

How Large is the Core of Language?

Václav Cvrček, vaclav.cvrcek@ff.cuni.cz

Institute of the Czech National Corpus

Faculty of Arts, Charles University in Prague

ABSTRACT: This paper deals with the delimitation of the size of a lexical core by corpus methods. It suggests the usage of the proportion of hapax legomena (i.e. words that occur only once) to all word-types in relation to the growing corpus size to identify the frequency range in which core elements occur. In a hypothetically small corpus (a few sentences) the hapax-type ratio will be equal to one (each word-type is also a hapax). As we add texts to a corpus (up to a few million words), the hapax-type ratio decreases (the number of new words including hapaxes is continuously increasing but the majority of added tokens are new instances of words already present in the corpus) from its maximal value (=1) to the local minimum (between 0.35 and 0.45). This is the turning point (in a graph it is represented by a plateau). Thereafter, extending the corpus increases the ratio because the amount of hapaxes grows at a faster pace than the number of types with frequency higher than one. The graph of the hapax-type ratio (which has similar shape in different languages, regardless of the types of texts or their order) resembles a pipe or chibouque (hence “pipe-graph”).

This empirical finding tested on corpora of Czech, English and Italian brings us closer to an exact determination of the range of the core lexicon. Subsequently, we can deduce the approximate size of a corpus sufficient for compiling a dictionary that covers the core lexicon.

Key words: corpora size, hapax, type, language core, corpora design

1. Introduction

Defining a core of a language is an important question, not only for corpus linguists but also for lexicographers and language teachers. The concept of Basic English (Ogden 1930; Crystal 1997: 358), which was designed as a small subset of English vocabulary and grammar with the highest possible text coverage, proved that frequency of phenomena plays a crucial role in determining what the core elements are.¹

Corpus research is based on the hypothesis that large and representative collections of texts reflect the language reality truly and precisely. For example, in lexicography we assume that a general corpus is a sufficient source for representation of core (and to some extent also peripheral) lexical units. However, comparison with traditionally elaborated dictionaries shows us that there are still lexemes missing both in dictionaries or, less frequently, in corpora (even if we take into account fundamental differences in the size of dictionary and corpus). These differences are caused, not only by the fact of unintentional omission of common words in dictionaries or by mistakes in

¹ Text coverage for English, see Francis and Kučera (1982) or Waring and Nation (1997).

corpus design which result in insufficient representativeness, but also by different approaches to delimitation of the range of the core (i.e. where should we draw the border line in frequency list separating core elements and peripheral?).

It is important to emphasise that core of a language has been typically delimited from the perspective of a user (often a non-native speaker) for pedagogical reasons (Waring and Nation 1997, Laufer 2010) or for NLP applications² (Zhang, Huang and Yu 2004). In this paper, I would like to adopt more descriptive approach in which a core is not the smallest set of elements capable of fulfilling basic communicative needs but the part of language which should be described by small or medium-sized dictionaries. In this sense, the main attention shifts from language learners and their needs to lexicographers and corpus linguists, and especially to their problem of size of dictionaries and corpora.

Although we could speculate about different “cores” of language (morphological, lexical, syntactical), in this paper I would like to focus on the **core of forms** of language which can be (to a large extent) identical to the lexical core, especially in languages with little or no inflection like English. However, it is important to emphasise that this approach does not deal with a language core composed of meanings (this would require totally different approach to the data).

2. How to determine the core of a language – the intuitive way

At first glance, it might seem that the simplest way to determine the size of a core of a language is to count all elements which are common to all or to the majority of texts (or authors). The results of this approach might be surprising: in the BNC there are 4049 different texts (or text samples). There is only one word (!) which can be found in all of

² Core lexicon is defined in this approach as the (largest possible) set of words that can be expected to be used regardless of the environments. (Zhang, Huang and Yu 2004: 1121)

them (it is the punctuation mark “.”). Obviously, this is because some of the texts in the BNC are too short. But when we limit the data to texts longer than 1000 words, we found that these 3828 texts in the BNC share only 58 words.

These findings do not match our concept of the core of language suggested above. Such a small core is useless for lexicographic purposes. We assume that the core consists, at least, of all grammatical words (i.e. tens or hundreds of units) and some basic lexical words (i.e. hundreds or thousands of words). This means that an intuitively estimated core of language should consist of at least several thousands of elements (unlike the core vocabulary elaborated for pedagogical purposes which does not exceed hundreds of types³). Or, in other words, there may be two types of core with respect to their size: the first one is designed to be the smallest possible⁴ (for understanding and producing), while the second is based on observation of real usage.

Let us begin with the fact that each text has elements which belong to the core of a language and elements which are peripheral or text-specific. They are specific for many reasons: the individuality of the author of the text, or they are connected with the topic of the text, its genre or communicative situation, or they are archaisms or neologisms, etc. Each text has its own specific elements, whereas the core remains the same. In other words, two texts will most probably differ in peripheral elements and will share the core language phenomena.

3. How to determine the core of a language – corpus style

The experiment which can help delimit the size of the core of a language is quite simple and straightforward. It models the process of building a corpus during which we

3 Ogden's original suggestion (basic English word list) was based on 850 most important lexemes. (Ogden 1930)

4 Such a small core designed for pedagogical purposes could be based on observations of H-point, which is in the rank-frequency presentation the point at which rank is equal to frequency, the point nearest to the origin [0;0] (Popescu-Altmann 2006). This point roughly divides the lexicon into two parts: the most frequent synsemantic words and the rest – autosemantic branch.

observe the change of hapax-type ratio. At the beginning, the corpus consists of one sentence, e.g. (examples are taken from the BNC)

THE ROMAN WAY OF LIFE .

There are 6 tokens and also 6 types (including the full-stop). All the types are used only once therefore they are at the same time hapaxes; the hapax-type ratio is hence $6/6 = 1$.

When we add one sentence more from this text to our corpus we get:

The Roman conquest of Britain started in earnest in the year AD 43 and by about AD 120 , the whole of Britain up to Hadrian 's wall in the north became a Roman province .

We have 42 tokens in our corpus above, which consists of 33 types (the analysis is case-sensitive, i.e. *Roman* and *ROMAN* are two different types) with 7 of them being non-hapaxes (*the*, *in*, *Roman*, *of*, *Britain*, *AD*, .); the rest, i.e. 26 types are hapaxes.

Therefore, the ratio is $26/33 = 0.7879$. We then add another sentence.

Then , for more than 300 years until the legions left , the Romans ruled Britain , bringing their talent for organisation to the country , building roads which made travel easier , and all the time controlling the province through the administration of Roman law - backed up by the powerful legions of the Roman army .

