Mapping the Time Continuum:
A Major Raison D'être for Diachronic Corpora

Karel Kučera[1]

1. Introduction

Historical corpora (i.e. corpora including other than contemporary texts) seem to be used as substitutions of manual excerption rather than tools for discovering really new facts, or types of facts, about history of languages. If this is true, one could provocatively but justifiably ask whether all historical corpora are good for is making linguistic work less time-consuming and more comfortable. It is not easy to argue convincingly against this view if the historical corpus in question is a collection of texts which only covers some parts of the history of a language and is built without any discernible conception. On the other hand, one can successfully argue for the corpus, if it is a diachronic one (i.e. one covering the entire history of a language, with the understandable exception of the contemporary stage, which is usually reflected in much larger synchronic corpora), built with a defensible conception of representativeness in mind.

However, before one really starts arguing for the corpus, one should be aware of the fact that the representativeness of diachronic corpora is a somewhat unsatisfactory concept rather different from the representativeness of synchronic corpora. At a general level (see Kučera, 1999b and 2002) one can say that representativeness of a synchronic corpus is derived (a) from the linguistic experience and intuition of the native speakers of the language (hence the synchronic corpus is considered representative if no more or less common word, phrase, sentence structure etc. is missing from it), (b) from the totality of the contemporary communication in the language (the corpus is representative if all more or less common contemporary types of texts and domains of communication are proportionally represented in it), and (c) from the degree of authenticity of the texts in the corpus (the corpus is representative if it represents the real language faithfully, i.e. without "corrections" or any other than purely formal changes like, for example, unification of letter fonts or styles). The concept of representativeness of a diachronic corpus has not been discussed in great detail so far, but it seems that in the end it can only be based on the body of preserved texts and the authenticity of those included in the corpus. However, the linking up of representativeness of diachronic corpora to the body of preserved texts means that the corpora reflect, in fact, the skewed stylistic, genre and other proportions in the body of texts rather than the characteristics of the real language of the time. This holds especially for the early periods of history of languages, where the number of texts is usually very limited and very often of the kind which was undoubtedly far removed from common communication (particularly texts written in verse).

Considering these inevitable limitations, one may well ask if such thing as the mapping of the time continuum of a language is at all possible. The obvious objection to this undertaking is that whatever data we can get from the corpus will reflect both

[1] Czech National Corpus Institute, Charles University, Prague, Czech Republic
e-mail: karel.kucera@ff.cuni.cz

changes in the language itself and changes in the proportions of various text types and domains at different periods of time, and it may be impossible to distinguish one from the other. Still, even at the present, rather elementary stage of development of diachronic corpora it seems to be possible to get some encouraging results showing facts about history of linguistic units and their combinations which have been unknown so far.

The Diachronic Part of the Czech National Corpus (DCNC), under construction, intended to cover the seven-century history of Czech written texts, has been used in this contribution to show some of the potential of diachronic corpora to map the historical continuum of languages. With its current modest size of over 2 million running words, about one-third of it accessible on the internet, the DCNC can hardly be called representative or sufficient for detailed analyses of the historical continuum of the Czech language, but even so it can be used to convincingly demonstrate the case with chosen examples.

2. Experimental

To minimize the abovementioned limitations as well as the problems associated with the very limited representativeness of the DCNC, the data for the examples below have been extracted in the following way:

(a) The texts included in DCNC were grouped in one-hundred-year clusters starting with the first year of the century and ending with its last year (thus, for example, all corpus texts written or printed from the year 1501 through 1600 were taken as one cluster labeled 1600); the frequency of selected forms or combinations of forms was then extracted from these clusters. In more frequent linguistic units (letters and sounds), fifty-year clusters of texts were used as sources of frequency data.

