The Czech National Corpus

Jan KOČEK, Marie KOPŘIVOVÁ, Věra SCHMIEDTOVÁ,
Praha, Czech Republic

Abstract
The paper deals with the history of the Czech National Corpus (CNC) project. It reports on the present stage of its development, describes what type of corpus it is, and the text processing methods and morphological annotation used in its compilation. It also briefly discusses the software used in the CNC. The Bank of Czech (BoC) has now 330 million word forms. It is the basis of a representative corpus (SYN2000 – 100 million word forms) which was created in spring 2000, and is intended as a material source for future dictionaries. At the moment the lexical saturation of the material is tested.

1 Introduction

The Czech National Corpus is the name for the whole project of text collection and computer text processing. The project includes a corpus of Old Czech, a dialect corpus, a corpus of spoken language in Prague and Brno and a corpus of contemporary written texts. The Bank of Czech (BoC) is a cover term for the sum of all the converted and marked-up texts of contemporary written language which can be examined using our software. SYN2000 is the name of a representative corpus which includes 100 million word forms.

2 History

At the time of political changes in 1989 Czech lexicography and lexicology were in a critical state. The lexical archives on the basis of which principal Czech dictionaries were compiled had not been kept up for several years. The teams of lexicographers generally consisted of people in pre-retirement age and there were no younger workers to ensure continuity. There was, and still is, no-one working on the description of contemporary lexis as a whole, the only lexical stratum being documented is neologisms.

In this difficult situation when the main problem was the absence of language data for lexicographic and linguistic work, representatives of several institutions gathered to establish Computer Collection of Czech (1991) with the aim to co-ordinate all people interested in the preparation of computer-readable and computer-processable material for Czech and to initiate radical changes in lexicographic work.

This team gradually managed to attract funding and succeeded in popularizing their goals in spite of the conservatism, lack of understanding, and sometimes mere envy of many colleagues. Most importantly, however, they carried on their work on the Computer Collection of Czech project in their spare time. It soon became obvious that a new institution devoted exclusively to this work would have to be established. In 1994, after long and unsuccessful negotiations with the Institute for the Czech Language under the Academy of Sciences of the Czech Republic, they became independent and succeeded in founding the Institute of the Czech National Corpus

(ICNC) at the Faculty of Arts, Charles University, in Prague. At first, all of them did their work in their free time as before since there were no means to pay the employees of the Institute. The first employee was a programmer paid by the leading Czech publisher Lidové Noviny. In the future, this publisher would like to take part in the compilation and publication of dictionaries that will come into existence on the basis of the corpus.

The actual work of the ICNC began in October 1996 when thanks to the grant provided by the Grant Agency of the Czech Republic and the Ministry of Education of the Czech Republic it was possible for the Institute to employ their own workers and purchase computer equipment. Furthermore, the sponsorship from Komercní Banka made it possible to convert a basement room in the main building of the Faculty of Arts, Charles University, and furnish it with new furniture and begin work on new 'underground' premises. At present the Institute uses four rooms which serve as offices for five linguists, three mathematicians-informatics specialists, and one administrative college-educated worker. The team is headed by a linguist, Professor František Cermák. In addition to these workers, the Institute relies on two part-time collaborators from the Faculty of Arts and two from the Institute for the Czech Language.

3 State of the Art

The collection of electronic texts began long before the Institute was established. Thanks to donations from Komercní Banka in the early nineties the team could buy 1000 diskettes to archive their first texts. The texts for the corpus were obtained by several means. The seemingly simplest one is the collection of electronic texts from their owners. Yet, at the start of the Institute's activities no-one believed that there would be enough electronic texts available in the general public to make the existence of the corpus possible. Reality surpassed all our expectations. Within a short span of time almost all editorial departments, and all private persons, began to record their texts on computers. It has its drawbacks, however – there are some typesetting programmes, such as QUARK, which our programmers cannot handle. It is extremely difficult to deal with PAGEMAKER. The texts which we want to include in the representative corpus or in the BoC and which are not available in electronic mode are accessed by another method – scanning. They include some works of fiction or texts from out-of-the-way fields. A minimum of texts are prepared by a third method – manual retyping. They include various ephemera or personal correspondence. Part of the texts are also obtained from the Internet.