Here, we have exactly 100 tokens with 65 types of which there are 50 hapaxes. It means that after adding 100 words the ratio dropped from an initial value (which is when counting the ratio after adding each token in all cases equal to one) to $50/65 = 0.7692$.

In the pre-corpus era, linguists would probably assume that by adding more texts the ratio would drop to a level which is equal to the proportion of hapaxes to all types in the lexicon. That would mean that the line representing the course of this ratio would have the shape of a big letter "L". Surprisingly however, it looks different:

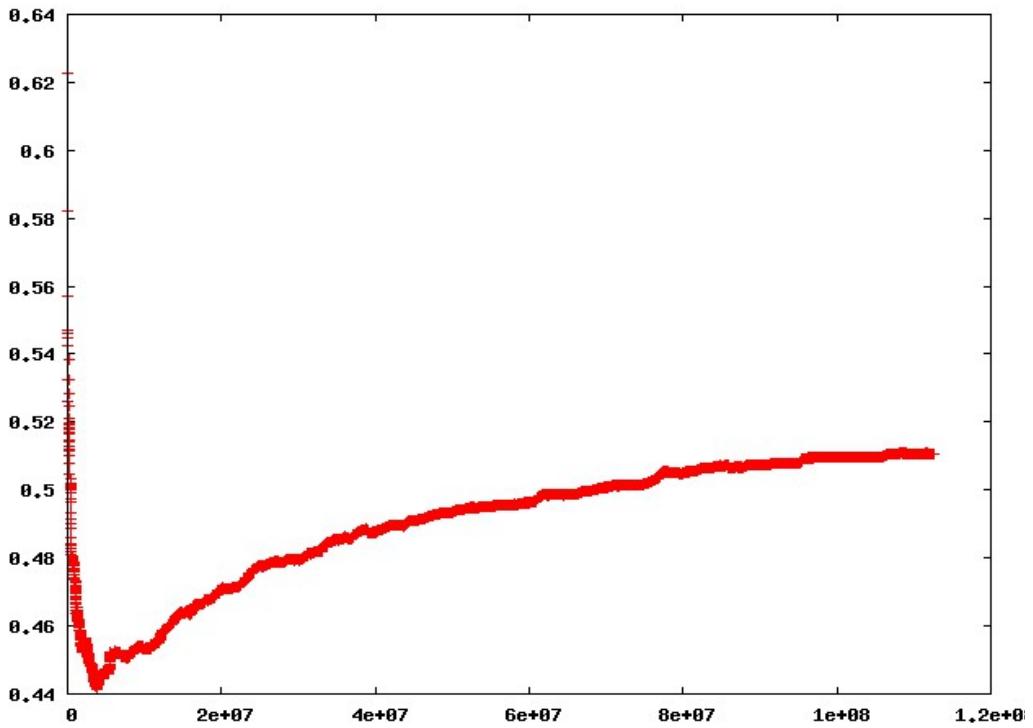


Fig. 1: Hapax-type ratio in BNC (types are case-sensitive word-forms). X-axis represents number of tokens added to corpus, Y-axis is hapax-type ratio.

As we add texts to the corpus (up to a few million tokens), the hapax-type ratio decreases from its initial and maximal value (=1) to the local minimum (which is between 0.43 and 0.47 for English word-forms). This is because although the number of new words – including hapaxes – is continuously increasing (during the whole process of adding texts to the corpus), the majority of added tokens in this phase are new instances of words already present in the corpus. This is the turning point (in a graph it could be represented by a small plateau) and from now on extending the corpus increases the ratio because the amount of hapaxes grows at a faster pace than the number of non-hapax types. Or, in other words, from that moment on with adding new texts the process of turning hapaxes into types with frequency higher than one slows down in comparison to the pace by which new hapaxes enter the corpus.

The graph of the hapax-type ratio resembles a pipe or chibouque (hence “pipe-graph”). Similar pipe-graphs can be found when examining the hapax-type ratio on English lemmas. The difference is only in the the position of the local minimal point and slightly higher value of the ratio (which is understandable because of the lesser variability of forms among lemmas in comparison to the number of different word-forms).

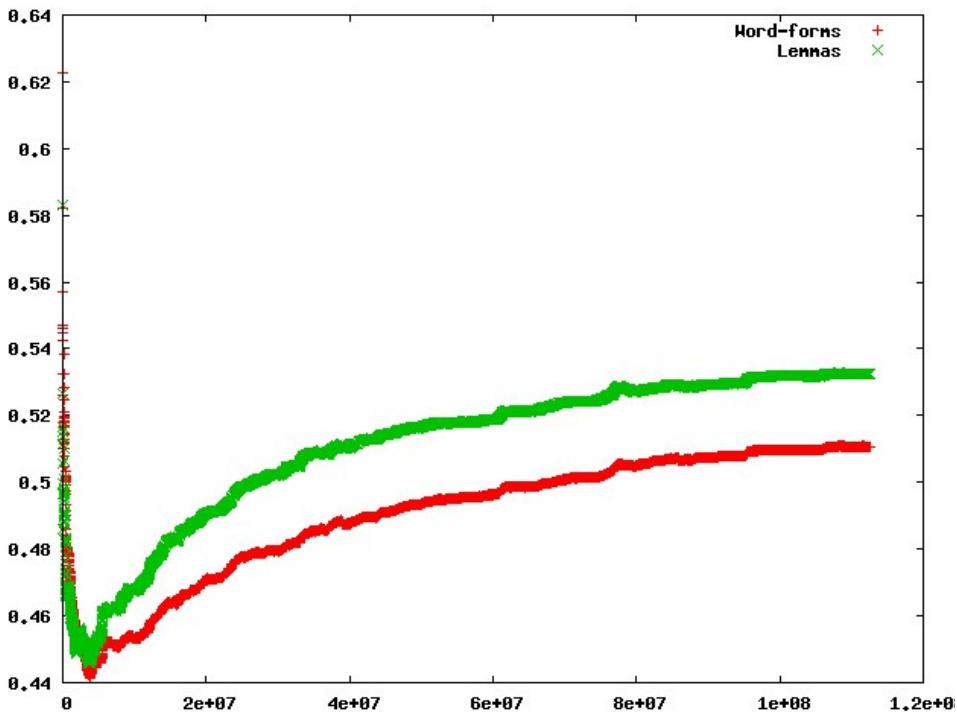


Fig. 2: Hapax-type ratio (y-axis) in BNC – difference between lemmas and word-forms.

