

The Czech National Corpus: Principles, Design, and Results

Karel Kučera

Charles University, Praha, Czech Republic

Abstract

This paper describes the general principles, design, and present state of the Czech National Corpus (CNC) project. The corpus has been designed to provide a firm basis for the study of both the contemporary written Czech (a goal well attainable with the present resources) and the Czech language beyond the limits of contemporary written texts (a long-term commitment including the building of a corpus of spoken Czech and diachronic and dialectal corpora). The work on the CNC project, now in the eighth year of its official existence, has resulted in the completion of SYN2000, a 100-million-word corpus of contemporary written Czech, the organization of the cores of spoken, diachronic, and dialectal corpora, and the finding of workable solutions to some general theoretical problems involved in the building of these corpora.

1 Introduction

The idea of constructing a representative corpus of the Czech language was first conceived at the beginning of the 1990s by a group of linguists from Czech universities and the Academy of Sciences, but was largely rejected by officials for several years. The work on the Czech National Corpus (CNC) started in 1994 when the Faculty of Arts at Charles University, Prague, founded the Czech National Corpus Institute and provided it with basic equipment (for more details about the institute see <http://ucnk.ff.cuni.cz>). In the following years, the construction of CNC gained momentum after the institute was awarded grants by the Ministry of Education and the Grant Agency of the Czech Republic; it was the financial support from these two sources that made it possible for the team of CNC builders to grow to its present size of ten researchers. The team co-operates with about twenty linguists and programmers from other Czech institutions and a number of corpus linguists in other countries.

From the outset CNC has been conceived as a continuous project, the concrete results of which (specific corpora and databases) will monitor the forms, varieties, and development of the Czech language and provide the users of CNC both with as representative and ample data as possible

Correspondence:

Karel Kučera,
Czech National Corpus Institute,
Filozofická fakulta UK, náměstí Jana
Palacha 2, 11638 Praha 1,
Czech Republic.
E-mail:
karel.kucera@ff.cuni.cz

and with tools for their exploitation. The project has been worked up in conformity with the SGML-based Text Encoding Initiative (TEI) guidelines (see Sperberg-McQueen and Burnard, 1994), the objective being to create, extend, and update a versatile source of information capable of meeting the variegated needs of linguists and non-linguists, teachers and students, scientists and the general public (see Čermák (1997, 1998, 2001); for a general background, see Atkins *et al.* (1992) and McNaught (1993)).

2 Principles: Aspects of Representativeness

A corpus built with these objectives in view has to meet certain criteria, which are usually subsumed under the general principle of representativeness (for a general discussion, see Biber (1993)). Although this is hardly a disputable statement, it is also extremely general, as the principle of representativeness, crucial as it is, has been used and referred to rather loosely and vaguely in both corpus and non-corpus linguistics, and the differences between existing corpora suggest that there are differing views on how the general concept translates into the size and structure of a large versatile corpus. (In smaller specialized corpora, such as those focused on the texts of one author, the representativeness of the corpus may be a relatively straightforward principle: often it simply means the inclusion of all the relevant texts in their authentic form.)

2.1 The synchronic section of CNC: written texts

The conception of representativeness applied in the synchronic section of CNC is based on the fact that qualified users of a large versatile synchronic corpus make judgements about its representativeness founded on

- (1) their linguistic experience and intuition (from this point of view the corpus is considered representative if no more or less common word, phrase, sentence structure, etc. is missing from it);
- (2) the totality of the contemporary communication in the language (from this perspective the corpus is representative if all more or less common contemporary types of texts and domains of communication are proportionally represented in it; a corpus having this quality is sometimes also described as 'well-balanced');
- (3) the degree of authenticity of the texts in the corpus (from this standpoint the corpus is representative if it represents the real language faithfully, i.e. if the texts in the corpus have not been 'corrected' or changed in any other than a purely formal way, which now usually includes—but may not necessarily include in the future—changes such as unification of fonts and styles, exclusion of pictures, or the cancelling of division of words at the ends of lines).

