OTTO: A Tool for Diplomatic Transcription of Historical Texts

Stefanie Dipper and Martin Schnurrenberger

Linguistics Department
Ruhr University Bochum, Germany
dipper@linguistics.rub.de | martin.schnurrenberger@rub.de

Abstract

In this paper, we present OTTO, a new transcription tool which is designed for diplomatic transcription of historical language data. The tool supports easy and fast typing and, at the same time, renders the transcription as close to the original manuscript as possible. It also allows for the annotation of rich, user-defined header information.

1. Introduction[1]

Since the first days of corpus-based linguistic investigations, historical language data has been in the focus of research. Starting with the Dominican cardinal Hugh of St Cher, who in 1230 compiled the first concordance of the Bible (more precisely, of the Latin translation *Vulgate*), up to Johann Jakob Griesbach, who published the first Greek Gospel synopsis in 1776. Concordances served (and still serve) as the basis for comparing the meaning and usage of specific words in different texts, such as the books of the Bible. Synopses are often used to reconstruct lost original sources, or to construct a stemma, i.e. the relationships and dependencies between different text witnesses (different text versions of the same underlying content).

With the advent of electronic corpora in the 1960s and 1970s, focus shifted to modern languages, with recent data, because machine-readable texts were more easily available for modern languages than historical ones. A notable exception is the *Helsinki Corpus of English Texts*, a corpus of diachronic English data, compiled at the University of Helsinki between 1984 and 1991.[2]

Early manuscripts (or prints) exhibit a large amount of peculiarities (special letters, punctuation marks, abbreviations, etc.), which are not easily encoded by, e.g., the ASCII encoding standard. Hence, an important issue with historical corpora is 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 (e.g. short vs. long s: <s> vs. <f>).

Another matter is the amount of variation in language: prior to the emergence of a standard national language with orthographic regulations, texts were written in dialects, rendering dialectal vocabulary and pronunciation in a more or less accurate way. Important texts, such as *The Song of the Nibelungs*, have often been handed down in a large variety of witnesses, which, to a greater or lesser extent, differ from each other with regard to content and language (dialect). In the 19th century, Karl Lachmann, one of the formative scientist in stemmatics, created a kind of "ideal", artificial language for texts written in different dialects from Middle High German (MHG). This language "normalizes" and levels out regional differences and thus facilitates comparison and understanding of MHG texts. Of course, on the other hand, it impedes in-depth linguistic research because the languages of these texts are, in a certain sense, corrupted.

Unfortunately, the normalized language has been widely used in editions for MHG texts. Hence, electronic corpora that are based on such editions are useful only to a certain extent. As a consequence, a new project has been launched at the Universities of Bochum and Bonn, entitled "Reference Corpus Middle High German (1050–1350)", which aims at creating a reference corpus of MHG texts that (i) does not make use of normalized text editions but (scans of) original manuscripts only, and (ii) applies diplomatic transcription. The project group has more than 20 years of experience with transcribing and annotating historical texts. Until recently, however, they used ordinary text processing tools for transcribing. The transcription process is followed by a semi-automatic annotation procedure, which includes word form normalization, lemmatization, morphological analysis, and POS tagging.

In this paper, we present a new tool, *OTTO* ("Online Transcription TOol"), which is designed for diplomatic transcription of historical texts. It provides interfaces for text viewing and editing and entering of header information. Output formats are XML or plain text.

The paper is organized as follows. In Sec. 2., we list requirements specific to historical language data that transcription tools have to meet. Sec. 3. presents related work. Sec. 4. introduces OTTO, followed by concluding remarks in Sec. 5..

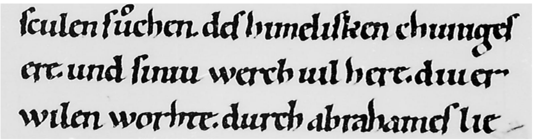
2. Requirements of transcription tools

2.1. Characteristics of historical texts

Diplomatic transcription aims at rendering a manuscript as original as possible, so that virtually no interpretation is involved in the transcription process. However, certain decisions still have to be made. Some of these decisions can be made once and for all, and apply to the transcription of the entire corpus. These are conventions, which are specified in the form of guidelines

[1] The research reported in this paper was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/1-1. All URLs provided in this paper have been accessed 2009, July 30.

