Corpus Frequency and Lexicographical Relevancy Czech Words with a Morfem Micro- (in Hundred Million Corpus of Czech Language - SYN2000)

Michal Šulc
Institute of the Czech National Corpus
Faculty of Arts
Charles University Prague
n. J. Palacha 2
Praha 1
116 38
michal.sulc@ff.cuni.cz

Abstract

Using the example of a Czech morpheme "micro", the paper shows several possibilities for how to decisively use language features to incorporate the item into a draft of dictionary-entry list. Quantity of occurrences, quantity of lemmas, frames of word-formation, parts of speech and especially quantity of lemmas inside frequency ranks are discussed. The result of the paper is that we can use a graph showing the quantity of lemmas inside frequency ranks as linguistic evidence and make the boundary for the draft entry list accordingly.

What is Corpus SYN2000?

SYN2000 is a representative 100 million-word corpus of electronic texts which was systematically collected at the Institute for Czech National Corpus (ICNC, Prague) between 1994 and 2000 as a reference source for the purposes of scientific study of language and for the future compilation of practical linguistic reference books. It comprises texts published between 1960 and 1999.

SYN2000 uses a uniform SGML format and is linguistically marked using stochastic methods (lemmatised and provided with morphological tags). Its composition reflects the current position of the ICNC staff on representativeness, based on studies from different areas and accessible statistical data. The composition of the corpus is as follows: journalism 60 %, imaginative literature 15 %, informative literature 25 %. The 25 % informative literature breaks down as follows: life style 5.55 %, technology 4.61 %, social sciences 3.67 %, natural sciences 3.37 %, art sciences 3.48 %, economy and management 2.27 %, law and security 0.82 %, faith and religion 0.74 % and administration 0.49 %.

What the results of a research of corpus frequency could be?

Since 1960s, when the age of corpora began, linguists have used such large amounts of language materials that they can be counted in tens or hundreds of millions of words. They have also used such effective software tools that hundreds and thousands of occurrences can

be seen and processed in a second. All this linguists can do for the first time ever. This enables them not only to solve several older questions and problems, but also to discover completely new questions and also kinds of questions.

Questions about a corpus frequency (of course together with other questions) are now taking up a central position because of the current possibility of dealing with huge collections of texts. By corpus frequency, I do not mean a T-score or a MI-score, which are nowadays frequently offered as common functions of linguistic software tools. What I mean is frequency of tokens, the related frequency of types, and a frequency of lemmas.

In this paper I will concentrate on lemma frequency.

Calculation of corpus frequency should offer measurable guidelines for interpreting the communicative importance of expressions and in that way help lexicographers compile lists of dictionary entries.

At this stage, questions concerning building lists of dictionary entries may include the following:

- *What corpus lemma frequency is of sufficient lexicographical relevance in order a word to be included into the first draft of a list of monolingual dictionary entries?*

- *Some morphological forms (with specific endings) occur more frequently within some lemma's while other do not. The obvious question to ask is what lexicographical relevancy, if any, do some of these morphological forms have, if viewed against the background of total of occurrences of this form in the whole of the corpus. Can this relevancy be measured by frequency?*

- *What frequency of a collocate or its proportional representation among other collocates is lexicographically relevant?*

Corpus linguistics attempts to solve several of the questions mentioned above and partial results are already available – mainly in the field of collocates.

In this paper I will be attempting to determine which lemma frequency is lexicographically relevant. The main question I will be trying to answer is: Can we find a point on the frequency axis, or several points for several types or subtypes of nomination, that will mark the border or borders of lexicographical relevancy? If we can, this information would be useful for the compilation of the first drafts of a list of dictionary entries. These drafts could eliminate much of the routine work for linguists and could be easily further processed (that is either shortened or expanded) by linguists.

I do not assume that lemma frequency will be the only decisive feature in compiling drafts of lists of dictionary entries in the future. This is because frequency of expression in communication is not necessarily always in harmony with notional-structural importance. Notional-structural importance that should also be considered in the compilation of monolingual dictionaries. High frequency is not the only condition for putting an expression into dictionary. The operation of this principle was demonstrated by several hyperonyms or terms, which occurred below expected frequency in corpus. Articulate animals (articulates) –

in the Czech language "členovci" – occurs in SYN2000 with the frequency of 74, while crayfish, even though a very rare animal occurs with the frequency 30 times higher (frequency of 2908).