Thus, it seems to be fair to conclude that the representativeness of a large versatile corpus has three major aspects, namely, size, proportionality, and authenticity, and in the view of the builders of CNC it cannot be reduced to any one or two of them. Authenticity is a *sine qua non* of representativeness in any corpus and cannot possibly be ignored. Neither

can the criterion of size be ignored in a large versatile corpus, as only a corpus of certain minimum size can include all common—as well as a variety of extremely uncommon—language elements (however, what the minimum size is remains an open question: the current stage of development of the British National Corpus indicates that it may be well over 500 million running words; which means that until we know more, a large versatile corpus should be simply as large as possible). Finally, proportionality cannot be ignored in a large versatile corpus because even very large corpora built of authentic texts without a view to proportionality cannot be said to be really representative of a particular language as a whole: they do not reflect the entire spectrum of text types and domains or their proportions in contemporary communication and as such they can hardly serve as really dependable bases for such areas of exploitation of corpora as the compiling of large dictionaries.

For the builders of a large versatile synchronic corpus, authenticity and size present practical rather than theoretical problems. The concept of proportionality, on the other hand, deserves closer examination and specification in every new case, as the proportions of various text types and domains may greatly differ in the communication in different languages, nations, and cultures. The proportionality of CNC has been based on sociological research into language reception, particularly into the size and frequency of exposure of Czech speakers to various topics and kinds of written language. The results (see Čermák *et al.* (1997), Králík (2001), and Šulc (2001) for a more detailed account) showed that the ratio of specialized or technical texts to nonspecialized texts was almost exactly 1:2 (33.5%:66.5%), a figure much more favourable to specialized texts than had been expected. In nonspecialized texts, broken down into three subgroups, the research showed an unexpectedly overwhelming predominance of journals (56%) over fiction and poetry (10%), and other types of nonspecialized texts such as letters or chronicles (0.5%).

These results have been adjusted and extended by further research and exploitation of available data such as the records of loans of different kinds of books in public libraries. Although the proportions of different text types and domains in contemporary Czech communication are still being studied from different angles, in 2000 the specification of the concept of proportionality based on language reception reached the level of practical applicability and the concept was applied in the first representative version of the Czech synchronic corpus of written texts called SYN2000 (see Section 4.1 for details). Its structure, reflecting some changes in the approach to the division of text types and domains, is as follows:

I. Imaginative texts (15%):

1. Fiction (11.02%)
2. Poetry (0.81%)
3. Drama (0.21%)
4. Other literary texts (0.36%)
5. Transitional types of texts (2.6%)

II. Informative texts (85%):

1. Journals (60%)
2. Technical and specialized texts (25%):
 - a. Lifestyle (5.55%)
 - b. Technology (4.61%)
 - c. Social sciences (3.67%)
 - d. Arts (3.48%)
 - e. Natural sciences (3.37%)
 - f. Economics and management (2.27%)
 - g. Law and security (0.82%)
 - h. Belief and religion (0.74%)
 - i. Administrative texts (0.49%).

2.2 The synchronic section of CNC: spoken texts

In the synchronic corpus of spoken language the above conception of representativeness has been modified in view of the specific problems and practical limitations connected with the task. In fact, one of the three aspects of representativeness—size—can only be considered theoretically in the construction of spoken language corpora, because the present demands of the work on time, manpower, and financial resources are so large that to build a spoken corpus of a size at least comparable with that of large written corpora is simply unfeasible. The above-mentioned sociological research into language reception among Czech speakers showed that, on average, they are exposed to speech about twice as much as to reading, which means that the ratio of the spoken component to the written component in the corpus should be 2 : 1. This, in turn, means that—like other existing spoken corpora—the Prague Spoken Corpus with its present size of 800,000 running words definitely cannot be called fully representative.

Faced with this fact, special attention has been given to the other two aspects of representativeness, i.e. proportionality and authenticity. In addition to the tenet that the texts included in the corpus should not be ‘corrected’ or changed in any other than a purely formal way, the authenticity of the Prague Spoken Corpus has one more aspect, namely, the exclusive use of prototypical spoken language. This means that no texts combining the characteristics of the spoken and written language, such as lectures, theatrical plays, or political speeches, have been included in the corpus (see Section 3.2 for more details).