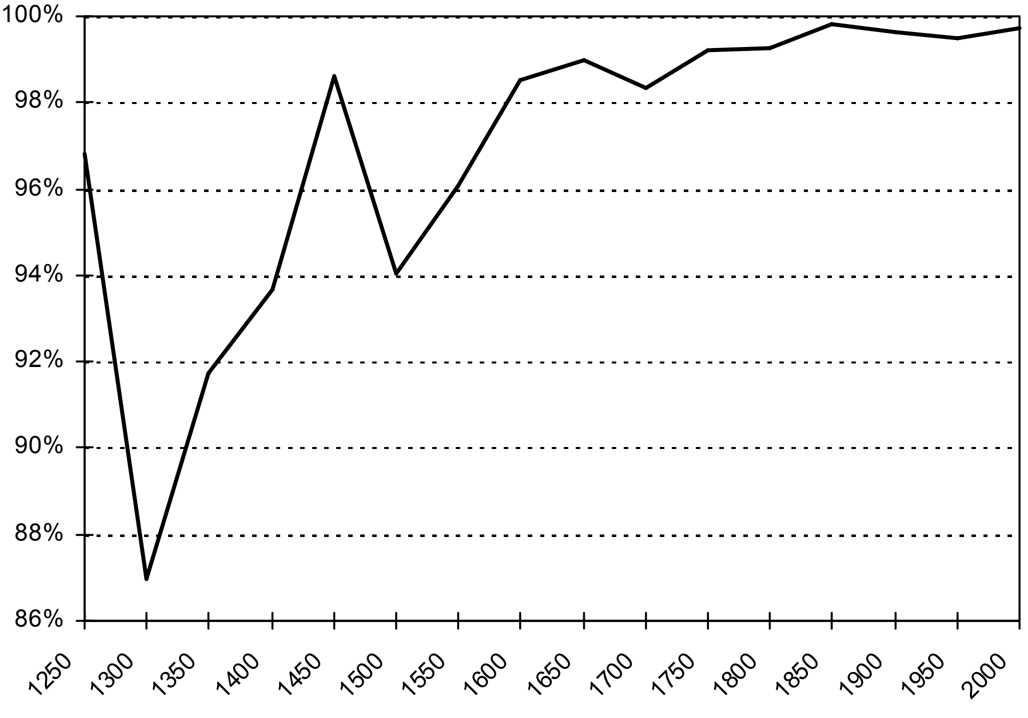
(b) The examples examined below were chosen to be as much independent of topics, literary styles and forms as possible. There was no problem with examples concerning the history of the Czech writing system or phonology, as letters and sounds are highly frequent units without any consistent association with different types of texts. In morphology, syntax and vocabulary only such high-frequency words, forms and structures were used as examples that can be said to be largely independent of topics, literary styles and forms; moreover, the choice of morphological, syntactic and lexical examples focused on groups (mostly pairs) of competing words, forms and structures, and the results have been computed as mutual ratios of the frequencies of the competing units to avoid the fluctuation of their absolute frequencies.

The following graphs represent the results.