Currently corpus linguists are working on a wide variety of problems and questions. But what about the future? Languages with the most developed corpus linguistics have already compiled megacorpora. Therefore questions concerning the lexicographical relevant part of corpus, and the automatic extraction of this part from the entire corpus especially, are going to gain in importance greatly.

This paper deals with the automatic extraction from a corpus and does so within a broader context.

What was chosen as a research field

The purpose of my research was to investigate corpus frequency of one limited area of the lexicon. For this purpose I chose words (compounds and their derivatives) in which the first part was formed by the morfeme MICRO-. This collection of words was investigated in corpus SYN2000.

First I should draw to your attention several facts, that will clarify the position of the morfeme MICRO in the frame of the lexicon as a whole. In reviewing my results it will be useful to have these facts in mind.

Even though frequency of any expression is principally affected by its communicative importance, we could expect its word-formation, the level of acclimatisation of borrowings and other factors to be influential. It is also important to keep in mind, that MICRO- is a morfem borrowed from foreign languages. Through its meaning, a morfeme is predestined to function in terminology of different – not necessarily related fields. Consequently, its functioning in non-terminological fields is limited.

Therefore, the conclusions that will be made here will be valid primarily for compounds with similar features, but, it is hoped, that the reasoning developed here could also apply to other language situations.

Research of corpus frequency and lexicographical relevancy

Approach A – Quantity of occurrences

In this approach my goal was to find out: What percentage of occurrences is it necessary to include?

(proceeding from the most frequent lemmas toward frequency 1)

	occurrences	lemmas	frequency of lemmas of micro-
total number:	11338	579	1 to 1177
% occurr.	**Num. of occurr.**	Num. of lemmas	Frequency of lemmas
95 %	**10771**	178	4 and more
90 %	**10204**	78	11 and more
85 %	**9637**	45	32 and more
80 %	**9070**	31	48 and more
75 %	**8504**	22 (less than 4 %)	75 and more

Table 1:Quantity of lemmas (and their frequencies)
according to the quantity of occurrences, expressed in percentages

However, why should one arbitrarily choose just 95 % as the starting point? Why not 94 % or 96 %? I think it is not possible to answer the question and, at the same time, remain in the group of responsible lexicographers. The point is that similar approach depends on culturally-established ideas about round numbers, symmetry and rhythm. It is not based upon any linguistic theory of lexicon; they do not use any linguistic feature that could be somehow measurable. From a linguistic point this is an unacceptable view.

Approach B – Quantity of lemmas

The goal of approach B was to find out: What percentage of lemmas is it necessary to include?

(proceeding from the most frequent lemmas to frequency 1)

	lemmas	occurrences	frequency of lemmas of micro-
total number:	579	11338	1 to 1177

% lemmas	Num. of lemmas	Num. of occurr.	frequency of lemmas
5 %	**29**	8977	53 and more
10 %	**58**	9925	18 and more
15 %	**87**	10286	9 and more
20 %	**116**	10496	6 and more
25 %	**145**	10639	4 and more

Table 2: Quantity of occurrences (and their frequencies) according to quantity of lemmas expressed in percentages

This approach is not only based on the frequency of tokens, but also on the notion of lemma, which is a linguistic abstraction. However, using B, the problem of approach A is not yet eliminated. The choice of 5 % or 20 % was made arbitrarily. No linguistic theory supports it. No measurable language feature would lead to such nice round numbers. Consequently, this approach is unacceptable from the linguistic point of view, too. For this reason, it is a good idea to begin with linguistic theories. We should take them as the basis for answering the following question: *What should the selection of lemmas take into account?*

Approach C – Frames of Word-formation

The communicative importance of an idea (expressed in words) is represented in language chiefly by simple frequency. This approach was seen in approaches A and B.
But there is a conceptual-structural importance of this idea, too. This could be represented by a variety of functional positions (parts of speech) that the idea could take up or also by a necessity to connect it into one whole together with other ideas and therefore differentiate and enrich it (word-formation).

