OTTO: A Transcription and Management Tool for Historical Texts

Stefanie Dipper, Lara Kresse, Martin Schnurrenberger & Seong-Eun Cho
Institute of Linguistics, Ruhr University Bochum
D – 44780 Bochum
dipper@linguistics.rub.de, lara.kresse@rub.de,
martin.schnurrenberger@rub.de, seong-eun.cho@rub.de

Abstract

This paper presents OTTO, a transcription tool designed for diplomatic transcription of historical language data. The tool supports easy and fast typing and instant rendering of transcription in order to gain a look as close to the original manuscript as possible. In addition, the tool provides support for the management of transcription projects which involve distributed, collaborative working of multiple parties on collections of documents.

1 Corpora of Historical Languages[1]

The only way to study historical languages is, of course, by looking at texts, or corpora from these languages. Compared to texts from modern languages, early manuscripts or prints pose particular challenges. Depending on physical condition of the manuscripts, passages can be hard to decipher, or pages can be damaged or missing completely. Some texts contain words or passages that have been added later, e.g., to clarify the meaning of a text segment, or to correct (real or assumed) errors.

Moreover, historical texts exhibit a large amount of character peculiarities (special letters, punctuation marks, abbreviations, etc.), which are not easily encoded by, e.g., the ASCII encoding standard. For instance, medieval German texts often use superscribed letters to represent emerging or remnant forms of diphthongs, e.g. $\overset{o}{u}$. Some texts distinguish two forms of the (modern) letter <s>, the so-called short vs. long s: <s> vs. < >. Conversely, some texts do not differentiate between the (modern) letters <u> and <v>.

The existence of letter variants is often attributed to aesthetic reasons or to save (expensive) space. Thus, when early manuscripts are to be transcribed, it must first be decided whether the differences between such variants are considered irrelevant and, hence, can be safely ignored, or whether they constitute a (possibly) interesting phenomenon and potential research issue.

This discussion relates to the *level of transcription*, i.e. "how much of the information in the original document is included (or otherwise noted) by the transcriber in his or her transcription" (Driscoll, 2006). *Diplomatic transcription* aims at reproducing a large range of features of the original manuscript or print, such as large initials or variant letter forms.

Another important issue with historical corpora is meta-information. A lot of research on historical texts focuses on the text proper and its content, rather than its language. For instance, researchers are interested in the history of a text ("who wrote this text and where?"), its relationship to other texts ("did the writer know about or copy another text?"), its provenance ("who were the owners of this text?"), or its role in the cultural context ("why did the author write about this subject, and why in this way?"). To answer such questions, information about past and current depositories of a manuscript, peculiarities of the material that the text is written on, etc. are collected. In addition, any indicator of the author (or writer) of the text is noted down. Here, the text's language becomes relevant as a means to gather information about the author. Linguistic features can be used to determine the text's date of origin and the author's social and regional affiliation. Usually, this kind of information is encoded in the *header* (see, e.g., the TEI header (TEI Consortium (eds), 2007)).[2]

From the above, we derive the following requirements:

Above all, use of *Unicode* is indispensable, to

[1]The research reported in this paper was financed by Deutsche Forschungsgemeinschaft, Grant DI 1558/1-1.
We would like to thank the anonymous reviewers for their helpful comments.

[2]Text Encoding Initiative, www.tei-c.org

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 182–185,
Uppsala, Sweden, 15-16 July 2010.

be able to encode and represent the numerous special symbols and characters in a reliable and sustainable way. Of course, not all characters that occur in historical texts are already covered by the current version of Unicode. This is especially true of character *combinations*, which are only supported partially (the main reason being that Unicode's Combining Diacritical Marks focus on superscribed diacritics rather than characters in general). Therefore, Unicode's Private Use Area has to be used as well.

Similarly, there are characters without glyphs defined and designed for them. Hence, an ideal transcription tool should support the user in creating new glyphs whenever needed.

Since there are many more characters in historical texts than keys on a keyboard, the transcription tool must provide some means to key in all characters and combinations (similar issues arise from logographic scripts, such as Chinese). In principle, there are two ways to do this:

(i) The transcriber uses a virtual keyboard, which supports various character sets simultaneously and is operated by the mouse. Virtual keyboards are "WYSIWYG" in that their keys are labeled by the special characters, which can then be selected by the user by mouse clicks. As is well known, virtual keyboards are often preferred by casual users, beginners, or non-experts, since they are straightforward to operate and do not require any extra knowledge. However, the drawback is that "typing" with a computer mouse is rather slow and tedious and, hence, not a long-term solution.

(ii) Alternatively, special characters, such as "$", "@", etc., are used as substitutes for historical characters, commonly in combination with ordinary characters, to yield a larger number of characters that can be represented. Regular and advanced users usually prefer substitute characters to virtual keyboards, because once the user knows the substitutes, typing them becomes very natural and fast. Of course, with this solution transcribers have to learn and memorize the substitutes.