3. Results and discussion

3.1 Writing system

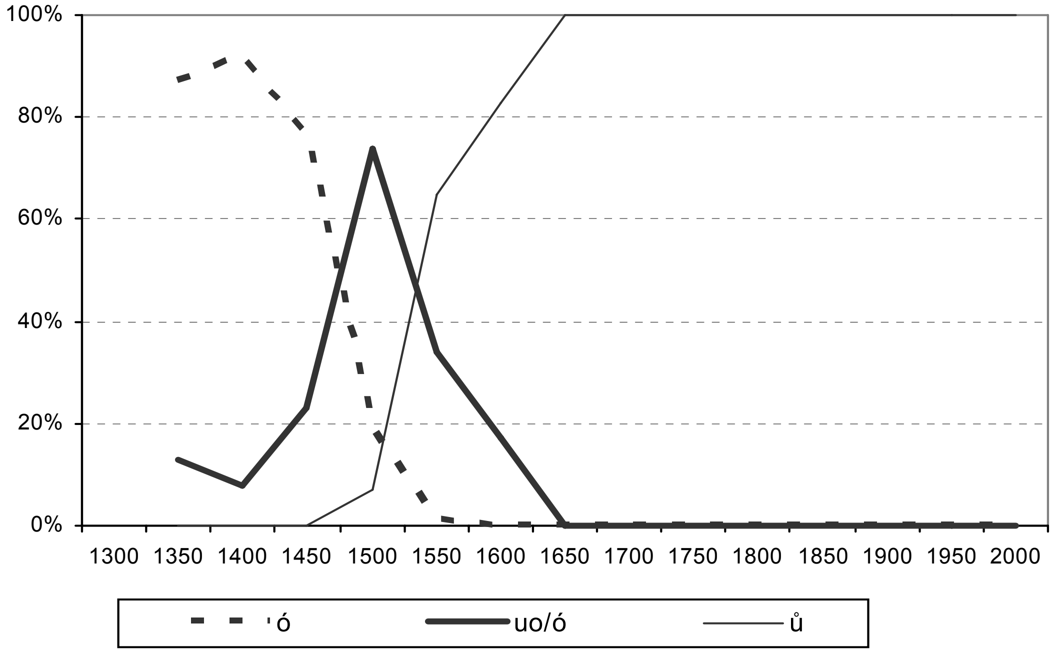
Until now, overall characteristics of the history of the Czech writing system focused primarily on different modifications of medieval Latin alphabet used for the writing down of Czech texts. Three different orthographies (viz. (a) primitive orthography, characterized by an ad hoc use of Latin letters to write down specific Czech sounds, (b) combinatorial orthography, marked by the use of digraphs, trigraphs or even more complicated combination of letters, and (c) diacritical orthography with its typical diacritical marks above Latin letters) are generally recognized as major stages of development of the Czech writing system, and their changes, varieties as well as historical, cultural and systemic contexts have been described in great detail. The DCNC made it possible to follow several alternative lines of development of the writing system more or less independent of the three orthographies (for a more detailed account see Kučera, 1998 and 1999a). Graph 1 shows one of the major lines, namely the development of efficiency of the changing Czech writing system, which has been based on 20,000-letter samples of texts taken at fifty-year intervals, and computed as the ratio of the number of sounds to the number of letters needed to write down the sounds,. The graph shows (a) how the relatively high efficiency of the primitive writing system was abandoned in favour of the less efficient, but much more unambiguous combinatorial system around 1300, which, in turn, was being simplified and made more efficient during the 14th century, until an attempt was made after the year 1400 to replace it with the highly efficient diacritical system used in Czech to the present day. What was virtually unknown, or at least unanalysed and unformulated before this analysis, was the rather surprising drop in efficiency between 1450 and 1600. The most likely explanation for the drop is the introduction of letter-print, which – given the fact that for decades the early printers had no letters with diacritical marks at their disposal – led to reintroduction of digraphs. Moreover, as the print quickly became the most prestigious form of texts, contemporary scribes started to imitate it, so that the digraphs reappeared not only in prints but also in manuscripts. However, with the types including diacritical marks becoming more and more available in the 16th century, the efficiency of both printed and handwritten texts grew constantly, until 1600. The moderate growth of efficiency between 1600 and 1850, when it reached today's level, was caused by the gradual abandonment of several surviving digraphs.



Graph 1: History of efficiency of the changing Czech writing system.