Note that this is a type of phenomenon that was not observable in the pre-corpus era or in the corpus “stone-age”. Even Kučera and Francis, with their 1 million Brown corpus, could not observe this type of phenomenon, as the change in slope is visible only when the number of tokens surpasses 4 million. Moreover, this shape was not expected even by quantitative linguists, as all variables explored in that time (token-type ratio, hapax-token ratio etc.) showed constant and unvarying development (either constantly rising or constantly declining). In the case of the hapax-type ratio, a fundamental change in

development of the ratio occurs while increasing the amount of data.

The local minimal point shows us the turning point in the hapax-type ratio. We can see that in the case of both graphs (for word-forms and lemmas, fig. 2), the minimal point of hapax-type ratio is somewhere between 3,000,000 and 4,000,000 tokens (when the corpus exceeds this size the number of hapaxes starts to rise at a higher pace than the number of non-hapax types, therefore the hapax-type ratio grows). In a corpus of this size there are approximately 100,000 different word-forms and 70,000 different lemmas (for exact, but average (!), values see table 1 below). The first conclusion of this experiment thus could be that the lexical core of English consists of 70,000 most frequent lemmas or 100,000 most frequent word-forms. It is also obvious that in order to observe the phenomenon of hapax-ratio ratio tendency change, one has to have at one's disposal large amount of language data (at least tens of millions of words).

The phenomenon of change in slope of the hapax-type ratio function can also be explained by theoretical distinguishing between two types of hapaxes. The first type (let us call it "incidental") are words that occur only once in the corpus because of its insufficiently small size. This was the case of all words in the first added sentence (see above). On extending the corpus by adding more sentences, sooner or later all of these words (as they are quite common in English) become more frequent (i.e. non-hapax types). The second kind of hapax (we can call it "permanent") consists of words which are so rare that they occur only once no matter how large corpus we have at our disposal. The local minimal point in pipe-graph is the moment beyond which the process of turning incidental hapaxes into non-hapaxes slowly ceases when more text is added, and newly introduced types to our corpus are more likely to be permanent hapaxes (because incidental hapaxes are already represented in the corpus and are no longer hapaxes).

In order to obtain a better and more precise picture of the crucial turning point in the pipe-graph slope, we have to test what factors can potentially bias its shape.

So far, we have been testing our hypothesis that the core of a language can be determined by the exploring of change of hapax-type ratio, only with English. However, similar and sometimes even more decisive results can be found in Czech and Italian.

The results in Czech are quite decisive especially in the case of lemmas.

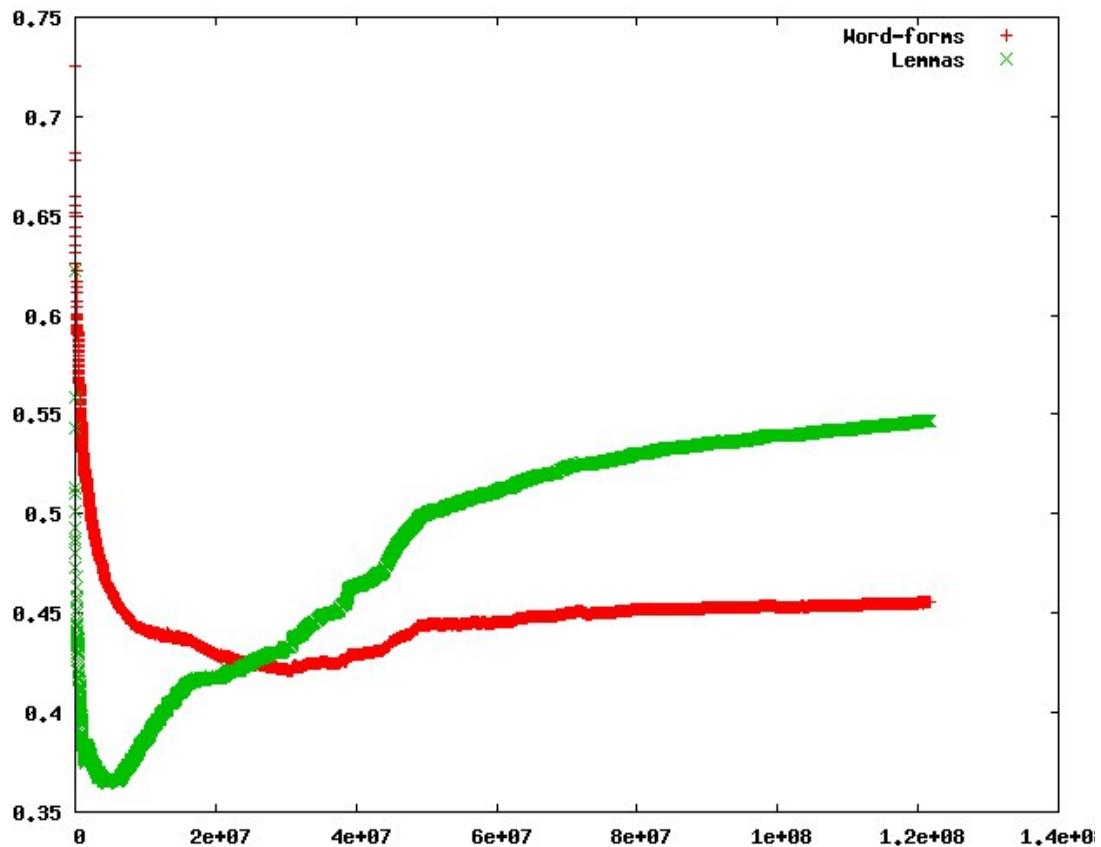


Fig. 3: Hapax-type ratio for Czech word-forms and lemmas. Source: SYN2010, representative and balanced corpus of contemporary written Czech.

The difference between Czech and English is caused by two main factors. 1) The difference between number of word-forms and lemmas is greater in Czech than in English. While there are approximately 2 word-forms for one English lemma, the ratio

in Czech is much greater; one lemma is assigned an average of 13 word-forms. 2) The greater text diversity of BNC and the sampling policy (no text is longer than 45,000 words, see <http://www.natcorp.ox.ac.uk/corpus/index.xml>) adopted in its creation, result in a smoother curve than can be observed in Czech. Corpus SYN2010 is based on whole texts and has three main parts (their boundaries are observable in the graph): fiction, scientific texts and newspapers.

Finally, for Italian, I have used an ad-hoc corpus of texts downloaded from the internet version of the newspaper *L'Arena*. This corpus was not lemmatised; I ran this experiment on word-forms only.