Some tools convert substitutes to the actual characters immediately after typing (this is the case, e.g., with shortcuts in Emacs), while others require additional post-processing by interpreters and viewers to display the intended glyphs (e.g., LaTeX encodings converted to postscript). Immediate preview seems advantageous in that it provides immediate feedback to the user. On the other hand, it might be easier to memorize substitutes if the user can actually see them.

Which input method is to be preferred for historical data? Transcription projects often involve both beginners and advanced users: having people (e.g. student assistants) join and leave the team is rather often the case, because transcribing is a very labor- and time-intensive task.

Our transcription tool OTTO faces this fact by combining the advantages of the two methods. The user types and views character substitutes but simultaneously gets feedback in a separate window about whether the input is correct or not. This lessens the uncertainty of new team members and helps avoiding typing mistakes, thus increasing the quality of transcription.

Another important requirement is the possibility to mark additions, deletions, uncertain readings, etc. To encode such information, TEI also provides a standardized representation format.

Finally, projects that involve multiple parties distributed over different sites add a further requirement. In such scenarios, tools are preferably hosted by a server and operated via a web browser. This way, there is no need of multiple installations at different sites, and data on the server does not need to be synchronized but is always up to date.

To our knowledge, there is no transcription tool that (i) would support Unicode, (ii) allow for fast typing, using character substitutes, and (iii) is web-based. In MS Word, special characters are usually inserted by means of virtual keyboards but character substitutes can be defined via macros. However, macros often pose problems when Word is upgraded. Moreover, Word is not web-based. LaTeX, which supports character substitutes, is often considered too complex for non-expert users, does not offer instant preview, and is not web-based.

2 The Transcription Tool OTTO[3]

OTTO is an online transcription tool for editing, viewing and storing information of historical language data. OTTO's data model is a directed graph. Nodes point to a (possibly empty) stretch of primary data and are labeled.

The tool is written in PHP and also uses some Java Script; data is stored in a mySQL database.

[3] A prior version of OTTO has been described in Dipper and Schnurrenberger (2009).

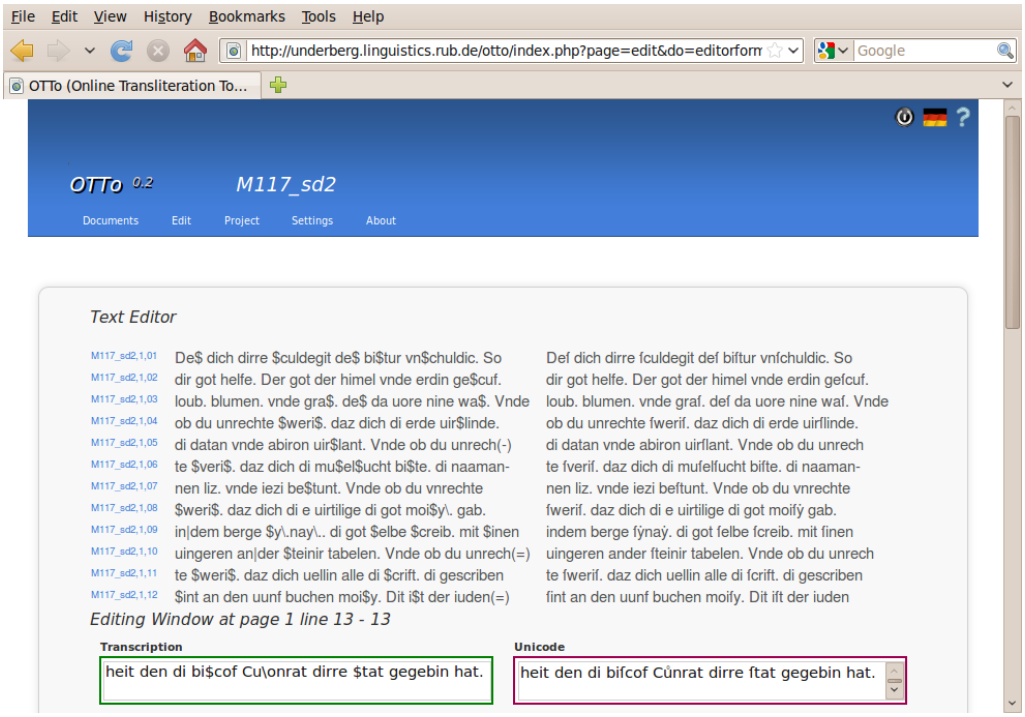

Figure 1: Screenshot of the text editor

Any server which runs PHP >5.2 can be a host for OTTO. Users can login to the tool from anywhere using a standard web browser. A live demo of OTTO, with slightly restricted functionality, can be tried out here: `http://underberg.linguistics.rub.de/ottolive`.

2.1 Transcribing with OTTO

OTTO integrates a user-definable header editor, to enter meta information about the manuscript, such as its title, author, date of origin, etc. However, the tool's core feature is the text editor. The upper part of the text editor in Fig. 1 displays the lines that have been transcribed and saved already. Each line is preceded by the bibliographic key, *M117_sd2*, the folio and line numbers, which are automatically generated.