3.2 Phonology

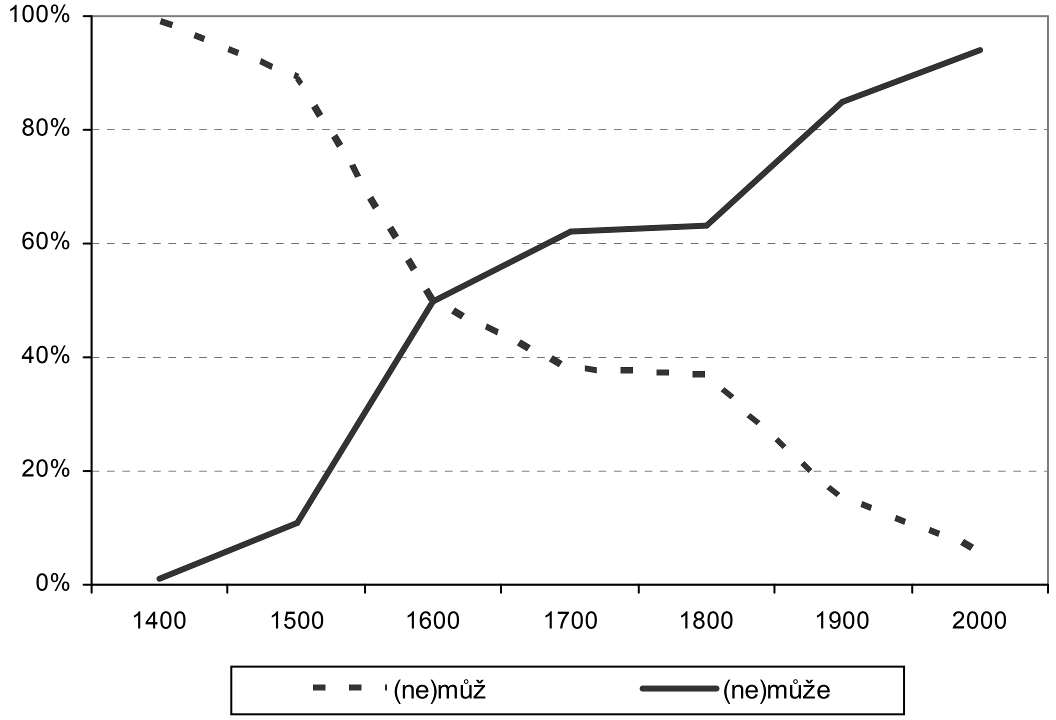
The DCNC has also been used to demonstrate some undiscussed aspects of sound changes in Old Czech, with the primary focus on differences in the dynamics of the changes in different positions in the word (see Kučera, 2006). Graph 2 shows a relatively simple case of two successive changes, namely *ó>uo* and *uo>ú/ů*, which were realised in Czech from the 14th through the 16th century. The general dynamics of the changes had been described in great detail long before the present analysis of corpus data, but what remained virtually unanalysed and unsaid, was that at a certain period (from about 1450 to about 1550, according to the graph), the two changes overlapped, so that one could find all the three alternants (*ó, uo* and *ú*/ů) in different texts of the same time.



Graph 2: Phonology: History of the changes *ó>uo* and *uo>ú*.

3.3 Morphology

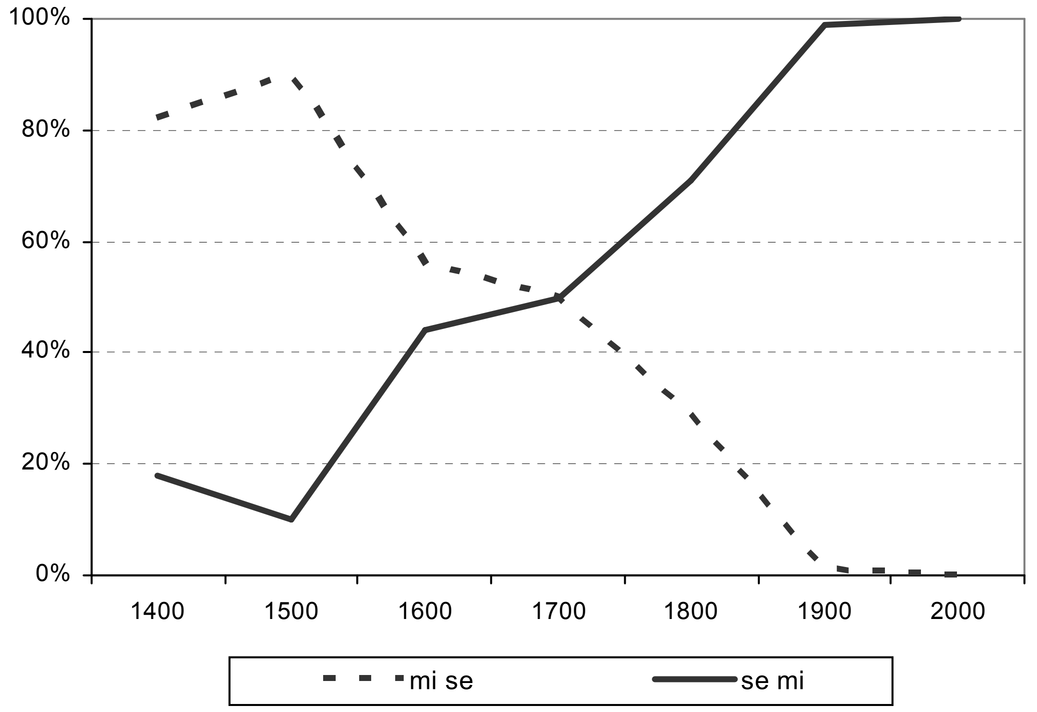
Morphological information that can be obtained from a diachronic corpus like the DCNC at the present time and stage of development is largely limited to the concurrent use of frequent endings and forms. Graph 3 represents the more than five hundred years long competition of two 3rd person singular present-tense forms *(ne)můž* and *(ne)může* of the verb *moci* (‘can’); the graph is focused on the ending, ignoring the different sound varieties of the forms (*(ne)móž, (ne)muož, (ne)můž*, *(ne)móže, (ne)muože, (ne)může*). Again, the history of the competition of the two forms has been virtually unknown, the impression being little more than the general notion that the form without the word final *–e* (*(ne)můž*) existed in the past and was rather frequent at some times. The graph, however, shows the competition of the two forms as a steady process going on through the whole history of Czech written texts. Here, as well as in all the following graphs, the lines representing the progress of the change can be expected to get smoother as the corpus grows and the number of occurrences of the forms in it increases.