[2] `http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/`



ſculen ſuͦchen. deſ himeliſken chuniges
ere und ſinu werch uil herte. diu er-
wilen worhte. durch abrahameſ lie

Figure 1: Fragment of the *Altdeutsche Exodus* (Vienna, Austrian National Library, Cod. 2721, f.159r, 12th century); for a transcription of the text, see Fig. 2.

that have to be followed. Other decisions are up to the transcriber and have to be made as the case arises in the course of the transcription process, e.g. if the language's alphabet contains characters and symbols that can be mixed up and the transcriber has to decide which character is at hand.

Transcription guidelines specify how to transcribe letters that do not have a modern equivalent. They also specify which letter forms represent variants of one and the same character, and which letters are to be transcribed as different characters. Relevant cases include "normal" and tailed z (<z> vs. <ʒ>) and short vs. long s (<s> vs. <ſ>). In both cases, one of the variants has been abolished in the course of time: in modern European alphabets, only <z> and <s> are still used. This means that there is no straightforward one-to-one mapping between medieval and modern alphabets. Abolished letters have to be encoded in a special way.

Frequently, (remnant or emerging forms of) diphthongs are rendered in a way that the second vowel is superscribed over the first vowel, as in *ſuͦchen* 'seek', see the second word of the fragment displayed in Fig. 1. In modern (European) alphabets, only diacritics such as accents or umlaut can be used as superscripts.

To save space and time, medieval writers used a lot of abbreviations. For instance, nasal <n> or <m> is often encoded by a superscribed horizontal bar ("Nasalstrich") as in *vō*, which stands for *von* 'from'. A frequently-used abbreviation is *vn̄* for the conjunction *und/unde* 'and'. A superscribed hook ("er-Kürzung") abbreviates *er* (or *re, r*, and rarely *ri, ir*), as in *mart'*, which represents *marter* 'martyrdom'.

Another kind of special characters are initials, which can range over the height of two, three or even five lines, and need also to be encoded in some way.

Further, medieval texts often 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. Such additions or corrections can be made either on top of the line that is concerned, or else at the margin of the page. A special case is provided by inter-linear translations or *glosses*, where, e.g., a German word-for-word translation is superscribed over the Latin original text.

Finally, the layout of the texts (lines, columns, front and back page) should be encoded in the transcription. First, this information provides the usual access to the texts; text positions are usually specified by these coordinates. Second, information about line breaks is essential for lyrics, and could be useful for determining word boundaries in prose.

Let us briefly summarize the main encoding issues with historical texts:

- Encoding of letters, symbols, and combinations thereof that do not exist in modern alphabets
- Encoding of abbreviations
- Encoding of later additions
- Encoding of layout information
- For bilingual glosses: encoding of alignment (word-for-word correspondences)

2.2. Meta-information: header and comments

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.

This kind of meta-information, which pertains to the entire text, is encoded in the *header*. Typical header information further includes observations of all kinds of peculiarities of the text under consideration, such as special writing conventions ("writer uses a peculiar 'ff' ligature") or uncertainties within the transcription ("exact placement of the circumflex accent is often unclear; in the transcription it is always placed on the first letter").

Similar meta-information can be encoded in the form of *comments*, if it only concerns specific parts within the text rather than the text as a whole. Comments are used, e.g., for passages that are not well readable, that are destroyed, or otherwise questionable. Transcriber use them to mark uncertainties, to mark remarkable properties of letter or word forms, or to mark later additions/corrections. This information could be used for later (semi)automatic creation of a critical apparatus.

To summarize the encoding issues related to meta-information:

- Encoding of information about the text, its author and/or writer (header)
- Marking of text peculiarities (header and comments)

2.3. Requirements for transcription tools

The characteristics of (research on) historical texts that we identified in the previous sections put specific requirements on transcription tools.

