Corpus of 19th-century Czech Texts: Problems and Solutions

Karel Kučera, Martin Stluka
Charles University, Czech National Corpus Institute
náměstí Jana Palacha 2, 116 38 Praha 1, Czech Republic
E-mail: karel.kucera@ff.cuni.cz, martin.stluka@ff.cuni.cz

Abstract

Although the Czech language of the 19th century represents the roots of modern Czech and many features of the 20th- and 21st-century language cannot be properly understood without this historical background, the 19th-century Czech has not been thoroughly and consistently researched so far. The long-term project of a corpus of 19th-century Czech printed texts, currently in its third year, is intended to stimulate the research as well as to provide a firm material basis for it. The reason why, in our opinion, the project is worth mentioning is that it is faced with an unusual concentration of problems following mostly from the fact that the 19th century was arguably the most tumultuous period in the history of Czech, as well as from the fact that Czech is a highly inflectional language with a long history of sound changes, orthography reforms and rather discontinuous development of its vocabulary. The paper will briefly characterize the general background of the problems and present the reasoning behind the solutions that have been implemented in the ongoing project.

Keywords: corpus, Czech, 19th century

1. General background

The Czech language found itself in a state of profound change in the 19th century. After being all but reduced to everyday use and homiletics for more than one century, it was still struggling for full-fledged existence in the early 1800s, being Germanized and practically unused in the prestigious domains of scientific and technical writing as well as 'high' poetry and prose.

However, in the course of the century, as a result of an extensive national revival movement, Czech was going through intensive de-Germanization and refinement, absorbing thousands of neologisms, many of them short-lived, which dramatically augmented and diversified its vocabulary. At the end of the 19th century, following several decades of rather turbulent development, Czech could be characterized as a fully developed language with an extensive word-stock which stood comparison with vocabularies of languages like German or French, was stabilized in several authoritative multi-volume dictionaries and was capable both of satisfying the terminological demands of technical or scientific writing and of providing a rich, stylistically diversified array of linguistic means for the then prose and poetry.

In addition to the extensive changes in the vocabulary, three cardinal spelling reforms were implemented in the Czech language during the first half of the 19th century (in 1809, 1843 and 1849) – reforms that fundamentally altered the appearance of Czech texts for the later centuries. Among the most conspicuous results of the three reforms were a radical change in the use of several high-frequency letters (*g*, *j*, *w*, *y*, and *ý*), the replacement of the digraphs *au* and *ſſ* respectively with *ou* and *š*, and the abandoning of the letters *ſ* and *g* with caron (Unicode U+01E7) counterbalanced by the introduction into the Czech alphabet of the letter *í* and re-introduction of the letter *v* (the letter *v* used to mark the word-initial phoneme /u/ before). The following four spellings of the same model sentence (meaning 'Borrowed words are usually printed in a different font in older prints.') illustrate the effects of implementation of the three orthographic reforms in 19th-century texts:

pre-1809: *We starſſjch tiſcých gſau cyzý ſlowa obwykle tiſſtěna odliſſným pjſmem.*
pre-1843: *We staršjch tiscjch gsau cizj slowa obwykle tištěna odlišným pjsmem.*
pre-1849: *We starších tiscích jsau cizí slowa obwykle tištěna odlišným písmem.*
post-1849: *Ve starších tiscích jsou cizí slova obvykle tištěna odlišným písmem.*

Understandably, in many texts printed or written in the years following each reform, the use of the letters influenced by the changes was rather unsystematic, with the pre-reform spellings gradually giving way to the post-reform ones. In unofficial handwritten texts the transition could have taken up to several decades.) Moreover, even in literary supplements of newspapers and magazines in which the use of post-reform orthography had come to be the norm, parts of extensive texts such as novels were reprinted for some time in their pre-reform spelling because they were readily available in the form in which they had been typeset for their book edition several years before.

2. The texts

Seen from the perspective of the present corpus project and its need for authentic 19th-century texts, both the changeable spelling and the rather unsettled ever-growing vocabulary of the 19th-century Czech represent major problems for optical character recognition (OCR) of the period prints, which is especially true of the OCR systems using extensive built-in modern lexica to reduce the number of OCR misinterpretations. In fact, the use of modern lexica in OCR proved to be largely counterproductive when the systems were applied to documents like Czech prints of the first half of the 19th century, that is, documents with a high percentage of obsolete words and word forms and a number of letters having different values than today. However, as has been convincingly shown,[1] the success rate of OCR can be significantly increased if period lexica, including the validated period grammatical forms and proper period spelling, are used instead of the modern ones in the OCR systems.

