Challenges in Accessing Information in Digitized 19th-Century Czech Texts

Karel Kučera
Charles University in Prague, Czech Republic
nám. Jana Palacha 2
11638 Praha 1
00420 224241490
karel.kucera@ff.cuni.cz

Martin Stluka
Charles University in Prague, Czech Republic
nám. Jana Palacha 2
11638 Praha 1
00420 224241490
martin.stluka@ff.cuni.cz

ABSTRACT

This short paper describes problems arising in optical character recognition of and information retrieval from historical texts in languages with rich morphology, rather discontinuous lexical development and a long history of spelling reforms. In a work-in-progress manner, the problems and proposed linguistic solutions are shown on the example of the current project focused on improving the access to digitized Czech prints from the 19th century and the first half of the 20th century.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing, dictionaries.*

General Terms

Languages

Keywords

Information Retrieval, Known-Item Retrieval, Historical Text, Lemma, Hyperlemma

1. INTRODUCTION

As has been recently pointed out, in spite of undeniable progress over the last few years, the state-of-the-art software for optical character recognition still does not provide satisfactory results in transformation of historical books, magazines and newspapers into searchable and editable text. [1] Low quality of old prints, use of historical typefaces (such as the Gothic script in its numerous regional variants), special characters and ligatures, ageing of paper and page curl are usually mentioned among the major technical OCR difficulties being worked upon. However, the whole problem also has a linguistic aspect, since the results of OCR can be substantially improved by linguistic information, as has been proved in OCR of modern texts in tens of languages where extensive language-specific lists of paradigmatic word forms have been used to optimize the OCR 'best guesses' by comparing the resulting interpretations of character strings to existing word forms.

Long overshadowed both by the abovementioned technical issues and the more urgent demand to achieve high dependability of OCR results in modern texts, the problems of using historical lexica in noisy old text data has been fundamentally addressed only lately [2]. At the same time, there has been designed a plausible way of building period-specific lexica from manually corrected ground-truth texts and/or from historical dictionaries (if available), [3] but so far few lexica have been compiled and tested in practice. One notable exception was the series of tests performed under the European IMPACT program, which included historical lexica for nine languages and showed that "deployment of historical lexica improves the state-of-the-art of both OCR and IR". [4]

Generally speaking, the deployment of historical lexica for OCR and IR purposes should help to solve the language-related noise coming from

- archaic words (such as *eftsoon* 'again; at once' or *thine*, to give English examples) and word formations (*disobediency, labourous* etc.)
- archaic inflectional forms (e.g. *maketh, makest, bespake*) and
- archaic spellings like *oeconomic, aeternal, to-morrow, applyed, fruitfull, hydraulick* etc.

To compile a historical lexicon may represent different degrees of challenge, depending on how numerous and complicated the differences from the present language are in the above three areas, as well as on some other factors such as the availability of dictionaries and grammars from the particular period or accessibility of computer processable editions of historical texts. Moreover, the challenge is different in different types languages: the compilation of a lexicon may be relatively trivial in predominantly isolating languages like English, where inflected words follow a very limited number of paradigms with a very limited number of forms in each of them, as compared to highly inflectional languages, with up to several tens of forms in each of tens or even hundreds of paradigms diversified by grammatical categories, sound changes, variations and fluctuations.

In the following, we elaborate on the specific problems connected with the historical lexicon building in Czech, as they are approached in the project *Tools for Accessibility of Printed Texts from the 19th Century and the First Half of the 20th Century*. [5]

2. THE CASE OF CZECH

2.1 General Background

The texts from the 19th century and the first half of the 20th century which are in the focus of the aforesaid Czech project, are not too far removed from the present texts, and given the availability of several 19th- and 20th-century Czech dictionaries and grammars, it may seem to be a relatively unsophisticated task to compile a historical lexicon for OCR and IR purposes. At a closer look, however, the task is not quite as trivial, mainly due to historical reasons. At the beginning of the 19th century, German and Latin were the high-status languages in the Czech lands, while Czech was struggling for full-fledged existence, being practically unused in technical and scientific writing, 'high' poetry or prose. However, only 50 years later, following a vocabulary explosion, intensive de-Germanization and wide-ranging refinement resulting from the National Revival movement, the situation was completely different. Generally, this line of development continued, if in a less intensive way, in the second half of the 19th century, but while the Czech vocabulary kept growing in a number of branches of technology and science, more German loan words and many of the unsuccessful neologisms coined in the earlier period were being abandoned. Considering the modern Czech language of the 1st half of the 20th century, with its fully developed terminology and variety of language styles, one can conclude that at least three different lexica should be created to accommodate the OCR and IR needs, each covering a period of about 50 years.