Graph 3: Morphology: History of competition of the 3rd person sg. forms *(ne)můž* and *(ne)může*.

3.4 Syntax

An example of a systemic change in the Czech word order has been chosen to represent syntactic information that can be obtained from a diachronic corpus. The fixed position of enclitics in the sentence has been one of the exceptions to the otherwise highly free Czech word order, both in the present and the past. In Graph 4, the focus is on the systemic change in the combination of the enclitic dative forms of personal pronouns *mi, ti, mu* and the enclitic reflexive pronoun *se*. The general information found in historical grammars is that in Old Czech the standard word order was "the dative forms followed by *se*" (that is, for example, *mi se*), while in New Czech the word order is "*se* followed by the dative forms" (e.g. *se mi*). The graph below shows the change as a rather slow process extending over more than five centuries, and adds a piece of new information to the above general statement, namely that even in the oldest Czech texts "the dative forms followed by *se*" was a strong tendency rather than a rule.

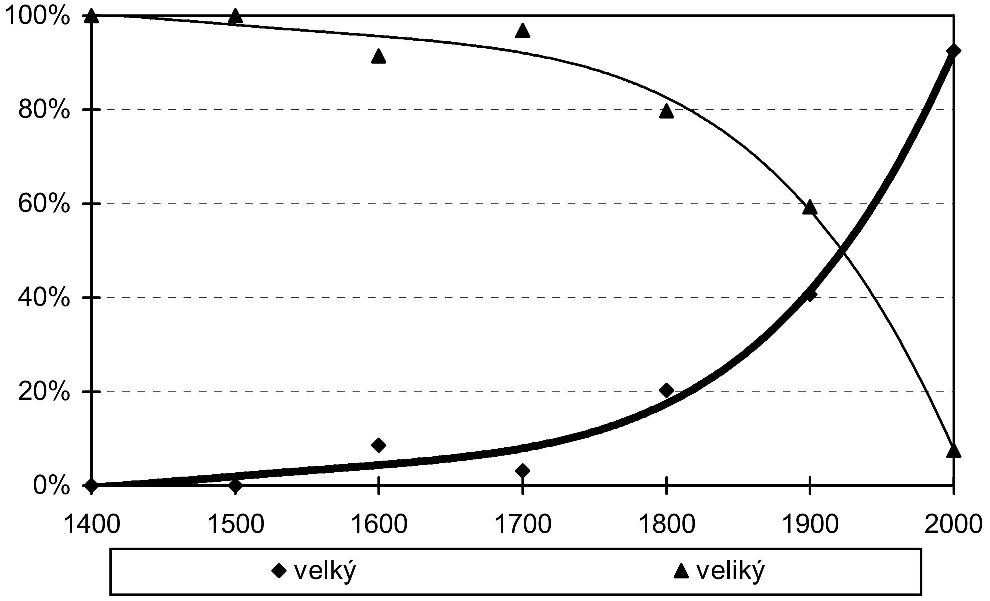


Graph 4: Syntax: History of competition of the word order in combinations of the dative forms of personal pronouns (*mi, ti, mu*) and the reflexive pronoun *se* (*se mi* and *mi se*).