Diplomatic transcription Above all, use of Unicode is indispensable, to 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. In principle, there are two ways to do this: the transcriber can use a virtual keyboard, which can support various character sets simultaneously and is operated by the mouse. Or else, special characters, such as "$", "@", "(", "#", etc., are used as substitutes for historical characters; these characters are commonly used in combination with ordinary characters, to yield a larger number of characters that can be represented. Of course, with this solution transcribers have to learn and memorize the substitutes.

Given the fact that each text can exhibit its own letter forms and writing conventions, it must be possible to customize the tool and adapt it to individual texts.

Meta-information The tool must provide suitable means for encoding header information. To promote use of standardized values (and to minimize the risk of typos), the header should provide drop-down menus or radio buttons wherever possible. For other features, the tool must provide free-text input. Again, these settings are highly dependent on the text that is transcribed and on the project's goal, and, hence, the tool should be customizable in these respects.

Work flow Projects that deal with the creation of historical corpora often distinguish two processes: (i) transcribing the manuscript, (ii) *collating* the manuscript, i.e., comparing the original text and its transcription in full detail. Often two people are involved: One person reads out the manuscript letter for letter, and also reports on any superscript, whitespace, etc. The other person simultaneously tracks the transcription, letter for letter. This way, high-quality diplomatic transcription can be achieved.

This kind of workflow implies for the tool that there be an input mode that supports easy entering of new text, from scratch. In addition, there should be a collation mode, which allows the user to view and navigate within the text in a comfortable way, and to easily jump to arbitrary text positions where transcription errors have to be corrected.

3. Related work

Many projects that create corpora of historical languages derive their electronic text basis from printed editions since this saves a lot of work. To them, collating is a prominent step (if they collate at all—not all projects have enough funding to collate or have access to the original manuscript).

To our knowledge, there is currently no tool available which supports collating a transcription with its manuscript. There are some tools that support collating multiple *electronic* texts with each other, such as transcriptions of different text witnesses, or different printed editions from one and the same source. These tools help the user by aligning text passages from the individual texts that correspond to each other, just like in a synopsis. Such tools are, e.g., Juxta[3], TUSTEP[4], or the UNIX command *diff*. Another technique of collating involves visual merging of copies of the texts that are to be compared (e.g., by overlays). This method presupposes that the texts are suffiently similar, at the visual level.

Hence, there is no tool that would work with handwritten texts using old scripts, which often require expert readers for deciphering.

Similarly, for transcribing historical texts from scratch, there are no specific tools, to our knowledge. A task which is somewhat similar is (phonetic) transcription of speech data. There is a range of linguistic tools for this task, which all focus on the alignment of audio and transcription data, such as Praat[5], Exmaralda[6], or ELAN[7]. In fact, canonical usage of the term "transcription" applies to convertion from sound to characters. By contrast, "transliteration" means transforming one script into another script. We nevertheless stick to the term "transcription" since transcription can be viewed as a mapping from analog to digital data, whereas transliteration usually involves digital-to-digital mapping. Manuscripts obviously represent analog data in this sense.

4. OTTO

OTTO is an online transcription tool which is used through a standard web browser. OTTO is designed for high-quality diplomatic transcription of historical language data and supports distributed, collaborative working of multiple parties. It is written in PHP and uses MySQL as the underlying database. In the following, the currently-implemented features of OTTO are described in brief.

Import Besides creating a transcription file and starting on an empty sheet directly in OTTO itself, there often are other sources for transcription files, such as electronic editions, which still need to be collated. For importing these transcription files, OTTO provides the Import tab. It lists all available import sources, which the project group can define to fit their individual needs. Once a transcription file has been imported to OTTO, all further editing takes place within OTTO.