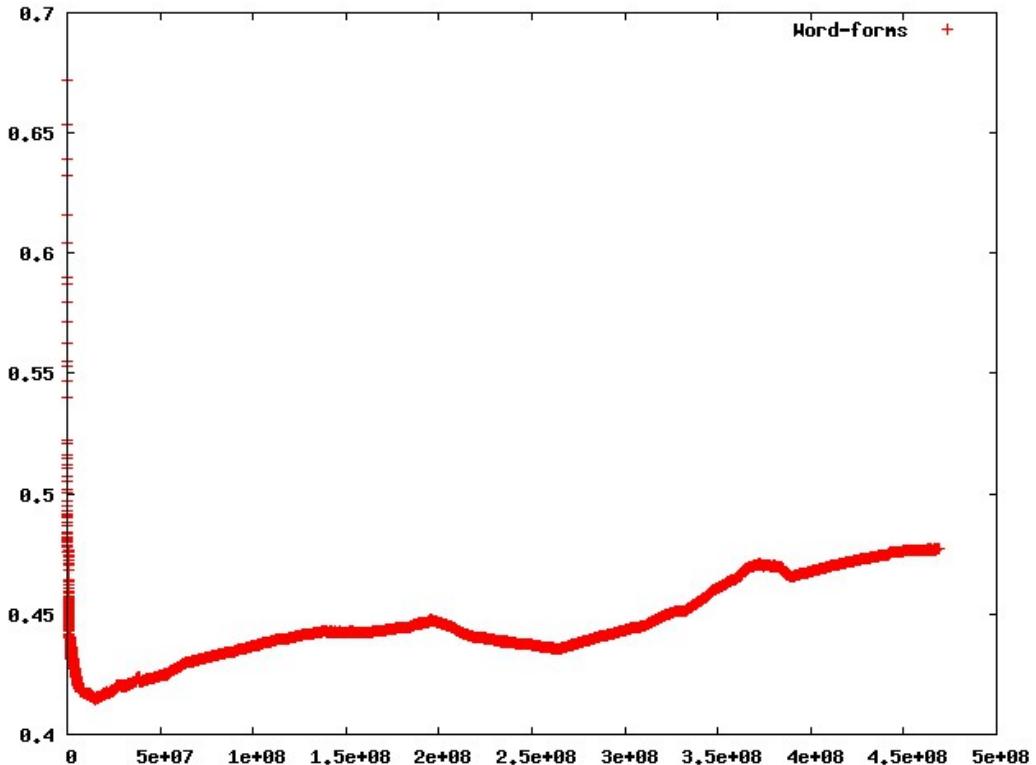


Fig. 4: Hapax-type ratio in the corpus of Italian newspapers. Source: *L'Arena*.

The graph shows that there are some other local minimal points. Their appearance can be explained by the fact that the texts in the corpus are from one source and are grouped in time spans; some areas of vocabulary (especially proper nouns of important people or

places) enter the corpus in short time spans which causes minor but abrupt changes in hapax-type ratio. Typologically, Italian is somewhere in the middle between Czech and English with respect to the richness of morphology.

Regardless of the morphological richness of the language, the phenomenon of the hapax-type ratio change during the process of adding texts remains the same in the examined languages. Nevertheless, the pipe-graphs above also show that the order of texts in the corpus is a rather important factor for the final shape of the curve. It is therefore necessary, for exact pinpointing of the local minimum, to randomize the text order to eliminate this influence. When this is done, we can also count the average shape of the curve, which now looks smoother for all languages and types of units.

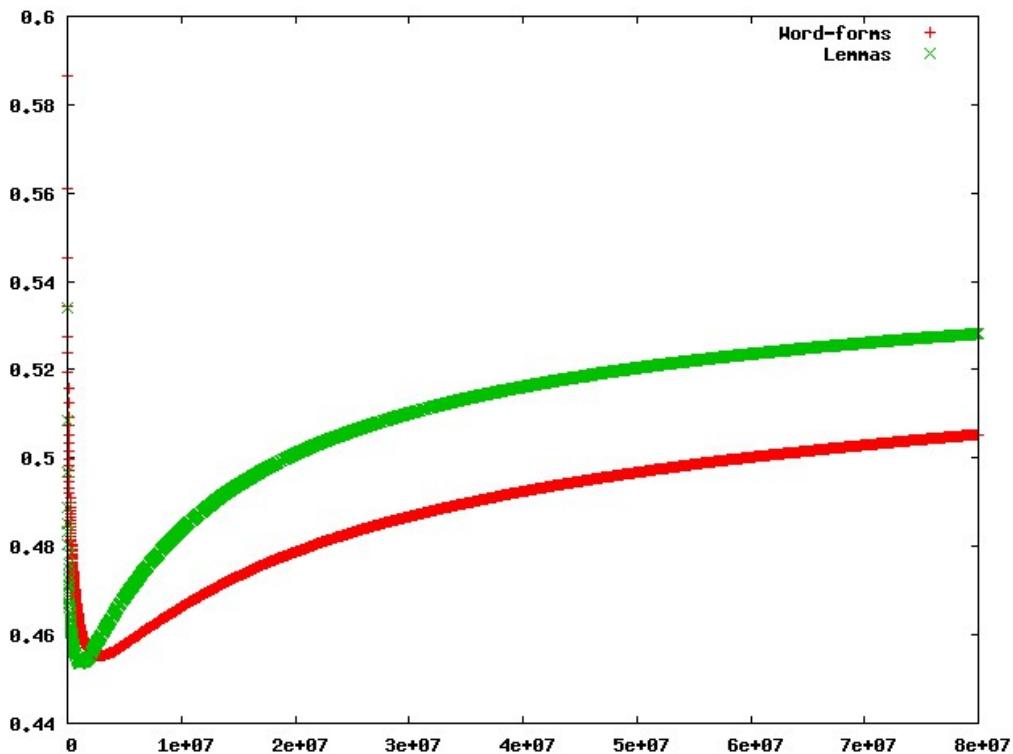


Fig. 5: Average values of hapax-type ratio for English lemmas and word-forms in BNC with randomized order of texts (number of random corpus compilations: N=600).

The situation in Czech (especially for word-forms) might look different, at first sight. But the truth is that the curve representing the hapax-type ratio for Czech word-forms has a local minimal point somewhere between 56 and 58 million tokens, and from that moment on it rises almost imperceptibly. The fact that the local minimal point for lemmas is far more closer to the beginning of the axes can be explained by the rich inflection of Czech (when we include differences in capitalization, some lemmas may be assigned to more than 200 different word-forms).