Nevertheless, one more important factor needs to be taken into consideration in the Czech case, namely the three deep-cutting reforms of orthography implemented in 1809, 1843 and 1849, which changed the use of several high-frequency letters and, consequently, the spelling of tens of thousands of word forms. The following four spellings of the same example sentence (meaning 'All this happened not by her fault, but someone else's') stand as telling samples of how pronounced the changes were:

until 1809:	*To wſſe ſe ſtalo ne gegj, ale cyzý winau.*
until 1843:	*To wše se stalo ne gegj, ale cizj winau.*
until 1849:	*To wše se stalo ne její, ale cizí winau.*
after 1849:	*To vše se stalo ne její, ale cizí vinou.*

As a consequence, four lexica, each of them reflecting different spellings and rather different vocabularies, are being worked on to cover the 150-year period. In fact, four more lexica will be compiled, each of them including both the pre-reform and post-reform spelling variants. These lexica will be used in OCR and IR with the prints from the short transitory periods when the orthographic reforms were only being introduced and the older and newer spellings were used in the same texts.

2.2 Building the Lexica

The compilation of each of the four Czech historical lexica is based on the combined use of lists of headwords obtained from 19th- and 20th-century dictionaries and/or lists of word forms extracted from available OCRed or manually transliterated historical texts. After a proofreading, the lists are processed in the following four steps:

- Each word form on the list is assigned a modern lemma. i.e. a citation/dictionary form written in modern spelling. Applying this approach, the English forms *make, makes, made, making* would be all assigned the lemma *make*; the modern lemma for historical spellings as *oeconomic, aeternal, to-morrow, applyed, fruitfull, hydraulick* would be *economic, eternal, tomorrow, apply, fruitful, hydraulic* etc. The unrecognized forms in all the lexica are reviewed and either discarded as noise or accepted, corrected (in the case of OCR misreadings) and manually lemmatized. The procedure for the words and word forms printed in one of the pre-1849 spellings is different in that they are first converted into modern spelling and only then (automatically or manually) assigned a lemma.
- The lemmata are then distributed into groups according to their paradigmatic characteristics, i.e. according to the way they inflect. Special attention is given to integrating all old forms (in English, for example, *maketh, makest*) into the paradigms.
- Using a paradigm-specific utility for each of the groups, the lemmata are expanded into full paradigms, many of which in the case of Czech include up to several tens of forms. The modern lemma accompanies each generated form, so that the resulting lines of the lexicon have the format "form;lemma", i.e. for example *vílou;víla*.
- Finally, the full paradigms based on the transcribed pre-1849 spelling forms (cf. step one above) are converted back to the spelling identical with the one originally used. Depending on the original spelling, the line quoted as an example in the previous paragraph would then be changed in one of the following: *wjlau;víla* (pre-1843 spelling), *wílau;víla* (pre-1849 spelling) or *vílou;víla* (post-1849 spelling).

Ideally, the resulting initial versions of the lexica at this point include complete paradigms of all the words found in the texts and/or dictionaries used for their compilation. However, the lexica are paradoxically far from being ideal, especially from the IR viewpoint.

2.3 Reductions and Additions

Experience with the lexica compiled in the above-described way showed that some rare or unused items (mostly archaisms and neologisms) tend to penetrate into the them as a result of the fact that such words had their own entries in Czech 19th-century dictionaries. This, again, had its historical reasons: especially in the first half of the century, the author of a dictionary might wish not just to reflect the real usage, but also to show that the richness of the Czech vocabulary was comparable to that of German, which may have not been quite true then. As a result, the dictionary in fact partly demonstrated the potential of Czech by including new coinages and centuries-old words, not just the contemporaneous usage.

Experience also showed that the lexica are overgenerated, especially in that they include all the low-frequency forms of low-frequency words. Out of context, such comprihensiveness may be desirable, but in practice it proved counterproductive. In Czech, this is primarily the case of transgressive forms of low-frequency verbs, which may have never been used in Czech texts but are often homonymous with forms of other words, many of them high-frequency ones, such as for example *podle* (transgressive of the rare verb *podlít* 'stay for a short time') and *podle* (high-frequency preposition meaning 'according to' or 'by'). As such, they are potential sources of noise in IR.