The conception of proportionality in the spoken corpus has been based on balancing four sociolinguistic criteria in the recorded sample of speakers:

- (1) sex (women and men are equally represented in the recordings);
- (2) age (only speakers over 20 years of age have been recorded, with an equal proportion of two age groups, namely, 21–35 years and over 35 years);
- (3) education (speakers with elementary education and those with higher (i.e. secondary or university) education have been equally represented in the corpus);

- (4) type of the discourse (equal proportion of ‘formal’ discourse, controlled by a series of general questions, and ‘informal’ discourse, represented by loose conversations).

In the Prague Spoken Corpus, all existing combinations of the above criteria are represented by a practically equal number of recordings.

2.3 The diachronic section of CNC

Still another conception of representativeness has been applied in the diachronic section of CNC (for a more detailed discussion, see Kučera (1999)). Understandably, the representativeness of the diachronic corpus can be based neither on speakers’ linguistic experience or intuition nor on the totality of communication within a particular period. Practically everything we know for sure about the past of the Czech (or any other) language is based on a limited number of preserved texts, and so it is the body of these texts and the authenticity of those included in the corpus that are the ultimate measures of representativeness in this section of CNC.

In this perspective, the diachronic corpus has been built on two somewhat different conceptions reflecting the estimates of the amount of historical Czech texts that can be (mostly manually) converted into a machine-readable form in the foreseeable future. According to the estimates based on the current pace of progress, in years (rather than decades) the diachronic section of CNC should reach full representativeness for the oldest period of Czech written records (from about 1250 to 1400), i.e. it should include all the preserved texts that were written in this period. In decades rather than years, the diachronic section should reach full coverage also for the fifteenth century.

From the sixteenth century onwards the number of preserved Czech texts is so large that it makes full representativeness of the corpus unfeasible. The sixteenth century is thus the first century that is represented in the diachronic corpus in a similar way as the present is represented in the synchronic written corpus. Practically this means that the period from the sixteenth to the twentieth centuries is being covered by stratified random sampling of the body of preserved texts, in which the stratification should ensure that all types of the preserved texts are included in the corpus. It is true, however, that in spite of the plan to steadily extend the number of texts in the diachronic corpus the satisfactory representation of certain text types or domains of communication will remain unattainable, especially in the older part of this period (sixteenth to eighteenth century), and the diachronic corpus will inevitably reflect the skewed stylistic, genre, and other proportions in the body of preserved texts. However, the closer to the point where it borders on the synchronic section of CNC (see Section 3.1) the more the diachronic corpus will conform with the proportions of text types and domains applied in the synchronic corpus. Given that sooner or later older texts will be moved from the synchronic to the diachronic section of CNC, the transition between the two corpora should be as smooth as possible.

The principle of authenticity, too, has a rather different meaning when used in connection with the diachronic corpus. For one thing, a distinction has been drawn between 'authentic' texts (texts written or printed at a certain time and preserved from that time to the present day) and 'unauthentic' texts written or printed at one time and preserved only in newer copies or reprints from a different period. The unauthentic texts are not included in the corpus, because their language is often a mixture of the original text and innovations, and does not belong fully to either of the two periods. The inclusion of newer copies and reprints cannot be completely avoided, however, especially in the early period, but the difference between the time at which the original was written or printed and the time at which it was copied or reprinted must not exceed certain pragmatic limits: so far, the maximum difference of 20 years has been tolerated up to the end of the 15th century and that of 10 years has been tolerated up to the end of the 17th century; only fully authentic texts have been included in the corpus from the period after the year 1700. However, this discrimination between authentic and unauthentic texts does not really mean that unauthentic texts are ignored by CNC. They are not included in the diachronic corpus proper, but those available in an electronic form are stored in the bank of historical texts for special research purposes.