3.5 Vocabulary

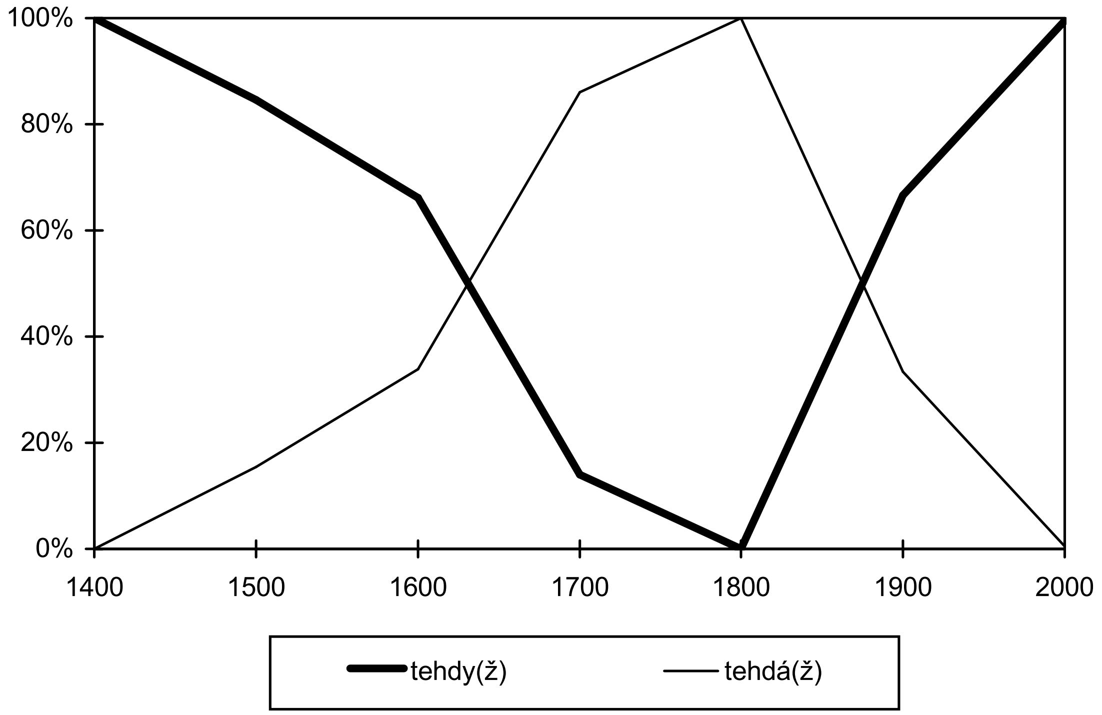
Vocabulary is arguably the area where the contribution of diachronic corpora to the mapping of time continuum of a language is most obvious. Also, it is the area where the corpus reveals a fair amount of new facts, since in most languages very little detail is known about the histories of individual words. Three cases of competing Czech synonyms have been chosen to demonstrate the potential.

The first case (competition of the words *veliký* and *velký*, both meaning ‘big, large’), presented in Graph 5, is a history of one word gradually replacing another in expressing the same meaning. The remarkably smooth trend line that goes through the values of relative frequencies of the two words and extends over the entire seven centuries of Czech texts represents a completely new perspective on the history of the two words.