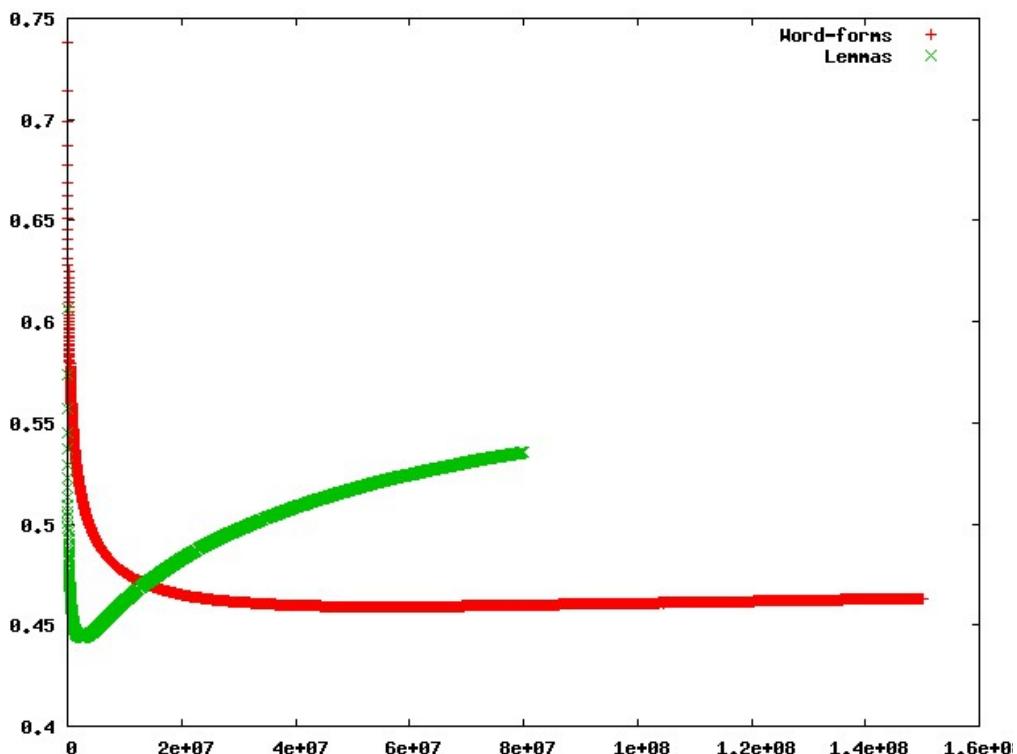


Fig. 6: Average values of hapax-type ratio for Czech lemmas and word-forms in SYN corpus with randomized order of texts (number of random corpus compilations: $N=100$).

The different positions of local minimal point for Czech word-forms and lemmas show us that different sizes of corpora are needed for different language units. The more variable the unit, the larger corpus we need. Therefore, for word-forms we have to have

a 22-times larger corpus than for delimiting the size of core for lemmas.

The average values for Italian look perfectly smooth, as the influence of the order of texts was completely neutralised by randomization.

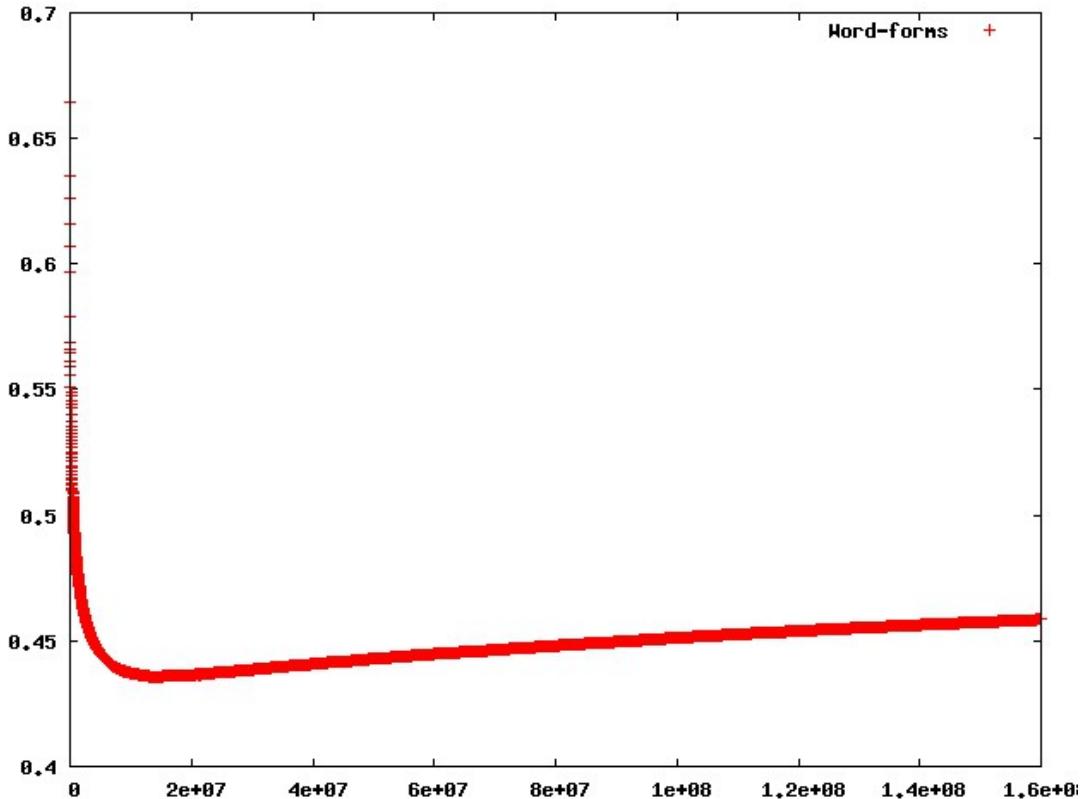


Fig. 7: Average values of hapax-type ratio for Italian word-forms in corpus L'Arena with randomized order of texts (number of random corpus compilations: N=100).

4. Results and interpretation

With randomised text order, we can draw some conclusions regarding the size of the core for each examined language. First of all we have to sum up the information about source data. For English, I used BNC XML edition corpus which has more than 100 million tokens and consists of written texts (or samples of them) of different genres (90 %) and also of spoken utterances (10 %). BNC is lemmatised and tagged. The Czech corpus SYN has 1.5 billion words and consists mainly of newspapers (almost 70 %) and

approximately 15 % fiction and 15 % scientific text of different genres, but it does not contain any spoken language. It is lemmatised and tagged. The Italian corpus based on web version of the newspaper L'Arena is 450 million words and is neither tagged nor lemmatised.

Important average values for each language are in table 1. In all cases, the hapax-type ratio was counted after adding each 10,000 tokens, the local minimal point was found as the smallest value of a floating average of ten consecutive hapax-type ratios.