Another specific aspect of authenticity in the diachronic corpus is a consequence of the fact that the Czech writing system has undergone radical changes and a great degree of fluctuation during its 700-year history (for example, the word spelled today as *tvář* 'face' could be spelled as *twarz*, *twarž*, *twařz*, *twárz*, *twárž*, *twář*, *twaarz*, ..., often with several different spellings used in one text). There seem to be only two ways of searching for such a variety of spellings without the transcription of the texts: one can either keep a database of all the various spellings of all the forms of all the lemmata in the corpus (which is, in fact, an extension of lemmatization), or create a search program that would generate, and consequently search for, all the possible spellings of the given word or form. However, several tests carried out in CNC showed that in Czech (and probably in most languages with morphologies as rich as Czech) both of these approaches are prohibitively ineffective, although the first has been successfully applied in a historical corpus of a different language (see Beltrami, 1998). As a consequence, the solution adopted in CNC was to transcribe the texts written or printed before the last major changes in the Czech writing system in 1849, which is standard practice in publishing Czech historical texts, even for research purposes. To counterbalance the loss of authenticity resulting from the practice, the CNC plans include the commitment to link the transcribed texts with their digitized copies, thus making it possible for the user of the corpus to check the authenticity of the transcriptions. Digitizing all the originals and linking them with the texts in the corpus is a costly and time-consuming task that could hardly be performed by a small institution such as CNC, especially if its attention is focused primarily on the contemporary language. However, there are hopes that it will be possible to realize the plan by using to this purpose the

future results of a broad long-term project focused on digitization of the Czech cultural heritage currently supported by several grants.

2.4 The dialectal section of CNC

Because of practical limitations, representativeness in the dialectal corpus has remained a theoretical concept. Ideally, this section of the corpus should include more than 2,000 transcripts of recordings covering the whole territory of the Czech Republic and balancing at the same time basic dialectological and sociolinguistic criteria such as indigenosity, age, education, and sex of the speakers. For comparative reasons the samples of dialects should be recorded, if possible, in the same towns and villages where the dialectal samples were taped during the full-scale, as yet unsurpassed, research of the Czech dialects organized in the 1950s.

However, at present CNC does not have the resources to launch even a pilot project of this kind, and the dialectal section of CNC is currently only a small random collection of short dialectal texts (see Sections 4.1 and 4.2).

3 Design

The design of the CNC project reflects the breaking down of the general objectives and principles described in Sections 1 and 2 into several more specific goals along three dividing lines: synchrony versus diachrony, writing versus speech, and dialects versus common language. Practical distinctions and solutions adopted in connection with the implementation of the design are discussed below and may be debatable to some extent. However, the bottom line is that CNC has been built as a flexible project and in the foreseeable future its data should be accessible through a software manager that will make it possible for the user to create virtual corpora. Thus, using the current markup of the texts, the users will be able to apply their own conceptions of the contemporary language, the boundaries between dialect and common language, and the types of texts properly representing the spoken language.

3.1 Synchrony versus diachrony

The concept of synchrony (or, rather, the concept of what is and is not the contemporary language), central for the division of CNC, is in part both arbitrary (as there are no universal clear-cut criteria for drawing a distinct line between two states, or stages of development, of a language) and changeable (the boundaries of the contemporary language will have to be repeatedly redefined in the future). The current solution adopted in CNC has been based both on linguistic and extralinguistic factors reflected in the attitudes of native Czech speakers towards what forms of language can be considered living or, on the other hand, dated, and what texts can be accepted as 'historically neutral' and what texts are 'historically marked'. As a result, the specific time boundaries applied in CNC are different in different types of Czech texts. The year 1990 has been accepted as a natural boundary in informative (journalistic, technical, and other specialized)

texts, as the political turnaround at the end of 1989 led to a substantial change both in their language and topics, with the communist jargon and pervasive indoctrination sinking fast into oblivion.

The year 1990 has been also recognized as one of the important points in time distinguishing contemporary imaginative texts (fiction, poetry, and drama) from historically marked ones, but the border is obviously more fuzzy here than in informative texts: many literary works created decades ago are still an essential part of the present Czech culture and communication, being repeatedly republished, read, and quoted. This is why two additional empirically based boundaries (1880 and 1945) have been adopted in CNC, leading to a system in which the core of contemporary imaginative texts consists of texts published for the first time in or after 1990, complemented by texts of authors born after 1880 published or republished between 1945 and 1990. So far the system has worked well for Czech: there seem to be no authors born before 1880 whose language is perceived as contemporary, and no texts published before 1945 (and never after 1945) that would play any noticeable part in contemporary Czech communication. On the other hand, the two additional boundaries make it possible for CNC to include in the pool of present texts such frequently republished and read modern classics as Karel Čapek or Jaroslav Hašek, whose language is still considered contemporary by a large part of Czech native speakers.