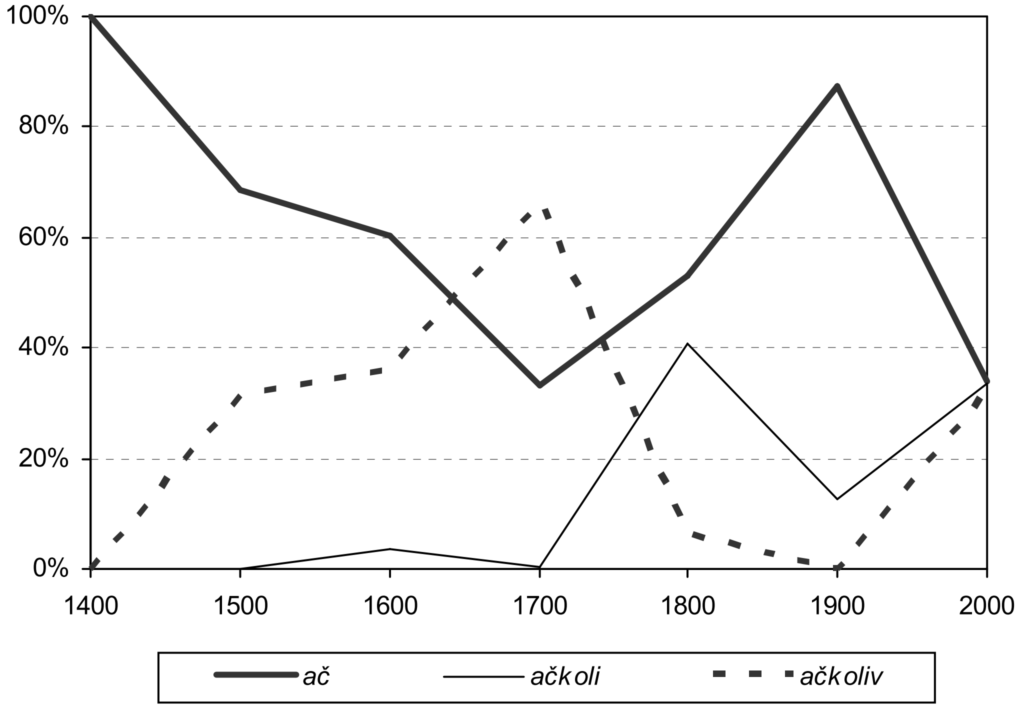
Graph 5: Vocabulary: History of competititon the words *veliký* and *velký* 'big, large'.

Another case (competition of the words *tehdy(ž)* and *tehdá(ž)*, both meaning 'then, at that time'), presented in Graph 6, shows a different history of two competing synonyms: one of them (*tehdá(ž)*), virtually nonexistent in the oldest Czech texts, started gradually replacing the other, but around the end of the 18th century, when it almost completely replaced *tehdy(ž)*, the frequency of the latter started to grow rapidly; in today's Czech *tehdá(ž)* is obsolete, its frequency being reduced to almost zero. This rather surprising, as yet unknown, turn was probably brought about by the national revivalists who searched for unused and unusual words to enrich the vocabulary of the contemporary literature.



Graph 6: Vocabulary: History of competition the varieties *tehdy(ž)* and *tehdá(ž)* ‘then, at that time’.

The third case, presented in Graph 7, represents still another course of development of competition of synonyms. As can be seen from the graph, each of the words *ač, ačkoli* and *ačkoliv*, sharing the meaning ‘(al)though’, had its special history, with its special ups and downs, which remain to be satisfactorily explained, but in contemporary Czech texts all of them have practically the same frequency.



Graph 7: Vocabulary: History of competition the forms *ač, ačkoli* and *ačkoliv* '(al)though'.

4. Conclusions

In our opinion, the examples given in graphs 1–7 demonstrated convincingly that diachronic corpora – in spite of their limitations as well as their present rather elementary stage of development – can be used to map the time continuum of languages and provide new facts about their histories. In the foreseeable future, a more general fcontribution to historical linguistics brought by the use of diachronic corpora could be seen in more emphasis on development, historical perspective, historical continuum. One can also hope that quantitative and statistical analysis of the information obtained from the corpora could lead to identification of new, as yet unknown turning points in the histories of individual languages, that is identification of periods where the development of a large number of linguistic units or their combinations changed markedly.

References

Kučera, K. (1998) Vývoj účinnosti a složitosti českého pravopisu od konce 13. do konce 20. století. *Slovo a slovesnost 59*, 178–199.

Kučera, K. (1999a) Dodatek ke kvantitativním charakteristikám vývoje českého pravopisu od 13. do 20. století. *Slovo a slovesnost 60*, 301–303.

Kučera, K. (1999b) The General Principles of the Diachronic Part of the Czech National Corpus, in *Text, Speech and Dialogue.* Matoušek, V., P. Mautner, J. Ocelíková and P. Sojka (ed.), pp. 62–65. Berlin, New York etc.: Springer.

Kučera, K. (2002) The Czech National Corpus: Principles, Design, and Results. *Literary and Linguistic Computing, 17, 2*, 245–57

Kučera, K. (2006) Kvantitativní charakteristika průběhu a uplatnění změn *ý>ej, ú>ou* a *aj>ej* v češtině od 14. století do současnosti, in Čermák, F., R. Blatná (ed.): *Korpusová lingvistika: Stav a modelové přístupy*, pp. 210–25. Praha: NLN.