We can see how communicative importance (here frequency) does not correspond with conceptual-structural importance (here width of derivation line):

---

mikrobus = **minibus** (276) – mikrobusek = **little minibus** (2) – mikrobusový = **adjective refering to minibus** (1)

compare with

mikrobiolog = **microbiologist** (47) – mikrobiložka = **microbiologist woman**(1) – mikrobiologie = **microbiology** (74) – mikrobiologický = **microbiological** (175) – mikrobiologicky = **microbiologically** (15)

---

mikropočítač = **microcomputer** (814) – mikropočítačový = **adjective referring to microcomputer** (53)

compare with

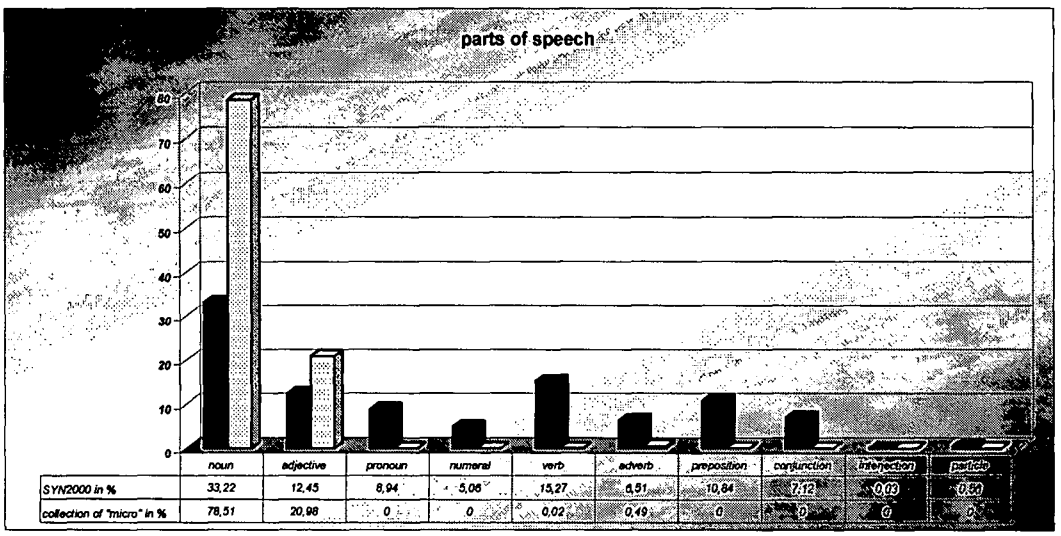
mikroskop = **microscope** (524) – mikroskopie = **microscopy** (57) – mikroskopický = **microscopical** (254) – mikroskopicky = **microscopically** (26) – mikroskopičnost = **microscopicallity** (1) – mikroskopový = **adjective refering to microscope** (1) – mikroskopik = **microscopist** (1) – mikroskopovací = **microscopiing /adjective/** (1) – mikroskopování = **microscopiing /noun/** (1)

If we want to study word-formation processes, we can quite easily obtain long lists of word forms with one common string of signs. The common string could be a stem morfeme as well as any other morfeme. But for processing this list of common strings, we cannot avoid demanding and often time consuming manual work. This is because words having a common string are not of the same kind as words from one derivational line. We cannot extract derivational lines from our corpus automatically, because we do not have such software tools yet. And it is not likely that we will have them in the foreseeable future.

Approach D – Parts of speech

As I have shown in approach C, it is difficult to use any theory concerning extending lexicons in corpus extraction. The situation with parts of speech is different. Copus SYN2000 is lemmatized and morphologically-tagged, so that we can work with parts of speech in a relatively efficient way. The question is whether this approach can help us in any way.