The year 1900 has been also accepted as the most important boundary between contemporary and historically marked spoken texts (both common language and dialects). However, allowances have been made for texts recorded in the 1980s, as 1990 is not really as important and obvious a milestone in spoken Czech as it is in written informative texts.

Spoken texts recorded before 1980, informative texts published before 1989, imaginative texts of authors born before 1880, and imaginative texts not published after 1944 are all funnelled to the diachronic section of CNC.

3.2 Writing versus speech

In Czech, as in many other languages, the once-distinct difference between written and spoken language has been blurred in the course of time by the growing number of texts in which both speech and writing have been applied in varying proportions. Some types of mixed texts, e.g. political speeches, public addresses, transcripts of parliamentary debates or radio lectures, have often been used as specimens of spoken language in linguistics, and it is not hard to understand why even in some contemporary large corpora, striving to provide the linguist with an authentic material basis, spoken language is represented mostly by such blends. The obvious reason is the above-mentioned fact that to record and transcribe prototypical, purely spoken texts today is a prohibitively expensive and time-consuming exercise that cannot but fail to gather enough authentic data for a large corpus to reflect realistically the speech/writing ratio in contemporary language communication. Even the British National Corpus, the project with by far the largest (about 10%) proportion of

purely spoken texts, is far from being ‘balanced’ or ‘representative’ in this respect.

Given the fact that at present a large versatile corpus cannot possibly reflect the real proportion of the spoken language in contemporary communication, the builders of CNC hold the opinion that the inadequately small spoken component of the corpus should at least represent the prototypical spoken language as faithfully as possible. As a result, the spoken part of CNC has been composed exclusively of samples of authentic spoken Czech (see Section 2.2). Texts representing various blends of written and spoken language are marked up accordingly and included in a special section of the written corpus.

3.3 Common language versus dialect

The distinction between a dialect (as a form of a language more or less clearly confined to a part of the territory of the language) and common language (as a literary or nonliterary form unconfined in this way) seems to be clear in theory, but may not be always easily applicable in practice. The traditional Czech dialects spoken in tight-knit rural communities have been disintegrating for decades, being gradually replaced by so-called interdialects, i.e. more or less unified forms that are spoken in larger regions and preserve only some features of the former local dialects. Generally, the regional interdialects are less different from the literary standard than the local dialects and they mix more often with literary Czech.

There are two basic criteria used in CNC to distinguish between dialectal and non-dialectal texts in this situation, namely, the frequency of dialectal features in the text and the degree of consistency of their use. Moreover, the degree of dialectal authenticity of the text is always taken into consideration, the authenticity ranging from spoken texts recorded in situations typical of the use of a dialect to formal spoken or written texts, in which the occasional dialectal forms or words are mostly either unwanted slips or elements used for contrast against the general literary background.

3.4 Corpora, banks, and archives

The division of CNC into corpora, banks, and archives reflects, for the most part, the progress of work on the texts. The incoming texts (transcribed texts, texts scanned into computers, or texts provided by publishers or authors) are stored in one of the two archives of CNC (the Synchronic Archive and the Diachronic Archive). The texts remain stored in the archives in their original formats even after subsequent processing and transfer to one of the corpora, because they often contain different kinds of textual information in the form of special codes, fonts, tables, etc. that are not routinely included in the corpus; such information is lost in the processing of the texts for the corpus, but later it may be recovered from the archives, if it is necessary. The archives of CNC are its largest sections; their current combined size is about 500 million running words.