On the other hand, in the course of time, thousands of words and forms will have to be added to the initial versions of lexica which, with over 500,000 word forms in each of the four of them, are still somewhat limited as a natural result of the fact that a rather limited number of computer-processable texts and dictionaries were available for their compilation. New items will be added to the lexica from a growing number of texts in the following four years of the project. The general expectation is that most additions will come from technical texts and poetry, but there will no doubt be one more, rather specific group coming from the prose, press and drama that partly reflected the colloquial stratum of the Czech vocabulary of the 19th century. Characterized by hundreds of German loan words, this largely unresearched part of the Czech word-stock was mostly ignored in the 19th-century dictionaries owing to the anti-German linguistic attitudes prevailing during the Czech National Revival and the following decades.

The difficulties presented by the lexica including rare or unused words and forms on the one hand, and missing colloquial words and forms on the other, are different in OCR and IR. In OCR, problems arise if the missing words or the rare/unused words happen to be formally similar to (but not identical with) some common forms, because the similarity may cause OCR misinterpretations. Formal identity (i.e. homonymy) of two or more forms is irrelevant because what matters in OCR is the mere existence of the form, not its meaning(s) or grammatical characteristic(s).

In IR, on the other hand, homonymy is the main source of difficulties as it may cause a considerable increase in the amount of noise in the results of end-users' queries. Formal similarity (not identity) of word forms itself does not present any direct problems for IR, but influences its results indirectly, through the abovementioned OCR misinterpretations.

To reduce these problems, a record will be kept of occurrences of words (lemmata) and their forms in the processed texts, with metadata including the ID of the text, page number and position of the word form on the page as well as information about the text including the year of its publication, text type (belles-lettres, press, science and technology) and domain (natural sciences, medicine, mathematics etc.). The reviewed record will be periodically used to add words and word forms to the existing lexica. Eventually, towards the end of the project it should also also be used for a realistic reduction of the initial lexica to words and forms attested in authentic texts. At the same time, the extensive record, estimated to include more than 5,000,000 word forms by the end of the project, should help to differentiate between generally used words and special vocabularies, as well as between words and forms used during the entire 150-year period and those with a limited life span.

3. LINGUISTIC INFORMATION AND IR

As shown above, in the Czech project the added linguistic information in the lexica consists in assigning a lemma to each word form. As a form representing the entire set of paradigmatic forms of a particular word, the lemma makes it possible to efficiently retrieve all the occurrences of all the forms of the searched word at once – a capacity especially appreciated by end-users performing searches in languages in which words may have numerous forms.

Assigning the correct lemma to all the word forms in the text can also help to remove many of the problems caused by homonymy: in this way, for example, the homonymy in the English *left* ('opposite of right' or past tense of the verb *leave*) can be eliminated. However, to assign the correct lemmata to homonymic words or word forms requires disambiguation, which in the case of historical texts can practically only be manual as, to our knowledge, there exist no acceptably functional historical disambiguation programs for old Czech or other old languages. Since manual disambiguation is far too inefficient in projects where the number of digitized and OCRed pages of old texts amounts to thousands a day, homonymy remains an interfering problem in IR. In the Czech case, for the time being, the homonymic forms are standardly assigned as many lemmata as many paradigms they are part of.

Nonetheless, if the strict linguistic definition of the lemma is stretched a little, the concept can accommodate more end-users' needs than just the clustering of all the forms of a word. Dubbed as "hyperlemma", the extended concept is being implemented in the ongoing lemmatization of the diachronic part of the Czech National Corpus, [6] representing not only the paradigmatic forms of words, but also their phonological and spelling variants used during the seven centuries of Czech texts. Thus, in a hyperlemma query, the user is free to use the modern phonological/spelling form of the lemma (e.g. *angažmá, téma*) to retrieve all the instances of its modern and historical forms and variants (in this case *engagement, engagementu, engagementem…, thema, thematu, thematem…*). The employment of the concept will arguably be even more important in the discussed Czech project than it is in the corpus, because unlike the corpus, the typical users of which are linguists, the body of texts which is in the focus of the project is expected to be used typically by historians and other scientists as well as by journalists and the general public, that is by people without a deeper knowledge of the historical changes in the Czech language.

In view of further problems they may experience when searching for a particular known item in the texts from the 19th and the first half of the 20th century, the following four general situations (and solutions) were considered:

- The word the user is searching for exists in just one phonological and spelling form used now as well as in the 19th century, and none of its paradigmatic, phonological or spelling forms overlaps with any form of any other word. The retrieved forms will be exactly those (s)he is looking for. This is the ideal (and, fortunately, also majority) case presenting no problems.
- The word the user is searching for exists in two or more modern phonological and/or spelling variants with the same meaning and about the same frequency (e.g. *sekera/sekyra* 'ax', *vzdechnout/vzdychnout* 'to sigh', the suffix

-ismus/-izmus '-ism'), or in two or more historical phonological and/or spelling variants of the same meaning and about the same frequency (*čiv/čiva* 'nerve'). There are hundreds of such cases in Czech; in English this is a relatively rare phenomenon (e.g. *ax/axe*) unless one considers the multitude of British and American spelling variants such as *humour/humor, theatre/theater, materialise/materialize* etc. To avoid the problems caused by the rather common situation that the user may not realize the parallel existence of the variants and consequently will miss part of the searched-for information, a record of these variants is being built and used by the search program. After one of such lemmata is keyed in (e.g. *ax*), the program will automatically retrieve all the forms of all the variants (i.e. *ax*, *axe* and *axes*), and the user will be informed about it.