Files The Files tab lets members of the transcription team see which transcription files have already been transcribed within OTTO (or imported to OTTO) and are available for further editing. With OTTO, there is no need of having numerous copies of the same transcription file spread across several computers, leaving the transcriber in doubt about whether or not the file she finds is up to date. Since there is only one of any transcription file, having two

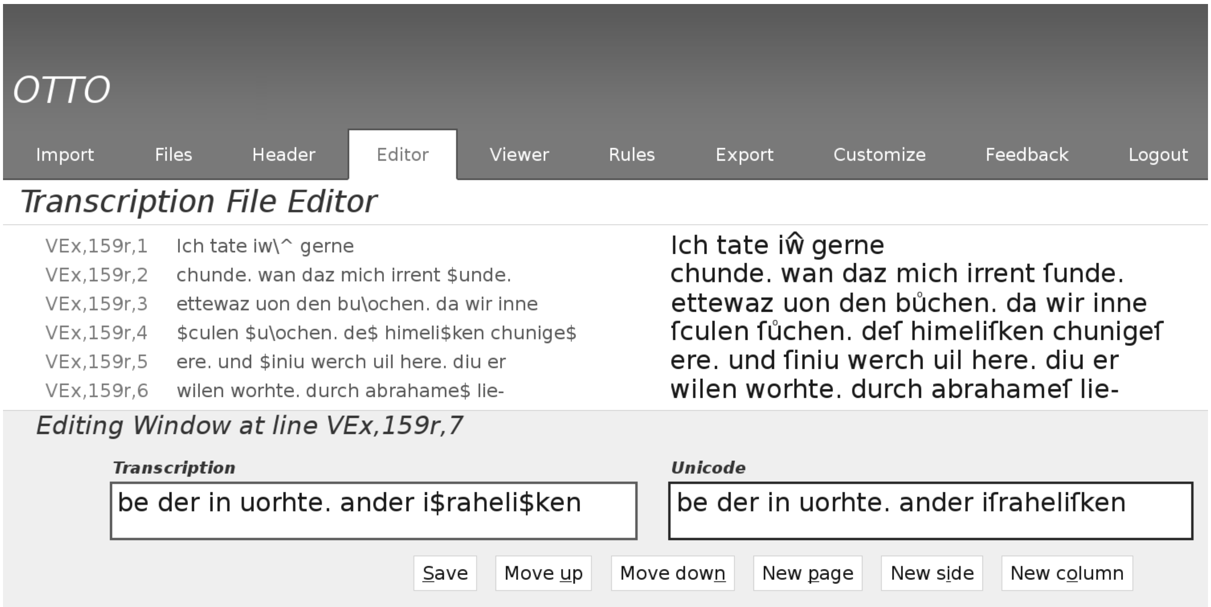
[3] `http://www.juxtasoftware.org`

[4] `http://www.zdv.uni-tuebingen.de/tustep`

[5] `http://www.fon.hum.uva.nl/praat`

[6] `http://www.exmaralda.org`

[7] `http://www.lat-mpi.eu/tools/tools/elan`



Transcription File Editor

VEx,159r,1	Ich tate iw\^ gerne	Ich tate iŵ gerne
VEx,159r,2	chunde. wan daz mich irrent $unde.	chunde. wan daz mich irrent ſunde.
VEx,159r,3	ettewaz uon den bu\ochen. da wir inne	ettewaz uon den bůchen. da wir inne
VEx,159r,4	$culen $u\ochen. de$ himeli$ken chunige$	ſculen ſůchen. deſ himeliſken chunigeſ
VEx,159r,5	ere. und $iniu werch uil here. diu er	ere. und ſiniu werch uil here. diu er
VEx,159r,6	wilen worhte. durch abrahame$ lie-	wilen worhte. durch abrahameſ lie-

Editing Window at line VEx,159r,7

Transcription

be der in uorhte. ander i$raheli$ken

Unicode

be der in uorhte. ander iſraheliſken

Figure 2: Screenshot of OTTO, displaying the editor interface with the text fragment of Fig. 1 in lines 4–6. Lines 1–6 have already been transcribed, line 7 is just being edited. Each line is preceded by the bibliographic key "VEx", the folio and line numbers, which are automatically generated.

transcribers work on the same file at the same time can lead to overwriting problems. OTTO faces this issue by keeping a file lock log. The moment a transcriber opens a transcription file it gets locked and the other members of the team will see that this file is in use by another transcriber. Also the name of the transcriber is displayed so members can negotiate turns.