The texts which are acquired for the Czech National Corpus (CNC) have to be covered by contracts made between the Institute and the original owners of the electronic texts. The contract is to guarantee that the texts will be used for academic research, that they will not be used for commercial purposes and will not be made available to a third party. These contracts are made with individual publishers, editorial boards of newspapers and magazines, sometimes with publishing corporations owning the copyright to several books. Some contracts are made directly with the authors of the texts, when the text is obtained via the Internet or by scanning. At any rate, by creating a corpus from which the texts are inseparable as a whole - inasmuch as the ultimate outcome of our work will be KWIC lists (Key Word In Context), or various types of frequency lists - the result of the work will be a new product to which the copyright will be the property of the Institute. Our activities do not infringe on the rights of the original owners of the texts for it does not restrict the use of their copyright.

4 The Type of Corpus

SYN2000 and the BoC are synchronic corpora. They include written texts of contemporary Czech. At present (spring 2000) the BoC comprises some 330 million word forms in a set of texts from about 260 publishers.
The texts in SYN2000 are divided into the following spheres: (1) Creative sphere (15%): fiction, poetry, drama. (2) Informative sphere (85%) includes non-fiction style (25%), life style, technology, humanities, natural sciences, arts, economy, law, security, religion; and newspapers (60%).

The representativeness of the SYN2000 corpus composition has been determined by sociological surveys. The guiding principle for inclusion is the way the language is perceived by users. The corpus will comprise journalistic and technical texts published after 1989, fiction will include texts dating from 1959 up to the present. The period of 1959 to 1989 will form 1/6 of the total of fiction texts. The year of birth of the author of a fiction text to be included in the contemporary corpus must not be earlier than 1883.

5 Further Stages of Text Processing

All electronic versions of the texts are first of all stored in the archives of the CNC and recorded in the database of text information. Then the texts are converted into a uniform text format (called intermediate format). This format is used by the linguists preparing external annotations when supplying the essential data about each text in the database in the form of abbreviations: *corpus type, verse form, text type, genre, medium, sex of author, original language of text, date of publication, date of first publication, opus.* Opus is the specific mark of each text which connects the text in the corpus with the database, allowing the future user to recover all database information about the text when searching the corpus. The programmers then append some of the information in the database to the text in an intermediate format and the text is then converted into an SGML form (Standard Generalized Mark-up Language):

- A sample of the head section of a document:

```
<doc file="S/B/1991/havel" id=001 >   # file designation and
                                      # the number of the document in it
<a >                                  # beginning of the head section of the document
<mod >S                               # synchronic (contemporary language)
<txtype >NOV                          # novel or another text unit
<genre >BIO                           # biography
<med >B                               # medium book
<authsex >F                           # the author is a woman
<temp >1991                           # year of publication 1991
<firsted >1991                        # year of first publication 1991
<authname >Kriseová Eda               # author's name
<opus >havel                          # name of the file in which the document is found
                                      #and its identification in the database
<id >001                              # number of the document in the file
</a >                                 # end of the head section
```

  Using texts prepared in this way as a basis, the program CQP (Corpus Query Processor) from IMS Stuttgart is used to compile the actual corpus.