- The word the user is searching for exists in one or more common modern phonological and/or spelling variants, with the same meaning and about the same frequency (e.g. *anděl* 'angel'*, myslet* 'to think') and in one or more infrequent or presently unused (mostly historical) variants of the same meaning (*anjel, myslit*). Many users will not be aware or think of the existence of the latter variant(s), so again, to avoid the risk of missing part of the searched-for information, a record of these variants is used, if in a slightly different procedure. The planned solution is that once the commonly/frequently used lemma (e.g. *anděl*) is keyed in, the search program will retrieve all the forms of all the lemmata (*anděl, anděla, andělovi, andělem…, anjel, anjela, anjelovi, anjelem…*), and the user will be informed about it. On the other hand, if the user keys in the currently unused/infrequent lemma (*anjel,* in this case), the program will only retrieve the forms of this lemma (i.e. *anjel, anjela, anjelovi, anjelem…*). The reasoning behind the latter procedure is that the user is obviously not a complete laymen, knows the form and has a reason to search for it. In case the user wants to retrieve just the forms of the more frequent variant (*anděl*), (s)he can revert to the string-matching query.

- The word the user is searching for only exists in one modern/historical phonological and spelling variant (i.e. it has one lemma), but one or more of its forms are homonymic, i.e. overlap with forms of another lemma, as in the example of *left* ('opposite of right' or past tense of the verb *leave*) given above. Czech as a highly inflectional language has thousands of such homonymic word forms, with some of them being part of four or even five different paradigms, and, as has been stated above, at present there is no practicable way to significantly reduce the noise such forms cause in IR from historical texts. A record of homonymic forms is being compiled for the future use in a disambiguator of historical Czech texts but in the nearest future its use will be mostly limited to informing the user about the problem whenever (s)he is searching for a lemma including homonymic forms.

4. CONCLUSION

While homonymy will remain one of the main problems of IR from historical texts in Czech as well as in many other languages, the expectation is that the results of the Czech project will make known-item retrieval easier for the end user, especially by implementing the abovementioned concept of hyperlemma and by modifying the query based on lists including both contemporary and historical variants. As a result, still on the linguistic ground, the user will be able to find, with a single query, all instances of all attested present and historical forms and spelling/phonological variants of a word – a feature which is not common in similar text collections (with very few exceptions like *encyclopedia* and *encyclopaedia*, several searches must be performed to find different forms like *go, goes, goeth; economy, oeconomy; medieval, mediaeval; peaceful, peacefull* etc. in Google books, Hathi Trust Digital Library, Open Library, the University of Michigan Collection and others). [7]

Last but not least, the lexica and lists being compiled under the Czech project will serve as a basis for the development of a disambiguator for the texts from the 19th century and the first half of the 20th century.

5. REFERENCES

[1] *IMPACT 2011 Project Periodic Report*, 5, http://www.impact-project.eu/uploads/media/IMPACT_Annual_report_2011_Publishable_summary_01.pdf.

[2] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. 2009. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica**.** In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09**.** ACM New York, NY, 69-76**.**

[3] Depuydt, K. 2009. Historical Lexicon Building and How it Improves Access to Text. (Conference *OCR in Mass Digitisation: Challenges between Full Text, Imaging and Language. The Hague*). See presentation at https://www.impact-project.eu/uploads/media/Katrien_Depuydt_Historical_Lexicon_Building.pdf.

[4] *IMPACT 2011 Project Periodic Report*, 11, http://www.impact-project.eu/uploads/media/IMPACT_Annual_report_2011_Publishable_summary_01.pdf.

[5] Part of the of the *Applied Research and Development of National and Cultural Identity Programme (NAKI)* funded by the Czech Ministry of Education. For details see http://www.isvav.cz/programmeDetail.do?rowId=DF and http://kramerius-info.nkp.cz/projekt-naki.

[6] See www.korpus.cz.

[7] Searches performed at http://books.google.cz/books, http://www.hathitrust.org, http://archive.org/details/texts, http://quod.lib.umich.edu/g/genpub?page=simple.