most frequent word forms) there are only 61 words with at least one r-difference greater than 1. All of them have very high rarities (the smallest rarity is 13.91), 59 of them are proper names, mainly foreign ones. They appear mainly in novels translated into Czech.

If we take into account words with lower frequency, the number of words with a great r-difference raises and values of appropriate rarity lower.

In the corpus N, there are only 3 word forms with the frequency > 400 (there are 4 696 such word forms) that have at least one r-difference greater than 1. And there are only 45 words with at least one r-difference greater than one among words with frequency higher than 100 (16 170 word forms). It is interesting that only half of them are proper names or abbreviations of proper names. This shows that individual proper names in newspapers are much more evenly distributed than proper names in fiction.

We can conclude this discussion with the following proposition:

The order of individual texts within the corpus has a minimal influence on the rarity of frequent words.

This conclusion justifies the use of rarity for measuring a real language rarity (in the sense of uncommonness or rareness) of words that have high frequency in the corpus.

Of course it is not possible to utter a similar proposition about words with low frequency.

4.2. Size of the corpus

Does the size of the corpus influence the rarity of words?

To answer this question, the rarity of lemmas was computed for two Czech corpora that differ in size. One of them was CNC - the Czech National Corpus, with 100 million word forms, the other one was DESAM - the manually desambiguated corpus from the Masaryk University in Brno (Pala et al., 1997) with only 1 million word forms. It must be stated here, that the two corpora did not differ only in size but in the content, too. DESAM was created with regard to a representativeness, while CNC was not balanced at the time of calculations (now it is, but the appropriate data are not yet available in the form needed for similar calculations). The great majority (about 80%) of the texts in CNC formed newspapers.

The table 3 shows the most frequent lemmas from the bigger corpus.

word	in English	frequency		rarity	
		CNC	DESAM	CNC	DESAM
a	and	2 809 239	26 126	1.59	1.57
v	in	2 665 425	27 239	1.66	1.66
být	to be	1 689 681	28 037	1.80	1.65
na	at	1 632 182	16 807	1.66	1.66
z	from	862 104	8 713	1.71	1.70
že	that	849 759	8 516	1.87	1.82
o	about	770 509	7 163	1.81	1.78
který	which	765 873	8 071	1.68	1.66
s	with	747 057	9 596	1.70	1.65
do	into	596 828	5 894	1.75	1.75
i	and	553 161	6 186	1.76	1.74
on	he	550 230	5 842	1.91	1.82

Table 3: Comparison of rarities from two corpora of different sizes

Again we can see that the most frequent lemmas have almost the same rarity, even when their frequencies in the corpora differ very much.

However, like in the first experiment, this is true only for the most frequent lemmas. With lower frequency, number of lemmas with greater differences between the two rarities raises. Among the 1000 most frequent lemmas from CNC there are 45 lemmas in DESAM the rarity of which differ by more than 1 from the rarity in CNC.

Table 4 lists some examples of frequent words with different rarities as explained in the previous paragraph.

We can finish the discussion about the influence of corpus size on rarity of words (lemmas) with a similar proposition like in the first experiment:

The rarity of very frequent lemmas is not influenced much by the size of the corpus.

However, again similarly to the results from the first experiment, nothing can be stated about words with low frequency.

If the contents of the two corpora were proportionally more similar, the agreement between the alternative rarities would probably be greater.

word	in English	frequency		rarity	
		CNC	DES.	CNC	DES.
Kč	abbr. of Czech crown	55 601	426	5,16	4,10
byt	flat	29 815	331	3,45	4,80
ČTK	Czech Press Agency	28 202	32	5,51	16,00
utkání	a match	23 348	2	4,25	1,00
zápas	a match	19 275	380	3,64	4,94
ty	you	17 943	44	5,14	1,91
odstavec	paragraph	14 342	90	9,34	3,10
vyhláška	public notice	13 712	57	5,00	2,85
román	novel	11 555	14	3,97	2,33

Table 4: Examples of different rarities in CNC and DESAM

4.3. Notes about the calculations

There are two ways how to calculate positions within a corpus. We can count only word forms or we can take into account punctuation, too. In the latter case the positions of the corpus are not occupied by word forms only, but also by punctuation marks. The first type of calculation was used in the first experiment, the second one in the experiment number 2.

I have not carried out the two types of calculations on the same data, so I do not know how it would influence the results. The rarities calculated for the most frequent Czech words in the two experiments described above (tables 2 and 3) are not directly comparable, even if the most frequent words in all the corpora involved are practically the same. It has two main reasons: In the first experiment word forms were counted while in the second one lemmas were counted. And secondly, types of the corpora were different. However, from an "approximate," comparison it follows, that the difference would not probably be very big. Another experiment should be processed to show which alternative describes the word rarity better – with punctuation marks or without them.

Also calculation of rarity for punctuation marks themselves could bring some interesting results.

5. Summarizations and some more reflections

I tried, on the basis of Czech textual data, to answer two questions concerning the legitimacy of using rarity as a measure of word rareness in a language. The results show (not surprisingly) that it is reasonable to consider only frequent words.

The first one was: "How much will reordering of the texts within the corpus influence the rarity value for individual words?"

The rarity of frequent words is not influenced much, but rarity of words with a low frequency can change significantly. We can even say: the more frequent word in the corpus, the more stable its rarity.

It would probably be better to make the same experiment once more, but this time with lemmas instead of word forms.

The second question was: "How large must the corpus be in order that the rarity calculated on it could describe properly distribution of words within it?"

This question is difficult to answer. We have seen that rarities calculated on very small corpus of about 1 million word forms showed nice agreement with the rarities calculated on the corpus 100 times larger, but this was true mainly for very frequent words. Words with lower frequency differed more often, not to mention, that some of them were not even found in the smaller corpus. More data give certainly the greater guarantee of reliability.

Table 5 shows very roughly what can be said about a word if it has any of four possible combinations of frequency and rarity in a language corpus.

		Frequency	
		high	low
Rarity	high	rare word	
	low	common word	?

Table 5: Combinations of frequency and rarity

Although the experiments were made only with Czech data, from the definitions it is clear, that calculation of rarity is language independent. It would be interesting to see, how much its values differ for the most frequent words in other languages. Do the language equivalents have the same rarity in different languages?

Maybe, that the rarity could serve as another tool for distinguishing senses of synonyms (see the example of two possible translations of the English word "match" in the table 4)...

6. Acknowledgement

This research was supported by the GACR, Grant Nr. 405/96/K214.

7. References

- Hlaváčová, J., Rychlý, P., 1999. Dispersion of words in a language corpus. Proc. TSD'99, Lecture Notes in Artificial Intelligence 1692, pp.321–324. Springer - Verlag Berlin Heidelberg New York ISBN 3-40-66494-74
- Pala, K., Rychlý, P., Smrž, P., 1997. Desam – approaches to disambiguation. Technical Report FI-MU-97-09. Faculty of Informatics, Masaryk University, Brno.