- Explanation of the marks in the sample:
  <f > this mark is followed by a concrete occurrence of the word
  <f cap > the word begins with a capital letter
  <MDl aj > this mark is followed by a lemma
  <MDt aj > this mark is followed by a grammatical mark
- Examples of grammatical marks:
  <MDt aj > NNMS3–A– noun. common, masculine, singular, dative,-,-,-,-,-,affirmative,-,-,-,-
- A concordance sample

#———

Corpus:: Syn2000 # Query: palec # CQP: "palec"

Size: 445

Context: left 25 Characters, right 25 Characters

User: Marie Kopřivová # Date: 07.04.00 10:12

———

Spustila okénko u řidiče o	**<palec >**	dolů kvůli vzduchu a zamáčkla
něco či nic , nebude to náš	**<palec >**	.Jistě že ze všech
zvolna nalevo a napravo a	**<palec >**	olizovala . Měla zavřené
jedinou marku ! " ukazoval	**<palec >**	pravé ruky druhý z nich ,
Laurie mu ukázala vztyčený	**<palec >**	. Přes svoji kocovinu se
své výši , což bylo dobře o	**<palec >**	víc , než měřili oni , a
Ukázal jsem mu vztyčený	**<palec >**	. Zatímco se na palubu
kluk si opřel ukazovák a	**<palec >**	o čelo a schválně vypadal
chtěla si jej zkusit . Ale	**<palec >**	se jí do boty nevešel , střevíček
má. Zbyl mi na ní akorát	**<palec >**	, řek chraplavě . Jak se hlubokou
komu přihodil, byl okopnutý	**<palec >**	nebo vyražený dech , jak
večer Obr Koloděj mi šlápl na	**<palec >**	. Co jste učinil proti tomuto
Drápovité prsty a vbočený	**<palec >**	byly zjištěny u více než
rukou . Do tlamy vsuneme	**<palec >**	, zbylé prsty sevřeme pod
(dots per inch - bodů na	**<palec >**	) . Výjimkou však již nejsou
o pánev tvaru umyvadla ,	**<palec >**	ve srovnání s ostatními primáty
část svíčkové se nazývá	**<palec >**	, tato část je vhodná na
vodič tak , aby odtažený	**<palec >**	byl ve směru proudu . Potom
Většina dětí však dumlá	**<palec >**	i v druhém , někdy i ve třetím
Na straně k dlani je	**<palec >**	opatřen navíc polštářkem
sám zvolí (např . svůj	**<palec >**	, tužku , atd .) . Zeptejte
fialové sako a na krku na	**<palec >**	tlustý zlatý řetěz . Dneska
Nilský kříž stiskneme mezi	**<palec >**	a ukazováček a za určitou
že hůl nebyla silnější než	**<palec >**	muže . Podle stejného zákona
jednotky má pouze zraněný	**<palec >**	u ruky. Demonstrace měly
k zemi - smrt . Dokud byl	**<palec >**	schovaný v dlani ,

6 Morphological Annotation

The present SYN2000 is morphologically tagged by means of a morphological lemmatizer and an automatic disambiguator operating on the basis of probabilistic methods. Accordingly, each word form in SYN2000 is provided with a lemma (= the representative dictionary form of a

word, e.g. the infinitive) and a tag (morphological mark) consisting of 15 positions for various morphological categories of particular word classes. But this part of our project has not been resolved in a satisfactory way, yet.

- A sample text with structural and morphological marks in the SGML format

```
<c >                                          # beginning of text
<p n=1 >                                      # first paragraph
<s id=id="S/B/1991/havel:001-p1s1" >          # beginning of sentence with its full identification
(data from head section]
<f cap >Eda <MDl aj >Eda_;Y<MDt aj >NNMS1——A—-
<f cap >Kriseová <MDl aj >Kriseová_;S <MDt aj >NNFS1——A—-
<s id=id="S/B/1991/havel:001-p1s2" ><i >i     # typographic mark - italics
<f cap >Vaškovi <MDl aj >Vašek_;Y<MDt aj >NNMS3——A—-
<i >/i                                        # typographic mark - end of italics
<p n=2 >                                      # second paragraph
<s id=id="S/B/1991/havel:001-p2s1" >          # end of sentence
<f cap >Motto<MDl aj >motto<MDt aj >NNNS1——A—-
<D >                                          # in front of the following sign
                                              # there was no space
<d >:<MDl aj >: <MDt aj >Z:————————-
<p n=3 >                                      # third paragraph
<s id=id="S/B/1991/havel:001-p3s1" >          # beginning of sentence
<f cap >Jednou<MDl aj >jeden‘1<MDt aj >ClFS7———-
<f >navštívil<MDl aj >navštívit_:W <MDt aj >VpYS—XR-AA—
```

