Lexicon construction and corpus annotation of historical language with the CoBaLT editor

Tom Kenter[1], Tomaž Erjavec[2], Maja Žorga Dulmin[3], Darja Fišer[4]

[1] Institute for Dutch Lexicology
Matthias de Vrieshof 3, gebouw 1171, 2311 BZ Leiden
tom.kenter@inl.nl
[2] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si
[3] maja.zorga@gmail.com
[4] Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

Abstract

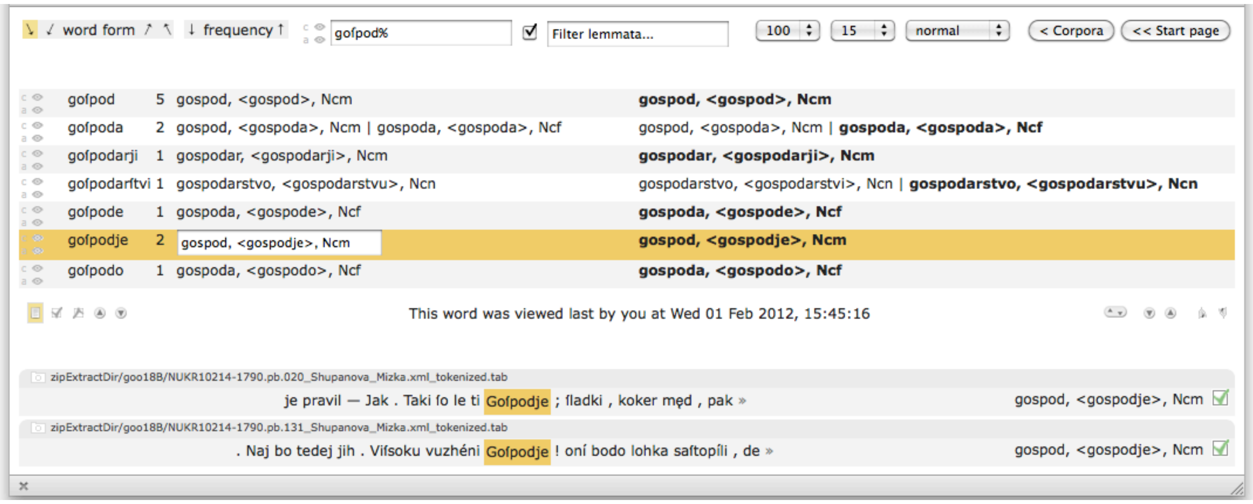
This paper describes a Web-based editor called CoBaLT (Corpus-Based Lexicon Tool), developed to construct corpus-based computational lexica and to correct word-level annotations and transcription errors in corpora. The paper describes the tool as well as our experience in using it to annotate a reference corpus and compile a large lexicon of historical Slovene. The annotations used in our project are modern-day word form equivalent, lemma, part-of-speech tag and optional gloss. The CoBaLT interface is word form oriented and compact. It enables wildcard word searching and sorting according to several criteria, which makes the editing process flexible and efficient. The tool accepts pre-annotated corpora in TEI P5 format and is able to export the corpus and lexicon in TEI P5 as well. The tool is implemented using the LAMP architecture and is freely available for research purposes.

1 Introduction

Processing tools as well as linguistic studies of historical language need language resources, which have to be developed separately for each language, and manually annotated or validated. The two basic resource types are hand-annotated corpora and lexica for historical language, which should contain (at least) information about the modern-day equivalent of a word form and its lemma and part-of-speech (PoS). The first of these is useful for easier reading of historical texts, as well as for enabling already developed modern-day PoS tagging and lemmatisation models to be applied to historical texts. PoS tags make for a better environment for linguistic exploration and enable further levels of annotation, such as tree-banking. They also facilitate lemmatisation, which is especially useful for highly inflecting languages as it abstracts away from the inflectional variants of words, thereby enabling better text searching.

To develop such resources, a good editor is needed that caters to the peculiarities of historical texts. Preferably it would combine the production of annotated corpora and corpus-based lexica. This paper presents CoBaLT, a Web-based editor which has already been used for developing language resources for several languages. We describe it within the framework of developing a gold-standard annotated reference corpus (Erjavec, 2012) and a large lexicon of historical Slovene.