We cannot assume that the distribution of frequencies of parts of speech remains the same in all domains of a national language. Distribution of frequencies differs in formal, familiar, slang and technical language. I did a comparison of the distribution of parts of speech in corpus SYN2000 and in our small sample of micro-.



	noun	adjective	pronoun	numeral	verb	adverb	preposition	conjunction	interjection	particle
SYN2000 in %	33,22	12,45	8,94	5,06	15,27	6,91	10,64	7,12	0,03	0,30
collection of "mikro" in %	78,51	20,98	0	0	0,02	0,49	0	0	0	0

Graph 1: Ratio of parts of speech in the corpus SYN2000 and in the collection of mikro-

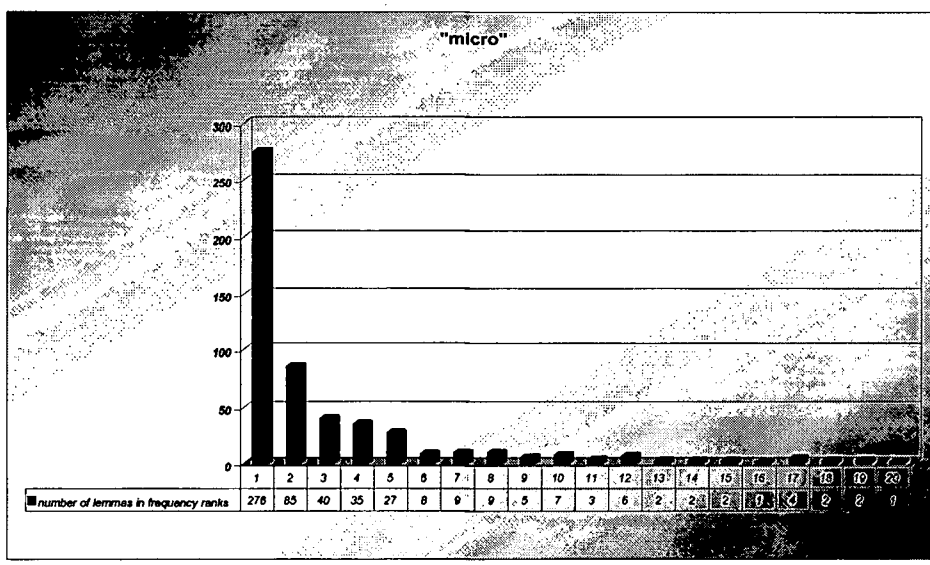
From the graph, it is clear that the word-formational field of *mikro-* is of nominal character. But. If we consider frequency of parts of speech in the whole corpus crucial to our work, we would incorporate all founded adverbs and verbs into the proposed dictionary. But verbs (*mikrovlnit, mikrodiseminovat, mikrofilmovat*) only occur once and to include all of them into a dictionary would be clearly improper.

On the other hand, I would like to point out the above mentioned consideration of the conceptual-structural importance of an idea. We should consider it not only a matter of semantics, but also a grammatically-categorial matter. From this standpoint we should carefully consider at least one verb with the morfeme micro-.
Graph 1 warns us about differences between the distribution of parts of speech in the whole corpus and in the subcorpus of micro-derivatives. It would also be useful to mediate this kind of information to dictionary-users at sometime in the future. Processing of subverbal lemmata and their part-of-speech-dependent frequencies would distinctly indicate language potentiality of the morfeme.

As far as corpus frequency and lexicographical relevancy is concerned, the parts of speech area is not significantly relevant – most likely with the minor exception of the grammatically-categorial connectivity mentioned above.

Approach E – Quantity of lemmas inside frequency ranks

In this approach, the goal is to find out: Is it possible /or sensible/ to say something about the boundary between a language's centre and its periphery by observing the quantity of lemmas inside frequency ranks?



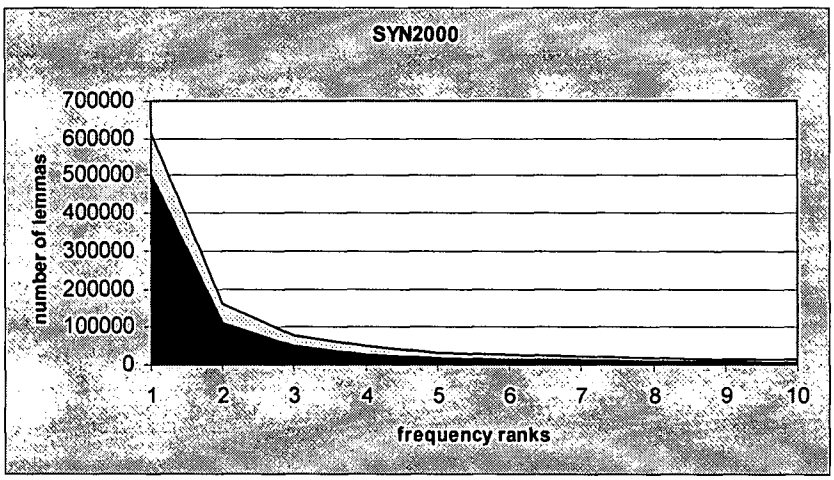
Graph 2: Quantification of lemmas from the same frequency rank in the collection of mikro-