- Explanation of the marks in the sample:

<f >	this mark is followed by a concrete occurrence of the word
<f cap >	the word begins with a capital letter
<MDl aj >	this mark is followed by a lemma
<MDt aj >	this mark is followed by a grammatical mark

7 Lexical Saturation

Several tests have been made to ascertain the lexical saturation of the corpus. The spheres selected for the testing included: abstract words - sounds, smells; parts of the human body; the two commonest Czech swear words. These tests will be further expanded and the results shown at our presentation.

8 Work To Be Done

The representative corpus SYN2000 is an ideal source for the compilation of a frequency list of the contemporary Czech language. It is the first task for which the corpus will be used. In addition, our Bank of Czech comprises the complete works of our foremost writers: Karel Capek,

Bohumil Hrabal and Josef Škvoreckorecký. Having their works in electronic form presents a challenge to make a thorough analysis of their language and prove that these particular writers have been of landmark importance for the development of modern Czech. Also, the representative corpus is an ideal material source for a future dictionary. Apart from this, there are other projects in the offing which at the moment are at the stage of contemplation.

There are plans for the Bank of Czech as well as the CNC itself to be further expanded. It will be continuously supplemented with new texts and, as long as we are able to carry on this work, we shall continue to prepare sources of material for the investigation of both present-day and historical Czech. We have no doubts that apart from linguists there will be specialists from other areas of study (e.g. sociology, history, psychology) who will find uses for this material. The application of the material for information science need not be even considered at this point.

9 Acknowledgment

The research reported on in this paper was carried out thanks to the financial support of the grant of the Czech Grant Agency GAČR 405/96/K214.

References

[Blatná 1998] Renata Blatná, "Textové korpusy slovanských jazykù", in *Slavica Pragensia* (Prague) 1993, 190-194

[Čermák 1995] František Čermák: "Jazykový korpus: Prostředek a zdroj poznání." *SaS* pp. 119-140

[Čermák 1997] František Čermák and Jan Králík and Karel Kučera: "Recepce současné češtiny a reprezentativnost korpusu." *SaS*, (Prague 1997): 117- 124

[Čermák 1997] František Čermák: "Czech National Corpus: A Case in Many Contexts." *International journal of Corpus Linguistics*, 1997, Vol. 2 (2): 181-197

[Čermák 1998] František Èermák and PetrKubíček, P.: "Jazykový korpus a škola." *Český jazyk a literatura*, (Prague): 84-92

[Čermák 1998] František Čermák: "Czech National Corpus: Its Character, Goal and Background." In *Text, Speech, Dialogue*, (Proceedings, Brno): 9-14.

[Čermák 1995] František Čermák and Jana Klímová and Vladimír Petkevič eds., *Studie z korpusové lingvistiky*, Studies in Corpus Linguistics: (in print).

[Hlaváčová 1998] Jaroslava Hlaváčová, "Technical Insights into the Birth of a Corpus", In: *Text, Speech, Dialogue*, (Proceedings, Brno 1998): 55- 60

[Hlaváčová 1999] Jaroslava Hlaváčová and Pavel Rychlý, "Dispersion of Words in a Language Corpus", In: *Text, Speech, Dialogue*, (Proceedings, Plzeò 1999): 321-325

[Šulc 1999] Michal Šulc, *Korpusová lingvistika*, (Karolinum, Praha 1999)