This paper is structured as follows: in the next section we describe the implementation and functionality of CoBaLT. In Section 3 we present the input and output corpus and lexicon formats, in particular from the perspective of our project. In Section 4 we compare existing tools serving a similar purpose to CoBaLT and discuss the advantages and disadvantages of the CoBaLT environment. The last section summarizes and lists our conclusions.



goſpod	5	gospod, <gospod>, Ncm	**gospod, <gospod>, Ncm**
goſpoda	2	gospod, <gospoda>, Ncm | gospoda, <gospoda>, Ncf	gospod, <gospoda>, Ncm | **gospoda, <gospoda>, Ncf**
goſpodarji	1	gospodar, <gospodarji>, Ncm	**gospodar, <gospodarji>, Ncm**
goſpodarſtvi	1	gospodarstvo, <gospodarstvu>, Ncn	gospodarstvo, <gospodarstvi>, Ncn | **gospodarstvo, <gospodarstvu>, Ncn**
goſpode	1	gospoda, <gospode>, Ncf	**gospoda, <gospode>, Ncf**
goſpodje	2	gospod, <gospodje>, Ncm	**gospod, <gospodje>, Ncm**
goſpodo	1	gospoda, <gospodo>, Ncf	**gospoda, <gospodo>, Ncf**

This word was viewed last by you at Wed 01 Feb 2012, 15:45:16

zipExtractDir/goo18B/NUKR10214-1790.pb.020_Shupanova_Mizka.xml_tokenized.tab

je pravil — Jak . Taki ſo le ti Goſpodje ; ſladki , koker mẹd , pak » gospod, <gospodje>, Ncm ✓

zipExtractDir/goo18B/NUKR10214-1790.pb.131_Shupanova_Mizka.xml_tokenized.tab

. Naj bo tedej jih . Viſoku vuzhéni Goſpodje ! oní bodo lohka saſtopíli , de » gospod, <gospodje>, Ncm ✓

Figure 1. CoBaLT interface

2 The CoBaLT tool

2.1 Implementation

CoBaLT is a Web-based editor using the classic LAMP architecture (Linux, Apache, MySQL and PHP). Ajax (Asynchronous JavaScript and XML) technology is used extensively as it enables updating only relevant parts of the screen which increases speed and usability. The code is optimised to work with large datasets and comes with documentation on various settings for MySQL and PHP that enhance handling large data collections. System, project, language-specific details (e.g. the list of valid PoS tags to enable their validation during editing) and some interface settings are encapsulated in a PHP file, making the adaptation of the tool to other environments very easy. However, some knowledge of MySQL is still required, e.g. to add new users to the system which is performed directly in MySQL.

Apart from an Internet browser, no additional software is required at the user side. The interface can be used from various browsers on all major operating systems, although it has been tested primarily on Mozilla Firefox.

2.2 User interface

Apart from logging into the tool and selecting the corpus or file to work on, the CoBaLT user interface is always contained on a single screen. The icons and fields on the screen have associated tooltips.

As shown in Figure 1, the screen is divided in four parts:

1. The upper, "dashboard" part enables ways of organizing the displayed information, i.e. how to sort the word forms, which ones to select, whether to hide certain less interesting word forms (such as numerals), the number of word forms shown, and links back to start pages.
2. The left side of the middle part shows the (selected) historical word forms with their corpus frequencies. This is followed by an editable window giving the modernised word form, lemma, PoS and potential gloss; if the corpus contains distinct annotations for the word form, they are all shown, separated by a pipe symbol. Finally, on the right-hand side, all the possible lexical annotations of the word form are given; those in bold have been validated.
3. The separator between the middle and lower parts shows who has worked last on the selected word form, and gives icons for sorting the word forms in context in the lower part according to a number of criteria: word form, right and left context, analysis and verification.
4. The lower part of the screen shows the selected word form tokens in context together with their analyses in that context and a tick box for validation next to each. Also displayed is the name of the document in which they appear. The arrows next to a context row allow for expanding the context. Clicking on the camera icon at the left side of the row opens the facsimile image.

The separator bar in the middle can be dragged for relative resizing of the middle and lower part.

2.3 Editing in CoBaLT

There is more than one way of editing the analyses assigned to a word form in CoBaLT. The user can work on a specific word form either in the middle screen or in the lower screen, with keyboard shortcuts making the process very efficient. Multiple rows of a word form in context can be quickly selected with the mouse. The user can assign the analysis to selected word form tokens a) in the middle part either by writing it in the editable window or by clicking on a proposed analysis; b) in the lower part by clicking on the word token, which opens a drop down menu. Further options are available, explained in the user manual.