Langauge	Type of units	Average minimal hapax-type ratio	Average hapax-type ratio at corpus size 80 million tokens	Number of tokens at minimal point	Number of types at minimal point	Number of random corpus compilations
English	word-forms	0.4492	0.5053	2 801 017	80 753	600
	lemmas	0.4409	0.5281	1 113 350	31 218	600
Czech	word-forms	0.4580	0.4591	58 183 030	1 091 966	100
	lemmas	0.4403	0.5354	2 642 727	86 157	100
Italian	word-forms	0.4332	0.4481	14 941 224	300 786	100

Table 1: Average results for each language and type of units.

It is important to emphasise that this experiment by its design does not account for the relatively common tendency (for example in English) to create multi-word units (see below). Numbers of types at the minimal point designating the size of the core are therefore not to be interpreted as the proposed number of entries in dictionary, only the number of core forms in language.

When we **compare the results** between lemmas and word-forms **within languages**, we can observe a not surprising relationship between the size of a corpus and variability of units. We need a larger corpus for reaching the turning point when counting with word-forms, in comparison to what is needed when counting the ratio for lemma-types. In other words, the size of a corpus covering all core elements depends on the variability of types used: the set of elements representing the core of language is

smaller in the case of lemmas than in the case of word-forms (this difference is especially noticeable in Czech, see fig. 6).

Comparison of the results between languages on the other hand shows us that these languages, although they are typologically quite distinct, share some basic features. Besides the fact that the shape of the curve is similar, the differences between values of hapax-type ratio at the turning point are quite low. It is thus likely that this could be some sort of general quantitative principle of large collections of texts.

5. Data versus mathematical model

Let us now look in more detail at the phenomenon of the pipe-graph. What causes this change in hapax-type ratio tendency and how can it be explained?

As mentioned above, there are core and peripheral elements in each language (for the case of this experiment we assume that hapaxes in texts are an accurate approximation of those peripheral units). The proportion of these two types of elements is different from language to language (it depends on formal variability, morphological richness, the tendency to create multi-word units instead of one-word expressions, etc.). Given that each language has its own proportion of core and peripheral elements to which the value of hapax-type ratio of each collection of texts should converge, we could imagine the pre-corpus concept of this phenomenon. In the time when large corpora were not as common as today, linguists probably assumed that the distribution of core and peripheral elements within texts is more or less even and equal to the proportion of hapaxes to types in the whole language. In this case, the hapax-type ratio curve would resemble a simple mathematical model which can be described as a mathematical sequence of two quantities (we can call them “M-types” and “M-hapaxes”).

Each sequence of each quantity is defined by constant but for each quantity different increment, e.g. 100 for M-types and 50 for M-hapaxes. The ratio of increments is therefore constant, $100/50 = 0.5$. Both sequences start from the same starting point, e.g. 1000. Before adding an increment, we count the ratio of the two quantities (M-hapaxes/M-types). At the beginning, the ratio is equal to one, as the starting point is the same for both of them. As we add more and more increments, the ratio between the two quantities converges to the ratio of their increments.

# of iterations	M-types	M-hapaxes	M-hapaxes/M-types
(starting point)	1000	1000	1.0000
1	1100	1050	0.9545
2	1200	1100	0.9167
3	1300	1150	0.8846
4	1400	1200	0.8571
...
990	100000	50500	0.5050

This model can also be described by the following formula:

$$\lim \frac{1000 + 50n}{1000 + 100n} = \lim \frac{1000/n + 50}{1000/n + 100} = \frac{1}{2}$$

which proves that the importance of the starting point slowly fades away in favour of the proportion of increments. If we adopt this formula for the specific purposes of hapax-type ratio, we obtain this formula (where n is the total number of tokens in the corpus):

$$\lim \frac{1 + \text{average_increment_of_hapaxes_per_token}}{1 + \text{average_increment_of_types_per_token}} = \frac{\text{average_increment_of_hapaxes_per_token}}{\text{average_increment_of_types_per_token}}$$

The comparison of the mathematical model with the reality derived from the data is shown below (the values of increments used for generating this graph were altered so that their proportion is equal to the hapax-type ratio of English word-forms, which is

0.51089):

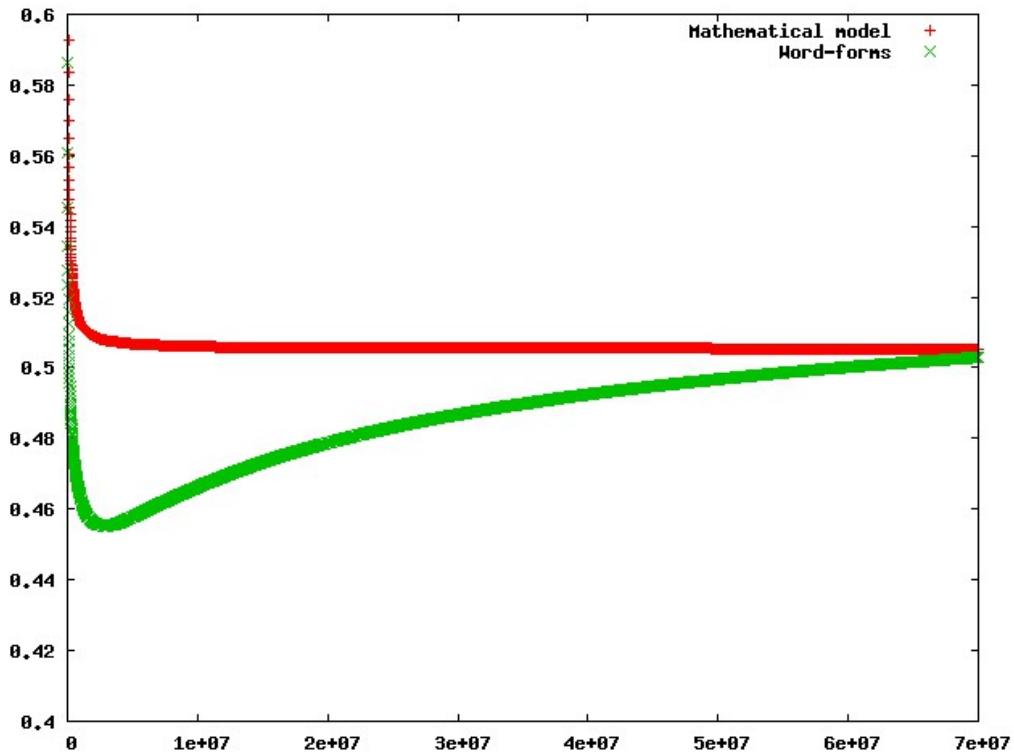


Fig. 8: Mathematical model versus hapax-type ratio of English word-forms.