In the graph there is an obvious zone of periphery (in my opinion with frequency 1 and 2) and a zone that we could name transitive (in my opinion with frequency 3, 4 and 5). Theoretical problems with defining the notion "core lexicon" should induce caution. We do not want to state that words with the frequency of 6 (and more) belong to a "core lexicon" with any certainty. But this may be so only because the notion of "core lexicon" has not yet been defined by corpus linguists. Possibly when corpus linguists define it, the definition will also provide a corpus frequency as a starting point.

Because it is not the goal of this article to define "core lexicon", the only thing I am suggesting is that to form a boundary for automatic extraction (from 100 million corpus) a graph which shows the increase in number of lemmas in frequency ranks would be useful. Words with frequencies of 6 (and more) might be included into background dictionary entry lists of new language dictionaries.

These lists may be rounded off according to some scheme other than frequency. If these 6-and-more-frequency words are put through a test where conceptual-structural importance is also taken into account, it seems possible, that some of them would be crossed out of the list. However, it is necessary to underline and to underline in bold that the graph deals only with separate words. Because it does not take note of very frequent type of collocative nomination it does not reflect all nominating processes, This type is disassembled and that fact slightly alters our results. Nevertheless we are aware of it and we still think that our conclusions are helpful. In case the nomination unit is binominal (or even longer), we can assume, that if the unit has communicative importance, both its parts (or all its parts) occur more than 6 times (in the same way as nominating units of one word).

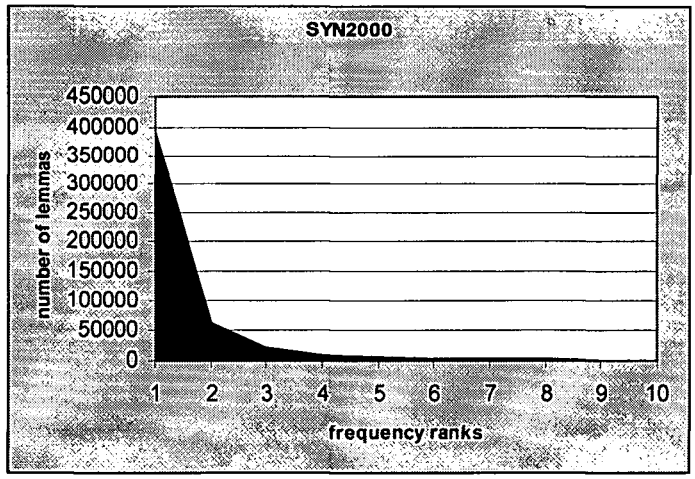
But we have to ask one more question. Is frequency boundary applicable to the whole corpus, or does it work only with compounds where the first part is made up of a foreign or loan word? Let us look at the graph made from the whole corpus SYN2000.



Graph 3: Quantity of lemmas of the same frequency rank in the corpus SYN2000
(The new growth above immediately preceding state.)

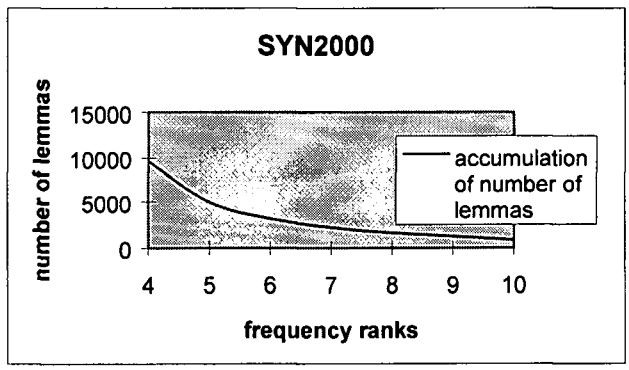
In the graph, the transition to a lexicon periphery seems slightly more gradual and does not offer any jumping-off point as graph 2 does. Let us examine then not only frequency, but also its accumulation. (In order to assess process the accumulation of numbers of lemmas in single ranks, we have to find out the relation between number of lemmas in the rank and number of lemmas in the preceding rank. This difference we will call "accumulation".)