A special feature is the ability to assign analyses to a group of word tokens, e.g. when multiple word tokens in the historical text correspond to a single modern word. Multiple analyses can also be assigned to a single word token, e.g. if one historical word form corresponds to several modern ones.

Working on historical language, the need occasionally arises to correct the transcription. This can be done by Ctrl-clicking the word form in context in the lower screen. An editable box will appear in which the user can correct a typo or separate merged words.

3 Data import and export

3.1 Corpus import and export

CoBaLT input corpus files can be in arbitrary formats, as long as the tokens, and possibly their annotations, are indicated in the texts, and appropriate import routines are in place. The tool currently accepts plain text and a parameterisation of TEI P5 XML (TEI Consortium, 2007). The latter option is more interesting for our case, as TEI files can already be structurally and linguistically annotated. Zip files are also supported, which enables uploading large datasets with many separate files.

The Slovene corpora are encoded in TEI, and each corpus file contains the transcription of a single page, together with the link to its facsimile image. The page is also annotated with paragraphs, line breaks, etc. Such annotation is imported into CoBaLT but not displayed or modified, and appears again only in the export.

The texts in our project were first automatically annotated (Erjavec, 2011): each text was sentence segmented and tokenised into words. Punctuation symbols (periods, commas, etc.) and white-spaces were preserved in the annotation so the original text and layout can be reconstructed from the annotated text. Each word form was assigned its modern-day equivalent, its PoS tag and modern day lemma.

Such files, a number of them together constituting one corpus, were then imported into CoBaLT and manually edited, with CoBaLT supporting the export of the annotated corpus as TEI P5 as well. In the export, each validated token is additionally annotated with the annotator's username and time of annotation.

One particular facet of the annotation concerns the word-boundary mismatch between the historical and modern-day word forms. As mentioned, CoBaLT supports joining two words in the transcription to give them a common annotation, as well as giving several successive annotations to a single word, and this is also reflected in the exported TEI annotation.

3.2 Lexicon export

While it is of course possible to produce a direct SQL dump of the lexicon, CoBaLT also supports lexicon export in TEI P5 using the TEI dictionaries module. This lexicon is headword (lemma) oriented. The lemma entry in the export consists of a headword, part of speech and optionally a gloss. The entry also contains all the modern word forms of the lemma as annotated in the corpus. For each modern word form one or more historical word forms are listed, including their normalised and cited forms. The difference between normalised and cited forms is that cited forms are the exact word forms as they appear in the corpus, while the normalised ones are lower-cased, and, in the case of Slovene, have vowel diacritics removed as these are not used in contemporary Slovene and are furthermore very inconsistently used in historical texts. These normalised forms are also what is listed in the left column of the middle part of the CoBaLT window. As illustrated in Figure 2, one cited form with examples of usage is "gláſnikam", the normalised form "glaſnikam", the modernised one "glasnikom" and the lemma form "glasnik", which is a common noun of masculine gender. This word does not exist anymore, so it is assigned a gloss, i.e. its contemporary equivalent "samoglasnik" (meaning "vowel").

```
<entry>
 <form type="lemma">
  <orth type="hypothetical">glasnik</orth>
  <gramGrp>
   <gram type="msd">Ncm</gram>
   <gram type="PoS">Noun</gram>
   <gram type="Type">common</gram>
   <gram type="Gender">masculine</gram>
  </gramGrp>
  <gloss>samoglasnik</gloss>
  <bibl>kontekst, Pleteršnik</bibl>
  <lbl type="occurrences">1</lbl>
 </form>
 <form type="wordform">
  <orth type="hypothetical">glasnikom</orth>
  <form type="historical">
   <orth type="normalised">glaſnikam</orth>
   <form type="cited">
    <orth type="exact">gláſnikam</orth>
    <cit>
     <quote>kadar beſeda, ktira naſléduje,
      sazhénja s' enim <oVar>gláſnikam</oVar>
      al tudi s' enim/quote>
     <bibl>NUK_10220-
1811.pb.007_Pozhetki_gramatike.xml</bibl>
    </cit>
   </form>
  </form>
 </form>
</entry>
```

Figure 2. Example of a TEI dictionary entry

The cited forms also contain examples of usage together with the file they occurred in. The export script can be limited as to how many usage examples get exported, as in the case of a fully annotated corpus the number of attestations for high-frequency words (typically function words) can easily go into the thousands, and there is little point in including all of them in the lexicon.