Clearly, the observed reality is different from the model which represents the situation in which with adding each new text the hapax-type ratio constantly converges from the starting value (=1) to the proportion which is in the language as a whole. There is no turning point in the model because the function is constantly decreasing. This difference is caused by the fact of uneven distribution of peripheral⁵ elements in texts.

Whereas language core elements are (more or less) the same in different texts, low frequency elements (not necessarily only hapaxes) are those which differentiate texts among themselves. When we start to build our corpus, all elements are hapaxes regardless of the fact whether they are (with respect the whole language) core or peripheral. As we add more texts, some of the former hapaxes become more frequent

⁵ It is important to emphasise, that peripheral elements in this experiment are approached by hapaxes, but if we dig deeper in the problem we shall also consider other low frequency words (e.g. with frequency up to 5) which are still peripheral and tend to be text-specific (see below).

types. These are mainly the core elements, because these are the type of elements that most texts usually share. On the other hand, elements which are text-specific remain, after enlargement of the corpus remaining peripheral with frequency close to (but not necessarily equal to) one.

The mathematical model described above is based on the assumption (empirically rejected) that the distribution of peripheral units in texts is even. We can model this distribution not only by sequences of variables but also by randomizing the order of words or sentences in the corpus. When we thoroughly mix up the words in the corpus we should get exactly the same even distribution of core and peripheral units across the data, as was assumed by the mathematical model. Let us look at the same experiment in which we do not randomize the order of texts in corpus, but instead we add sentences or words randomly picked from the whole corpus.

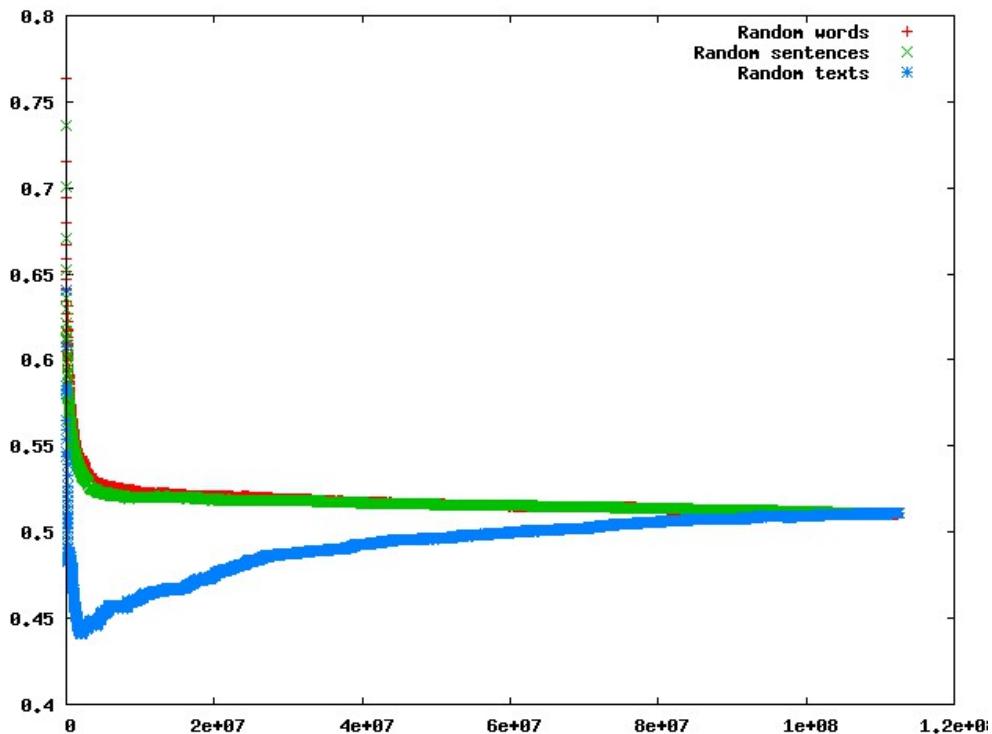


Fig. 9: Hapax-type ratio of English word-forms – random order of texts, sentences and words.

The smaller units we are using for randomizing the more the graph resembles the mathematical model (with no turning point). This is mainly because of the fact that the peripheral elements added to the corpus are from all possible parts of language. With the random order of texts, the corpus was enriched by peripheral words e.g. from the genre of love-story and then from economics. Conversely, a random order of sentences or words combined all genres, topics and author peculiarities together in a steady and constant proportion. The difference between the results of these two types of randomisation is in those low frequency types which are not hapaxes. These words tend to be text-specific and therefore they usually occur in one document or genre, e.g. 30 % of Czech word-forms (types) with frequency of 2 occurrences are to be found in one text, 39 % of Czech word-forms with frequency of 3 occur in less than three documents, etc. With a completely mixed order of words or sentences, there is no way to recognize the point at which almost all of the core elements cease to be hapaxes, and the new types entering the corpus are more likely to be peripheral and text-specific hapaxes than core elements.

Another important factor is that a corpus with a random order of sentences (or words) has higher initial values of hapax-type ratio than a corpus sampled from texts. This is because in the former case sentences are from completely different genres and domains; their diversity is thus much higher in comparison to diversity between sentences of one text (the initial text of the corpus). As a consequence, the ratio forms a constantly decreasing function with no turning points.

The situation in this aspect differs in English and in languages with rich morphology, especially Czech. Although the graph of hapax-type ratio in the corpus with a random order of sentences or words resembles the mathematical model described above, these graphs still have some properties of non-random corpora (the declining

part, the local minimal point, the plateau and, finally, the part in which the hapax-type ratio grows). But even so, the difference between graphs made with a random order of words or sentences and graphs made by randomizing text order is substantial and decisive.