Beneath the list of existing files, a new, empty transcription file can be created by one click. The transcriber is then asked to first enter the folio and line number of the first line that she is going to transcribe. This information is used to automatically create line counts for further lines.

Header The header of transcription files contains data about the file itself, its original corpus, its original corpus' origin, its transcription process, etc. Which information will be recorded in the header depends on the individual project's goals and resources. Hence, OTTO lets transcription teams define a customized but fixed header, which can for example contain preformatted values, thus reducing typing mistakes. Using fixed headers will make it more easy—or even possible—to exploit the information for further processing of the transcription files, or for use in a corpus search tool.

The header can be enriched with exactly-timed data on which person did which changes to any file.

Editor The Editor (see Fig. 2) is OTTO's core feature. The look and feel is highly customizable (see Customize further down below). It provides an editing window which resides at the current editing position. The editing position, when just having opened a file, is usually at the end of the file, so the transcriber can continue working right away. Usually she will enter a new line into the input field denoted as 'Transcription' (left frame). While she is doing this, the input field denoted as 'Unicode' (right frame) does a live (hence 'Online') transformation of her entered line into its actual diplomatic transcription form, using a set of rules (see Rules further down below). By keeping an eye on this online transformation, the transcriber gets feedback on whether her input was correct or not.

When the line or several lines have been transcribed, the new entry can be saved by a click on the 'Save' button. 'Move up' and 'Move down' will navigate the editing window up or down, if the transcriber wishes to edit the entire file in either direction. 'New Page', 'New side' and 'New column' will add marks to the current line, which are used for the automatically generated line counts (denoted in Fig. 2 as 'WEx,159r,7' for example).

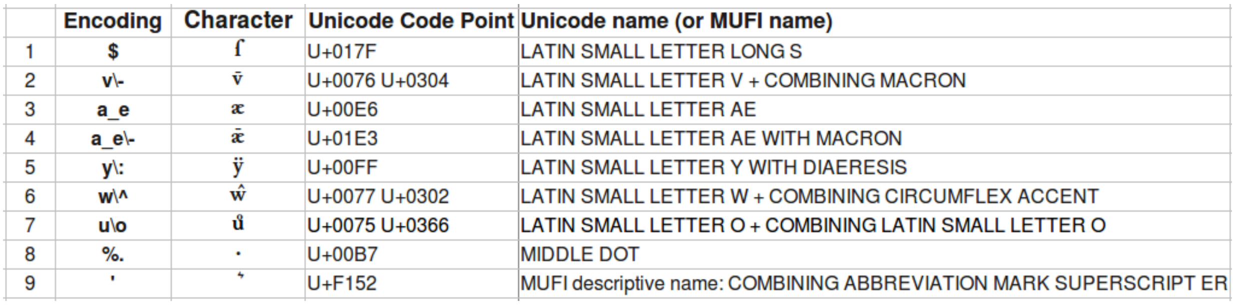
Above and below the editing window, all currently transcribed lines are displayed with their line count, the entered line and the diplomatic line generated by applying the transformation rules. The line counts also function as links for moving the editing window to a line of one's choice, in the act of proof reading or collating, for example.

Viewer The Viewer tab shows the transcription file in its original layout. It displays the diplomatic transcription in form of pages, page sides and columns. This format is well suited for collating and can be used to print out a paper version.

Rules The transcription group may define rules for transforming the entered lines into the diplomatic lines. These rules can be set up to be valid for all transcription files or just for the current file. Fig. 4. shows an excerpt of the rules that are currently defined in the Bochum MHG project.

In our project, abbreviations such as the horizontal bar or the *er* hook are not solved since we aim at diplomatic transcription. Other projects might want to define rules that replace abbreviations by the respective full forms.