The export script also accepts parameters that determine which word forms should be exported – all, or only the attested or verified ones.

As in the corpus, the special case of multiword units and split words arises in the lexicon as well. Multiword units have the lemma and modern day forms composed of multiple words, and multiple grammatical descriptions, one for each lemma, while split words have the historical word forms composed of two or more words.

Also included with CoBaLT is a script to merge two TEI lexica (e.g. derived from different corpora) into a single TEI lexicon and to convert the TEI lexicon into HTML for web browsing. We extended this script for the case of Slovene to also give direct links to several on-line dictionaries and to the concordancer that hosts our corpora.

4 Discussion

4.1 Strengths and weakness of CoBaLT

First, it should be noted that CoBaLT is not limited to working with corpora of historical language – it could also be used for non-standard language varieties (e.g. tweets) or for standard contemporary language, by slightly modifying the import/export and the parsing of the word annotation in the editor. Nevertheless, it incorporates several features that make it particularly suitable for handling historical texts:

- CoBaLT supports both corpus annotation and corpus-based lexicon construction; extensive lexica are, at least from the point of view of good processing of historical language, much more important than annotated corpora.
- The texts of historical corpora are typically first produced by optical character recognition (OCR) software and then manually corrected. In spite of corrections, some errors will invariably remain in the text and will be, for the most part, noticed during the annotation process. While not meant for major editing of the transcription, CoBaLT does offer the possibility to correct the transcription of individual words. This is a rare functionality in other annotation editors, which typically take the base text as read-only. The current version of CoBaLT offers support for editing, splitting, and joining word tokens. Deleting word forms altogether, however, is not supported – an option that should be added in the future.
- Related to the previous point is CoBaLT's feature to display the facsimile of a particular page, making it possible to check the transcription or OCR result against the original image of the page.

As regards the functioning of the tool, it is important to note that almost all linguistic processing occurs outside of CoBaLT making it more lightweight as well as more language independent. In previous work (Erjavec et al., 2010) a different editor was used which had linguistic processing built in and proved to be more difficult to adapt to Slovene than CoBaLT.

In this particular project we decided to organise the files around the concept of a facsimile page. This has a number of advantages, in particular a straight-forward mapping between files and facsimile images, a simple unit of sampling for the corpus, and small files, which makes it easier to manage the work of annotators. However, this arrangement causes some problems from a linguistic point of view, namely that the page will often start or end in the middle of a paragraph, sentence or even word. We decided to start and end each page with a paragraph or sentence boundary, while split words are marked by a special PoS tag. It should be noted that this is used only at page-breaks – split words at line-breaks are joined before importing the texts into CoBaLT.

From a user-interface perspective, a distinguishing feature of CoBaLT is that there is a single editor window, with keyboard shortcuts making the jumps between the parts of the screen faster than moving a mouse, allowing for quick and efficient editing. Adding or deleting a number of analyses is also just a click away. This again makes the tool very efficient but also means that the user has to be quite careful not to accidentally destroy already existing annotations – this proved to be a problem in the annotation round.

From an implementation standpoint, we should note that the level of security offered by CoBaLT is limited. Only a user name is needed to log in and have access to the data. While this can be easily circumvented by placing the entire interface behind a secure page, a higher level of security, e.g. just adding passwords to the login procedure, should be implemented in the future. On the other hand, access should not be too restricted, as simple access does allow for easy crowdsourcing.

4.2 Related work

Historical corpora have been compiled, annotated and made available for searching in a number of projects, such as Corpus of Historical American English (Davies, 2010), Penn Corpora of Historical English (Kroch et al., 2004), GermanC historical corpus (Durrell et al., 2007), Historical Corpus of the Welsh Language (Mittendorf and Willis, 2004) and Icelandic Parsed Historical Corpus (Wallenberg et al., 2011), etc. Surprisingly few of these initiatives have developed or discussed the need for a historical text platform that would enable manual correction of pre-annotated corpora, facilities for lexicon building, and a standardized annotation format.

As the simplest solution, some of the projects used general-purpose XML. However, human annotators usually have a hard time working in XML directly to revise word-level annotations and transcription errors. This is one of the reasons why automatic and manual corpus-development tasks were integrated into the same environment in the GermanC project (Scheible et al., 2010), where the GATE platform (Cunningham et al., 2002) was used to produce the initial annotations and to perform manual corrections. However, GATE does not provide explicit support for texts encoded according to the TEI P5 guidelines, which is why the GermanC team spent a lot of time on writing scripts to deal with formatting issues. As GATE has automatic processing integrated into it, it is also not trivial to adapt it to a new language.