From the progress of the graph we see that somewhere near frequency 5 or 4 both graphs begin to set themselves apart. When we isolate the accumulation of numbers of lemmas on a separate graph, we will see the separation more clearly.



Graph 4: Accumulation of number of lemmas in the corpus SYN2000
(The difference between number of lemmas in a rank and number of lemmas in immediately preceding rank.)

It is helpful to enlarge one part of the graph. Frequency ranks 4 to 10 are in detailed view below:



Graph 5 Accumulation of number of lemmas in the corpus SYN2000 – extract (4 to 10)
(The difference between number of lemmas in a rank and number of lemmas in immediately preceding rank.)

The graph shows that the accumulation of the numbers of lemmas progresses evenly until rank 6. The situation changes near rank 5. Hence, for the 100 million corpus we suggest the corpus frequency of 5 as the boundary for entering lemmas to draft lists of dictionary entries.

Comparison of approach E with The New Oxford Dictionary of English (morfeme micro-)

To show my approach E in a more comprehensible manner, I extracted from my original collection of "micro-" morpheme only lemmas with frequencies of 5 or more. I compared the numbers of these lemmas with the number of dictionary entries (containing "micro-") in The New Oxford Dictionary of English (NODE). The NODE is considered to be up-to-date.

In the corpus SYN2000 there was 143 lemmas containing morpheme "micro-" and occurring in frequency higher than 5. The New Oxford Dictionary of English contained 125 lexemes containing "micro-". I do not wish to assert that this method will lead to a similar result if applied to the British National Corpus (BNC). This means to such a number of dictionary-entries that go slightly beyond the number of dictionary-entries published in a dictionary. SYN2000 and BNC were built in different ways and typological differences of languages would play a certain role, too. Nor do I want to declare that the 143 automatically extracted lemmas will contain all of lemmas finally published in a dictionary.

But I think that the proportion between these two numbers enables us to assume that all lemmas relevant for dictionary making relevant from the point of view of frequency will be included. I believe that among the 125 lexemes found in NODE we can find also lemmas incorporated into the dictionary from the point of view of conceptual-structural importance. That slightly alters the above mentioned proportion so that the 143 lemmas will more probably include all frequently important lemmas. The difference between these numbers enables me too to say that dictionary-makers will have the possibility to make their own choice and in this way they can slightly influence the selection of the final dictionary-entry list. The number is also not so large as to flood the linguist with ballast whose manual inspection would be only a waste of time and energy, because the ballast will not be included in the dictionary at the end of the day.

Conclusion

According to overviews presented here, and as far as 100 million-word corpus is concerned, I recommend using a corpus frequency of 5 as the first entrance-frequency for the drafting of dictionary-entry list. This would guarantee to linguists the communicative importance of the item processed and would possibly also reflect the core lexicon of a nonspecialized written standard language.

When we are speaking about frequency and its importance we should also pay attention to the distribution of occurrences within a corpus [Savicky & Hlavacova to come]. Frequencies higher than 5 are not very useful when all occurrences are from one document. So, I hope, that the distribution factor will also be taken into consideration in preparing the draft.

Linguists could also add to the drafts by incorporating items about which we have spoken in this article such as about items of conceptual-structural importance. I believe that the need for such additions will be minimal. Finally, the lists will not be complete if we do not have any means for enlarging the draft by including items from the spoken language.

Caveat

My goal was to show how the automatic extraction of some basic information from a corpus could be useful towards making a certain stage of linguist's work quicker and more effective. Certainly, I did not want to excuse linguists from the responsibility for incorporating certain items into a dictionary-entry list or excluding them.

Additionally, it does not follow from this article that I consider the corpus frequency of 5 as sufficient basis for writing a dictionary entry.

References

Savicky, P., Hlavacova, J.: Measures of Word Commonness. (To come in *Journal of Quantitative Linguistics*.)