OTTO uses UTF-8-encoded Unicode and Junicode charac-



	Encoding	Character	Unicode Code Point	Unicode name (or MUFI name)
1	$	ſ	U+017F	LATIN SMALL LETTER LONG S
2	v\-	v̄	U+0076 U+0304	LATIN SMALL LETTER V + COMBINING MACRON
3	a_e	æ	U+00E6	LATIN SMALL LETTER AE
4	a_e\-	ǣ	U+01E3	LATIN SMALL LETTER AE WITH MACRON
5	y\:	ÿ	U+00FF	LATIN SMALL LETTER Y WITH DIAERESIS
6	w\^	ŵ	U+0077 U+0302	LATIN SMALL LETTER W + COMBINING CIRCUMFLEX ACCENT
7	u\o	uͦ	U+0075 U+0366	LATIN SMALL LETTER O + COMBINING LATIN SMALL LETTER O
8	%.	·	U+00B7	MIDDLE DOT
9	'		U+F152	MUFI descriptive name: COMBINING ABBREVIATION MARK SUPERSCRIPT ER

Figure 3: Sample rules as used in the sample text in Fig. 2. For example, line 1 specifies “$” as a substitute for long <ſ>. Column 1 displays the character that the transcriber types, column 2 shows the target character in Junicode[8] font. Columns 3 and 4 supply the code points and names as defined by Unicode. Line 9 specifies the apostrophe as a substitute of the *er* hook, as defined in Unicode’s Private Use Area by MUFI (Medieval Unicode Font Initiative)[9] .

ters. This makes it robust for all types of transcriptions.

Export Transcriptions can be exported to a plain text format or XML.

Customize The Customize tab lets each user customize the look and feel of OTTO. For example, displaying font sizes can be set to fit the needs of every individual. The transcriber can also customize the number of lines she would like to edit at once.

Feedback The Feedback tab is a place for open communication within the transcribing team.

Logout Logging out is essential. It frees the files that have been locked by the transcriber and finalizes the additional header information on who did what and when.

We conclude this section with some considerations that led us to the design of OTTO as described above.

Any text-processing system that deals with special characters, which are not part of common keyboards, has to supply the user with some means as to input these characters. A frequently-chosen option is to provide a virtual keyboard. 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 an alternative, (combinations of) keys provided by standard keyboards can serve as substitutes of special characters. In such systems, a sequence such as “"a” would be automatically replaced, e.g., by the character “ä”. 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. By contrast, regular and advanced users usually prefer a system that provides character substitutes, because once the user knows the substitutes, typing them becomes very natural and fast.

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. OTTO faces these facts by combining the two methods. The user types and simultaneously gets feedback 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.

Line-by-line processing, as provided by OTTO, is modeled after the line-based way of transcribing diplomatically. The lines of text that are currently not part of the editing window are write-protected. This reduces the risk of accidentally modifying parts of the transcription.

5. Conclusion and future work

We have presented OTTO, a new transcription tool designed for diplomatic transcription of historical texts. Its main feature is to support easy and fast typing, by use of user-defined special characters, and, simultaneously, provide a view on the manuscript that is as close to the original as possible. It is planned to support users in creating their own Unicode glyphs.

Future steps include an XML export that is compliant to the TEI standards, with respect to the encoding of properties of the proper text (Burnard and Bauman, 2007a) as well as header information (Burnard and Bauman, 2007b).

To further support collating, we plan to experiment with integrating manuscript scans into the tool. Putting scan and transcription side by side, or even as overlays, could considerably facilitate collating, especially if a project cannot afford employing two people for this task.

The tool will be made freely available for non-commercial research purposes.

6. References

Burnard, Lou and Syd Bauman, 2007a. Representation of primary sources. In *P5: Guidelines for Electronic Text Encoding and Interchange*, chapter 11. TEI Consortium.

Burnard, Lou and Syd Bauman, 2007b. The TEI header. In *P5: Guidelines for Electronic Text Encoding and Interchange*, chapter 2. TEI Consortium.

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

[8] `http://junicode.sourceforge.net`

[9] `http://www.mufi.info`