The bottom part is dominated by two separate frames. The frame on the left, called *Transcription*, is the currently "active" field, where the user enters the transcription (or edits an existing one). The transcriber can use substitute characters to encode non-ASCII characters. In the figure, the dollar sign (`$`) serves as a substitute for long s (< >), see the first word of the text, *De$*), and `u\o` stands for ů (see *Cu\onrat* in the Transcription field at the bottom).

The frame on the right, called *Unicode*, directly transforms the user input to its diplomatic transcription form, using a set of transcription rules. The diplomatic Unicode view thus provides immediate feedback to the transcriber whether the input is correct or not.

Transcription rules have the form of "search-and-replace" patterns. The first entity specifies the character "to be searched" (e.g. `$`), the second entity specifies the diplomatic Unicode character that "replaces" the actual character. Transcription rules are defined by the user, who can consult a database such as the ENRICH Gaiji Bank[4] to look up Unicode code points and standardized mappings for them, or define new ones. OTTO uses the Junicode font, which supports many of MUFI's medieval characters, partly defined in Unicode's Private Use Area.[5]

Rules can be defined locally—i.e., applying to the current transcription only—or globally, i.e., applying to all documents contained in OTTO's database.[6] The rules are used to map the lines entered in the Transcription frame to the lines in diplomatic form in the Unicode frame.

OTTO allows for the use of comments, which

[4] `http://beta.manuscriptorium.com/`

[5] Junicode: `http://junicode.sourceforge.net/`; MUFI (Medieval Unicode Font Initiative): `http://www.mufi.info/`

[6] Global rules can be thought of as the application of a project's transcription criteria; local rules can be viewed as handy abbreviations defined by individual users.

can be inserted at any point of the text. Since the current version of OTTO does not provide special means to take record of passages that have been added, deleted, or modified otherwise, the comment mechanism could be exploited for this purpose.

The transcription, both in original (typed) and in Unicode version, can be exported to a (customized) TEI-conform XML format. Transcription rules are optionally included in the header.

2.2 Transcription Projects

Projects that deal with the creation of historical corpora often involve a cascade of successive processing steps that a transcription has to undergo. For instance, high-quality transcriptions are often entered twice, by two transcribers independently from each other, and their outcomes are compared and adjusted. In the case of diplomatic transcriptions, a further step called *collating* is necessary. Collating means comparing the transcription and the original manuscript in full detail. Often two people are involved: One person reads out the manuscript letter for letter, and also reports on any superscript, white-space, etc. The other person simultaneously tracks the transcription, letter for letter. This way, high-quality diplomatic transcription can be achieved.

To cope with the numerous processing steps, transcription projects often involve a lot of people, who work on different manuscripts (or different pages of the same manuscript), in different processing states.

OTTO supports such transcription projects in several aspects: First, it allows for remote access to the database, via standard web browsers. Second, documents that are currently edited by some user are locked, i.e., cannot be edited or modified otherwise by another user. Third, OTTO provides facilities to support and promote communication among project members. Finally, graphical progress bars show the progress for each transcription, measuring the ratio of the subtasks already completed to all subtasks,

3 Conclusion and Future Work

This paper presented OTTO, an online transcription tool for easy and fast typing, by the use of user-defined special characters, and, simultaneously, providing a view on the manuscript that is as close to the original as possible. OTTO also supports distributed, collaborative working of multiple parties on collections of documents.

Future work includes adding further support for transcribing special characters. First, we plan to integrate a virtual keyboard for casual users. The keyboard can also be used in the creation of transcription rules, in order to specify the Unicode replacement characters, or if the user wants to look up the substitute character defined for a specific Unicode character in the set of transcription rules.

We plan to use the TEI *gaiji* module for the representation of transcription rules and substitute characters; similarly, elements from the TEI *transcr* module could be used for the encoding of additions, deletions, etc.[7]

For facilitating the collation process, we plan to integrate transparent overlays. The user would have to rescale an image of the original manuscript and adjust it to the transcription, so that corresponding characters would match.

OTTO is designed as to allow for adding custom functions, by being programmed according to the paradigm of object-oriented programming. Additional functionality can easily be integrated (known as Plug-Ins). We currently work on integrating a normalizer into OTTO which maps spelling and dialectal variants of word forms to a standardized word form (Schnurrenberger, 2010).

OTTO will be made freely available to the research community.

References

Stefanie Dipper and Martin Schnurrenberger. 2009. OTTO: A tool for diplomatic transcription of historical texts. In *Proceedings of 4th Language & Technology Conference*, Poznan, Poland. To appear.

Matthew J. Driscoll. 2006. Levels of transcription. In Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, editors, *Electronic Textual Editing*, pages 254–261. New York: Modern Language Association of America. URL: `http://www.tei-c.org/About/Archive_new/ETE/Preview/driscoll.xml`.

Martin Schnurrenberger. 2010. Methods for graphemic normalization of unstandardized written lang uage from Middle High German Corpora. Master's thesis, Ruhr University Bochum.

TEI Consortium (eds). 2007. TEI P5: Guidelines for electronic text encoding and interchange. `http://www.tei-c.org/Guidelines/P5/`.

[7] `http://www.tei-c.org/release/doc/tei-p5-doc/html/WD.html` and `PH.html`