At the time being, given the unsatisfactory results of OCR

[1] For testing, results and discussion of the overall role of the period lexica in OCR see, for example, IMPACT (2011).

and a rather limited number of 19th-century authentic texts available in electronic form, the Czech corpus is being built of texts which have been OCRed and manually corrected (most post-1850 texts) and of texts which have been manually transcribed. The latter are practically all pre-1850 texts, in which the performance of OCR is extremely poor not just because of the abovementioned language-related reasons but also because of technical problems like ageing and crumbling of paper, rather low quality of many pre-1850 Czech prints and last but not least because of the use of obsolete typefaces. This includes Czech Fraktur, a typeface noticeably different from the German standards and consequently more problematic for existing OCR systems, even those capable of handling the German Fraktur.

A large proportion of manual processing of the texts slows down the progress of the project but, on the other hand, it makes it possible to equip them not only with standard metadata such as the title of the document, time of its origin, author's name or pagination, but also with a number of codes designed to mark headings, footnotes, tables or other special arrangement of parts of the text, emendations of typographical errors, parts of texts written in verse or a foreign language, illegible or missing parts of texts etc.

3. Transcription vs. transliteration

Within the 19th-century corpus project, the pre-1850 Czech texts have been transcribed (in the narrow sense of the word), i.e. the modern Czech writing system – contemporary Czech alphabet and the standard phonological values of its letters – has been used to represent the phonological values of the original spelling found in the texts. Thus, for example, the pre-1850 ways of spelling of the word meaning 'related, connected' (*ſauwiſýcý*, *sauwisýcý*, *ſauwiſjcj, sauwisjcj*, *sauwisící…*) are all transcribed using a single modern form *souvisící*. In this way, transcription filters out the multitude of period spelling variants and irregularities. This is in conformity with the principle applied in the entire diachronic section of the Czech National Corpus (CNC), according to which transcription is employed in all the documents written or printed prior to the last systemic change of Czech spelling, that is before the 1849 reform which was the last to change the phonological values of individual letters of the Czech alphabet.

On the other hand, the documents employing the post-1849 spelling are transliterated, not transcribed, i.e. their period spelling is preserved, with the exception of the letters non-existent in the Czech alphabet of today (in the 19th century, this is the case of the abovementioned letter *ſ* which is standardly replaced with its newer substitute *s*). However, as indicated in more detail below, the fact that the phonological values of Czech letters has not changed since 1849 in no way means that modern spellings of all words and word forms are now completely identical with the ones used one hundred or one hundred fifty years ago.

The reasoning behind the above choice made between transliteration and transcription is that for a great part of the seven-century history of Czech texts the Czech authors, scribes and printers did not adhere to any strict orthographic rules in the modern sense. The use of transliteration in such documents makes text searches extremely complicated if not impossible, especially in the early Czech documents preserved from the end 13th through 15th centuries.

The use of transcription in corpus texts effectively means that the corpora included in the diachronic part of CNC, including the pre-1850 part of the 19th-century corpus, are not intended for spelling-oriented research. However, to at least partly accommodate such research needs a small specialized repository of pre-1850 transliterated text samples, kept side by side with their transcribed versions, is being compiled in CNC.

4. Lemmatization and hyperlemmatization

Generally speaking, lemmatization plays a much more important part in highly inflectional languages like Czech than in isolating languages such as English. In the case of Czech, some verbs could have had up to one hundred inflected (including negated) forms in the past, the amount being further increased in the verbs the paradigms of which encompass a number of coexisting phonological variants resulting from a number of historical sound changes. By way of example: in Old Czech the present-tense indicative 1st person singular form of the verb *pokračovat* 'to proceed, to continue' could be used in the four phonological variants, namely *pokračuju, pokračuju, pokračiju* or *pokračiji*, with two of these variants – *pokračuju* and *pokračuji* – still coexisting in modern Czech.

The way to lemmatization of the 19th-century Czech texts follows a procedure consisting in the following three steps.

4.1 Sorting out the vocabulary of 19th-century texts

In order to build a basis for lemmatization word lists, a standard lemmatizer using modern-Czech lexicon was applied to a sample of 19th-century texts encompassing more then one million of word forms. The tool succeeded in identifying most word forms used in the texts but more than 100,000 word forms remained unidentified. At present, the results are still being manually proof-read and corrected.

Predictably, a large proportion of the unrecognized forms were those of proper names, especially personal names, primarily surnames – a phenomenon reflecting the importance or fame of particular individuals in the 19th century, as well as their diminishing popularity and/or the decline of their historical significance in the 20th and 21st centuries. Thus, understandably, the modern Czech lemmatizer was able to identify names of such personages as for example *Darwin* or *Cézanne*, but did not recognize names of many 19th-century politicians such as e.g. *Broglie* (French minister of external affairs), *Isturic* (Spanish state attorney), military leaders as *Balazé* (French general) or aristocrats like counts *Montalembert*, *Lavradius* or *Wurmbrand*.