After the processing, which includes conversions into a unified SGML format with a unified DTD (a data type definition including structured information about the text) and cleaning (removal of foreign-language paragraphs, pictures, numerical tables, etc., but no corrections), the texts are funnelled to one of the two banks of CNC (Synchronic or Diachronic). The texts stored in the banks are ready to be transferred to the synchronic or diachronic corpora, but whether they will or will not be really used in a corpus depends on their characteristics. Each of the corpora represents a certain conception of representativeness and is composed of different types of texts in certain proportions (see Section 2); consequently, if there are too many texts of a particular type in a bank, all of them may not be used in the corpus. In most cases such texts are stored in the bank until the size of the corpus grows to the point where more texts of the type are needed, but there are also texts that may never be part of a corpus, because they are not in conformity with the CNC principles (they may be, for example, ‘unauthentic’ texts, i.e. texts copied or printed a long time after the lost original text was written; see Section 2.3). However, even these texts are kept in the bank for possible future use: once the corpus manager enables the user to create virtual corpora from the texts in the bank, researchers will be able to use all the texts, not only those found in the ready-made CNC corpora. In this perspective, the existing CNC corpora may be viewed as realizations of the conceptions of representativeness described above; the banks of texts may be viewed as opportunities to extend one’s research beyond the confines of the conceptions.

4 Results

As has been noted above, the general objectives of the CNC project have been broken down into several more specific goals along three dividing lines, namely synchrony versus diachrony, writing versus speech, and regional dialects versus common language. As a result, CNC has two basic sections (a synchronic section and a diachronic section), each of which includes written, spoken, and dialectal subsections. However, the six corpora that should be the end results of the work in the subsections differ (and no doubt will continue to differ) greatly in size and other characteristics, some of them being hardly more than blueprints for future work, which simply cannot be done now with the present funds and manpower.

4.1 The synchronic section

In CNC, as in most other large versatile corpora, it is the synchronic section that is the main focus of the project, and the corpus of contemporary written texts is its central and fastest-growing part. The first official result of the work in this section is SYN2000 (SYN as in *synchronic*), a 100-million-word corpus of contemporary written Czech completed in 2000, built on the conception of representativeness described in Section 2.1. A 20-million-word sample of the corpus, called PUBLIC, is freely accessible at <http://ucnk.ff.cuni.cz> (PUBLIC can be searched only for words, not

phrases, and there are certain limitations on the size and number of concordance lines). The full use of SYN2000 is free of charge, but is conditional on the user's undertaking in their registration statement not to use the data for commercial purposes. SYN2000 is a lemmatized, tagged corpus with an extensive morphological tag set indicating part-of-speech categories and subcategories, gender, number, case, person, tense, grade, negation, voice, and a characteristic distinguishing basic or 'neutral' forms from less frequent synonymous forms or forms that are archaic, bookish, colloquial, etc. New, larger versions of the corpus of contemporary written Czech are planned for release in the future.

The result of the work in the spoken subsection of CNC is the Prague Spoken Corpus encompassing 800,000 running words. It consists of about 300 transcripts of recordings of authentic spoken language, which represent, in balanced proportions, all existing combinations of the four basic criteria (sex, age, and education of the speakers, and type of discourse) that are at the core of the CNC conception of representativeness described in Section 2.2. The spoken corpus, too, has been tagged; its tagged version will be available in the near future. Its tag set, in addition to the characteristics mentioned in the previous paragraph in connection with SYN2000, includes some more detailed morphologico-semantic tags reflecting the specific, often much more complex way in which some parts of speech, such as particles, for example, are used in spoken Czech. The plans for the Prague Spoken Corpus include its further growth, but the pace of progress, dependent largely on funding, will inevitably be very slow. Two other spoken corpora, representing the spoken language of Brno and Plzeň regions, are being prepared for release.

The corpus of contemporary Czech dialects is the least developed of the contemporary corpora in CNC. As has been noted, CNC does not have the resources to launch even a pilot project covering the whole of the Czech language territory, and, as a result, the dialectal section of CNC has been so far just a minuscule unrepresentative random collection of short dialectal texts. A lexical database has been organized as a supplement of the small text collection, recording dialectal words and phrases currently used in a variety of texts, including written texts, which are not incorporated into the dialectal corpus. Considering the extreme demands of dialectal recording and transcription on funding, time, and manpower, it is difficult to foresee any great progress in this subsection of CNC in the future.