The only special-purpose tools for historical corpus development we could find is E-Dictor, a specialized tool for encoding, applying levels of editions and assigning PoS tags to ancient texts for building the Tycho Brahe Parsed Corpus of Historical Portuguese (de Faria et al., 2010). It is similar to CoBaLT in that it too has a WYSIWYG interface and allows annotators to check transcriptions and assign several layers of annotations to the tokens. E-Dictor enables export of the encoded text XML and the lexicon of editions in HTML and CSV. This is an interesting tool although it does not seem to support a lexical view of the data or merging and splitting word forms, and it is not quite clear how it interacts with automatic processing of the texts, or if a user manual is available.

As the review of related work shows, there is a general lack of tools such as CoBaLT which can significantly simplify and speed up most historical corpus and lexicon development projects. We believe CoBaLT has a number of qualities that will make it attractive for other researchers.

5 Conclusions

The paper presented CoBaLT, an editor for constructing corpus-based lexica and correcting word-level annotations and transcription errors in corpora. The editor has been extensively tested in a project in which a historical corpus was manually annotated and used to produce a lexicon, with the lexicon being further extended on the basis of a much larger corpus. Seven annotators have worked on the resources for over half a year, which put the tool through a good stress test. CoBaLT has also been used in several similar projects for other languages, in particular in producing historical lexica for Czech, Polish, Dutch and Spanish (de Does et al., 2012).[1]

With the help of CoBaLT Slovene now has two essential historical language resources, both encoded in TEI P5. The resources will be used to build better models for (re)tokenisation, transcription, tagging and lemmatisation, and to facilitate corpus-based diachronic language studies. We also plan to continue using CoBaLT to further extend the hand-annotated corpus and lexicon.

CoBaLT is freely available for research use from the Web site of the Impact Centre of Competence, http://www.digitisation.eu. The distribution contains the code, user manual, and associated scripts mentioned in this paper.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. The work presented in this paper has been supported by the EU IMPACT project "Improving Access to Text", http://www.impact-project.eu.

References

Mark Davies. 2010. *The Corpus of Historical American English (COHA): 400+ Million Words, 1810–2009.* http://corpus.byu.edu/coha

Jesse de Does, Katrien Depuyd, Klaus Schulz, Annette Gotscharek, Christoph Ringlstetter, Janusz S. Bień, Tomaž Erjavec, Karel Kučera, Isabel Martinez, Stoyan Mihov, and Gilles Souvay. 2012. *Cross-language Perspective on Lexicon Building and Deployment in IMPACT.* Project Report. IMPACT.

Tomaž Erjavec, Christoph Ringlstetter, Maja Žorga, and Annette Gotscharek. 2010. Towards a Lexicon of XIXth Century Slovene. In Proceedings of the Seventh Language Technologies Conference, Ljubljana, Slovenia. Jožef Stefan Institute.

Tomaž Erjavec. 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ACL.

Tomaž Erjavec. 2012. The goo300k corpus of historical Slovene. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12*, Paris, ELRA.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME).* Department of Linguistics, University of Pennsylvania. CD-ROM, first edition. http://www.ling.upenn.edu/hist-corpora/

Martin Durrell, Astrid Ensslin, and Paul Bennett. 2007. The GerManC project. *Sprache und Datenverarbeitung,* 31:71–80.

Ingo Mittendorf, and David Willis, eds. 2004. *Corpws hanesyddol yr iaith Gymraeg 1500–1850 / A historical corpus of the Welsh language 1500–1850.* http://people.pwf.cam.ac.uk/dwew2/hcwl/menu.htm

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.9. http://www.linguist.is/icelandic_treebank

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett, 2010. Annotating a Historical Corpus of German: A Case Study. *Proceedings of the LREC 2010 Workshop on Language Resources and Language Technology Standards*, Valletta, Malta.

Hamish Cunningham. 2002. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254.

Pablo Picasso Feliciano de Faria, Fabio Natanael Kepler, and Maria Clara Paixão de Sousa. 2010. An integrated tool for annotating historical corpora. *Proceedings of the Fourth Linguistic Annotation Workshop*, ACL'10, 217–221.

TEI Consortium, eds. 2007. *Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/P5

[1] For more information on these projects please see the Impact Centre of Competence: http://www.digitisation.eu/