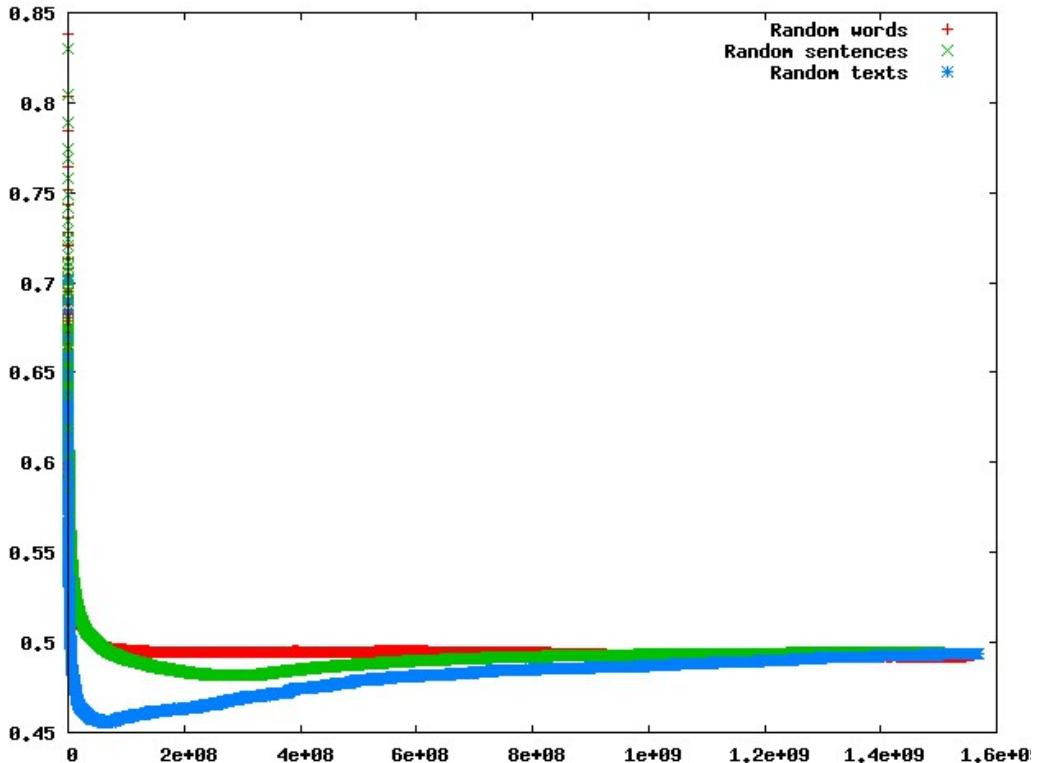


Fig. 10: Hapax-type ratio of Czech word-forms – random order of texts, sentences and words.

What causes the difference between languages here when other described features seem to be language-independent? Morphology richness and variability of forms play a crucial role here. Whereas most of the variability of forms in English is derived from lexicon (different forms are usually different lexemes), in Czech it is the morphology which causes the repertoire of types to be even more diverse. Each word-form of a frequent lexeme in English has thus sooner more than one occurrence, whereas a Czech

lexeme can occur in many unequally usual word-forms therefore it requires more data in order to be turned from hapax word form to more frequent one. Some word-forms of quite frequent Czech lexemes can therefore remain hapaxes in quite a large corpus because the grammatical category they are representing is rare or domain specific. Cores delimited by hapax-type ratio in these languages are thus of a different type; with one being more lexical (English) and the other a combination of lexical and morphological (Czech).

6. Conclusions

Hapax-type ratio can reveal all sorts of interesting facts about language and corpus, with the most important of them being the identification of the size of the language core. It is important to emphasise that the experiment described above can tell us something about the size of the core, not about its exact composition or its repertoire. It is indisputable that the core elements are always the most frequent ones, and this experiment is only an attempt to draw the border line between the core and the periphery elements in a hypothetical frequency word list. Determination of the size of the core is only the first step in contemplating the appropriate size of a given corpus. In order to have all core elements sufficiently represented in a corpus for (lexical or grammatical) description we have to start with multiples of the size of the core. If the core of 80,000 lemmas is delimited by a corpus with 3 million tokens we can base our description of these core elements on corpus of the size at least 30 million tokens.⁶

The shape of the hapax-type ratio function (pipe-graph) also suggests that there are still some unknown differences between text and language. Some of the quantitative laws discovered by exploring individual texts which claim to be valid for an entire

⁶ Even this size of corpus does not guarantee that all the core elements will be sufficiently represented (it largely depends on type of language).

language might therefore be biased by this phenomenon of qualitative change in data structure when the corpus exceeds certain size. Results derived from observing a small amount of data might thus be biased by this phenomenon and may reflect not the language as a whole but only properties of the text (or set of texts) on which the research was based.

We might thus interpret the pipe-graph (with slight exaggeration) as a special case of the parole-langue distinction: the first part of the graph (decreasing function) reflects the properties of texts (parole), followed by the transient part (plateau), and the third part (increasing function) reflects the whole language or domain (with a corpus large enough to eliminate idiosyncrasies of individual texts or speakers). While the first values of hapax-type ratio reflect the variability of specific texts, the last ones are driven by the properties of language, especially by the proportion of core and peripheral elements in the whole language (represented here by a large corpus).

This experiment also shows important differences between languages and units within them. Based on the results, we can try to postulate a new coefficient of inflection for each language which could serve as a starting point for yet another language comparison on a corpus basis.

This comparison should also include the question of derivation on the one hand and the tendency to create a multi-word unit on the other. It would be therefore interesting to confront findings about the low formal variability of English with the hapax-type ratio of bigrams or trigrams. As the repertoire of bigrams or trigrams is several times larger than the number of word types, this kind of research would need far more resources (recent experiments show that we will need at least tens of billions of words).

Finally, change of hapax-type ratio has taught us an important lesson – we have

to limit our conclusions to those parts of language that we actually can observe. Perhaps, with growing corpora, future researchers may discover similar phenomena to the local minimal point which could reveal something more about the distinction between individual texts, collections of texts and language.

Corpora used

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at: <http://www.natcorp.ox.ac.uk/>

Czech National Corpus – Corpus of internet version of L'Arena (1992-2000). Institute of the Czech National Corpus FF UK, Praha 2010.

Czech National Corpus – SYN. Institute of the Czech National Corpus FF UK, Praha 2010. Available at: <<http://www.korpus.cz>>.

Czech National Corpus – SYN2010. Institute of the Czech National Corpus FF UK, Praha 2010. Available at: <<http://www.korpus.cz>>.

References

Crystal, D. (1997). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press,

Francis, W.N. and Kučera, H. (1982). *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company.

Laufer, B. (2010). 'Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension'. *Reading in a Foreign Language*, Vol. 22,

No. 1, 15-30.

Ogden, C. K. (1930). *Basic English. A General Introduction with Rules and Grammar.*

London: Kegan Paul.

Popescu, I.I. and Altmann, G. (2006): 'Some aspects of word frequencies'.

Glottometrics, 13, 23-46

Waring, R. and Nation, P. (1997) 'Vocabulary size, text coverage, and word lists'. In *Vocabulary: Description, Acquisition and Pedagogy*, N. Schmitt and M. McCarthy (eds.). Cambridge: Cambridge University Press, 6-19.

Zhang, H., Huang, C. and Yu, S. (2004): 'Distributional consistency: A general method for defining a core lexicon'. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC2004), 1119-1222.