4.2 The diachronic section

Understandably, authentic prototypical spoken language (see Section 3.2) is rare in the diachronic section of CNC: its scarce recordings can be found in just a few decades of the twentieth century, so that the vast majority of the time span covered by the diachronic section of CNC—seven centuries—is represented by prototypical written texts and occasional blends of written and spoken language (court records, records of parliamentary debates, etc.), which, however, are rarer the farther back in time one goes. Consequently, the diachronic spoken 'corpus' is hardly

more than a receptacle prepared for the transcripts that now are part of the synchronic Prague Spoken Corpus, but sooner or later will be transferred to the diachronic section of CNC. These texts will be complemented—no doubt rather unsystematically, just to provide an elementary historical perspective on the newer spoken texts—by the transcripts of whatever recordings of authentic conversations before the 1980s the CNC builders are able to find and transcribe.

Rather paradoxically, the diachronic dialectal corpus, also built of transcripts of authentic spoken language, is now larger than both its synchronic counterpart and the diachronic spoken ‘corpus’ mentioned in the preceding paragraph. The reason is that it includes transcripts of practically all the samples of Czech dialects that have ever been published in dialectological studies, altogether about 200,000 running words. However, this source of historical dialectal texts has been almost fully exploited, so no significant growth of the corpus can be expected.

The core of the diachronic section is the diachronic written corpus, comprising texts from the late 13th century to the 1980s. Although the present corpus includes texts covering the entire time span, most of the work has been focused on manual transcription of texts from the fourteenth to eighteenth century, which represents the most time-consuming task in this section of CNC. There are two main reasons for the preference, namely, the fact that Middle Czech (that is, the language from about 1500 until about 1800) has not yet been properly researched or described, and the fact that the converting into a machine-readable form of Czech texts written or printed before 1500 will smooth the way for further work on the Old Czech Dictionary, a major Czech lexicographical project that has been in progress since 1968. At the end of 2001, the diachronic written corpus had about 2 million running words and has been growing at a rate of about 250,000 running words a year.

References

- Atkins, B. T. S., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7: 1–16.
- Beltrami, P. G. (1998). Norme per la redazione del Tesoro della Lingua Italiana delle Origini. *Bollettino dell'Opera del Vocabolario Italiano*, 1998: 277–330.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8: 243–57.
- Čermák, F. (1997). Czech National Corpus: a case in many contexts. *International Journal of Corpus Linguistics*, 2: 181–97.
- Čermák, F. (1998). Czech National Corpus: its character, goal and background. In Sojka, P., Matoušek, V., Pala, K., and Kopeček, I. (eds), *Text, Speech, Dialogue. Proceedings of the First Workshop on Text, Speech, Dialogue—TSD'98*. Brno: Masaryk University, pp. 9–14.
- Čermák, F. (2001). Language corpora: the Czech case. In Matoušek, V., Mautner, P., Mouček, R., and Taušer, K. (eds), *Text, Speech, Dialogue. 4th International Conference. Proceedings*. Berlin: Springer, pp. 21–30.

- Čermák, F., Králík, J., and Kučera, K. (1997). Recepce současné češtiny a reprezentativnost korpusu. (Reception of contemporary Czech and representativeness of the corpus.) *Slovo a slovesnost*, 58: 117–24.
- Králík, J. (2001). Vyváženost zdrojů Synchronního korpusu češtiny SYN2000. (Balancing the sources of the synchronic corpus of Czech SYN2000.) *Slovo a slovesnost*, 1: 38–53.
- Kučera, K. (1999). The general principles of the diachronic part of the Czech National Corpus. In Matoušek, V., Mautner, P., Ocelíková, J., and Sojka, P. (eds), *Text, Speech and Dialogue. Proceedings of the 2nd International Conference on Text, Speech, Dialogue—TSD'99*. Berlin: Springer, pp. 62–5.
- McNaught, J. (1993). User needs for textual corpora in natural language processing. *Literary and Linguistic Computing*, 8: 227–34.
- Sperberg-McQueen, C. M. and Burnard, L. (1994). *Guidelines for electronic encoding and interchange (TEI P3)*. Text Encoding Initiative. Available at: <http://www.uic.edu/orgs/tei/p3/doc/p3.html>.
- Šulc, M. (2001). Tematická reprezentativnost korpusů. (Text types and domains and representativeness of corpora.) *Slovo a slovesnost*, 1: 53–61.