A rather specific group of unidentified personal names consisted of unofficial patriotic middle names composed of Czech or Slavic roots, such as *Krasoslav* (from the roots of the words *krásný* 'beautiful' and *slavný* 'famous, glorious') or *Silorád* (*silný* 'strong' and *rád* 'liking'), the use of which was mostly limited to the national revivalists active in the 19th century.

Now obsolete Czech geographical names used to denote

foreign countries and cities like *Kitaj/Kytaj* 'China', *Francouzy* 'France', *Angora* 'Ankara' or *Frankobrod* 'Frankfurt' formed another significant group among proper names unrecognized by the modern Czech lemmatizer.

Apart from proper names, most of the unidentified word forms fell into the following six groups:

(a) pronounced archaisms and no longer used words (*špehéř* 'spy', *blahověst* 'apostle'), including on the one hand older borrowings from German like *frajmaur* 'freemason', *fraucimora* '(young) lady', and on the other unsuccessful coinages offered by 19th-century national revivalists attempting to replace German and other foreign nouns with Czech neologisms (*bijna* 'battery', *krasouma* 'aesthetics') and to enrich the vocabulary of 'high' literature, particularly that of poetry, with hundreds of poetic compounds, mostly adjectives, such as *divovábný* ('miraculously alluring'), *blahočarný* ('miraculously delighting'), *hrdoslavný* ('proud and famous') or *hvězdooký* ('star-eyed');

(b) now obsolete parallel word formations, some of them inherited from Old Czech (e.g. *náhrdlek* 'necklace'), some typical of the 19th century language (among them tens or perhaps hundreds of borrowed adjectives derived from international roots with the suffix *-ný* (e.g. *neutrálný* 'neutral', *kriminálný* 'criminal', *materiálný* 'material' etc.), which were later replaced with equivalent derivations employing the suffix *-ní* (*neutrální, kriminální, materální*);

(c) no longer used phonological variants of words like *kněhkupectví* 'bookstore' (modern equivalent *knihkupectví*), *pondrava* 'grub' (modern *ponrava*), *citedlný* ('perceptible, considerable', modern *citelný* etc.) or *jenerální* (modern *generální*);

(d) no longer used spelling variants of words such as *ssát* 'to suck'(modern spelling *sát*) or *neurosa* 'neurosis' (modern *neuróza*), including a number of borrowings spelled with the original double consonants in the 19th century, such as *irracionální*, *illegální*, later replaced with the spelling using single consonants (*iracionální, ilegální*);

(e) combinations of word forms with non-independent particles such as *-ť, -tě, -ž* (e.g. *dámť, byltě, dejž*) the use of which is practically non-existent in modern Czech;

(f) obsolete grammatical forms like the dual number in nouns (e.g. instrumental forms *rtoma* 'lips', *kačátkoma* 'ducklings') or now unusual combinations of particular inflectional endings with particular words, especially nouns and verbs, as for example in *adventě* ('Advent', locative singular form), *koněmi* ('horse', instrumental plural form), *andělmi* ('angel', instrumental plural form), *Achillesa* ('Achilles', genitive and accusative singular form), *honosejí se* ('to pride oneself', 3rd person plural, present tense).

4.2 Expansion and corrections

The second step in the building of the basis for lemmatization word lists consisted in assigning proper lemmata to the word forms that had remained unrecognized by the modern-Czech lemmatizer. The assigning of the lemmata was based on 19th-century grammars, vocabularies and linguistic historical studies. On the same basis of information about the period language, each lemma was then expanded into a full paradigm, with each paradigm encompassing forms no longer used in modern Czech and each form of the paradigm linked to its lemma.

The second step also included the handling of the word forms that had been recognized by the lemmatizer. The forms were checked for possible misinterpretations and corrected accordingly. The misinterpretations were typically anachronisms, i.e. lemmata that were chronologically out of place, the 19th-century forms being sometimes misinterpreted as forms of words that came into use later on. A conspicuous example was the interpretation of the form *robota* ('corvée, unpaid labor exacted by a feudal lord') as a form of the word *robot* which was coined at the end of the second decade of the 20th century.

4.3 Building the word lists

The third step consisted in distributing the lemmatized paradigms to three word lists intended to be used by the program to lemmatize OCRed or transcribed 19th-century texts and to identify in them further forms missing from the lists. The three word lists that are being built as part of the project include two special lists (one encompassing proper names, the other abbreviations) and lastly, the main general list that includes the rest of the vocabulary found in the processed texts.

The three lemmatization lists grow with every 19th-century text prepared for the corpus. With every processed text, the above three steps are repeated so that further unrecognized word forms found in the text are subsequently lemmatized, expanded into full paradigms and distributed to the three lemmatization lists. On the other hand, as the lists keep growing, close inspection and potential disposal of some of their items is becoming more and more important, in particular to check for and prevent potential overgeneration. Rare forms (such as present or past transgressives, in the Czech case) of extremely rare words are among first candidates for removal from the general list, especially if they are homonymous with common forms of high-frequency words without themselves having been attested in the texts. The homonymy of the Czech preposition *podle*, roughly corresponding to the English *along* or *according to*, and the transgressive form *podle* of the uncommon verb *podlít* 'to spend a short time somewhere' stand as good examples.

As to the lemmatized corpus texts, step 3 is going to be followed by computer-aided disambiguation of the word forms that have been assigned more than one lemma by the lemmatizer. The disambiguation tool is currently in the state of being tested and fine-tuned to make it as efficient and user-friendly as possible. The fine-tuning is being given a great deal of attention because homonymy among word forms found in the inflectional Czech language is much more extensive than in analytical languages like English, and morphological disambiguation of forms found in Czech texts is consequently much more complicated as well as time-consuming.

4.4 The hyperlemma

An extended concept of the lemma, dubbed *hyperlemma*, is applied in the whole of the diachronic section of CNC, including the 19th-century corpus; for discussion see Kučera (2007). Unlike the lemma, the hyperlemma (formally identical with the canonical modern form of the word) groups together not only the contemporary and past paradigmatic forms belonging to the same lemma, but also their present and historical phonological variants. Thus, for example, in the diachronic corpus including texts from the beginning of the 14th century through the 1980s, the Old Czech nominative singular form of the word *kóň* 'horse', the newer form *kuoň* (found mainly in the texts of the 15th and 16th centuries) and the modern form *kůň* would be all assigned the same lemma *kůň*. As a result, the end-user of the corpus will be able to use a single hyperlemma query (*kůň*, in this case) to retrieve all the above forms as well as all the present, historical, literary, colloquial or dialectal inflected forms such as *koňa, koňu, koňmi, koňoma, koněma, konima* etc.

Although the primary function of hyperlemmatization is to group the inflected forms and their phonological variants under one canonical form, hyperlemmata in the 19th-century corpus will also include all the spelling variants of words found in post-1850 texts. As mentioned above, the texts from the second half of the 19th century are transliterated, not transcribed, so that the peculiarities of their orthography are being preserved. Again, the hyperlemma makes it possible to retrieve all the spelling variants of the searched-for word by performing just one hyperlemma query. This is especially useful in many words starting in *zp-, zt-* in modern Czech (spelled, often unsystematically, with *sp-* or *st-* still in the second half of the 19th century), and in borrowings, including a large number of technical terms, in which their original foreign spelling used to be preserved in the 19th century – a convention that in most borrowings was abandoned for the standard New-Czech phonological spelling after the reform of 1957. As a result of this reform, for example, the 19th-century spellings like *aesthet, coelibat/caelibat, engagement, affaira, jalousie, isotherma* were changed to *estét, celibát, angažmá, aféra, žaluzie, izoterma.*

5. The current state of the project

19th century texts including more than 1,000,000 word forms have been transcribed or OCRed and manually corrected. As indicated above, the texts have been lemmatized with a modern-Czech lemmatizer and the list of unrecognized forms (at present amounting to more than 100,000 items) is being proof-read, filtered and lemmatized, with about 30,000 items still remaining to be processed. A set of 19th-century Czech model paradigms is being completed to be used to expand into full paradigms all the lemmata found in the texts.

The corpus of 19th-century Czech printed texts is planned to be accessible, including minimum 700,000 lemmatized word forms, by 2016 and to grow further afterwards, possibly with morphological tagging added.

Modified spin-offs of the work on the 19th-century corpus (namely modified word lists including spellings and morphological forms of authentic 19th-century words) are planned to be used, in cooperation with the National Library in Prague, to improve the access to 19th-century printed texts within the project *Tools for Accessibility of Printed Texts from the 19th Century*, part of the *Applied Research and Development of National and Cultural Identity Programme (NAKI)* funded by the Czech Ministry of Education. For details see http://kramerius-info.nkp.cz/projekt-naki or http://www.isvav.cz/programmeDetail. do?rowId=DF.).

6. Acknowledgements

This paper was written under the scope of the infrastructure *Czech National Corpus* (LM2011023), a part of the the *Projects of Large Infrastructures* financed by the Ministry of Education, Youth and Sports of the Czech Republic.

7. References

IMPACT 2011 Project Periodic Report (http://www.impact-project.eu/uploads/media/IMPACT_Annual report_2011_Publishable_summary_01.pdf)

Kučera, K. (2007). Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora. In Levická, J., Garabík, R. (eds): *Computer Treatment of Slavic and East European Languages*. 2007, Bratislava: Tribun, pp. 121-125.